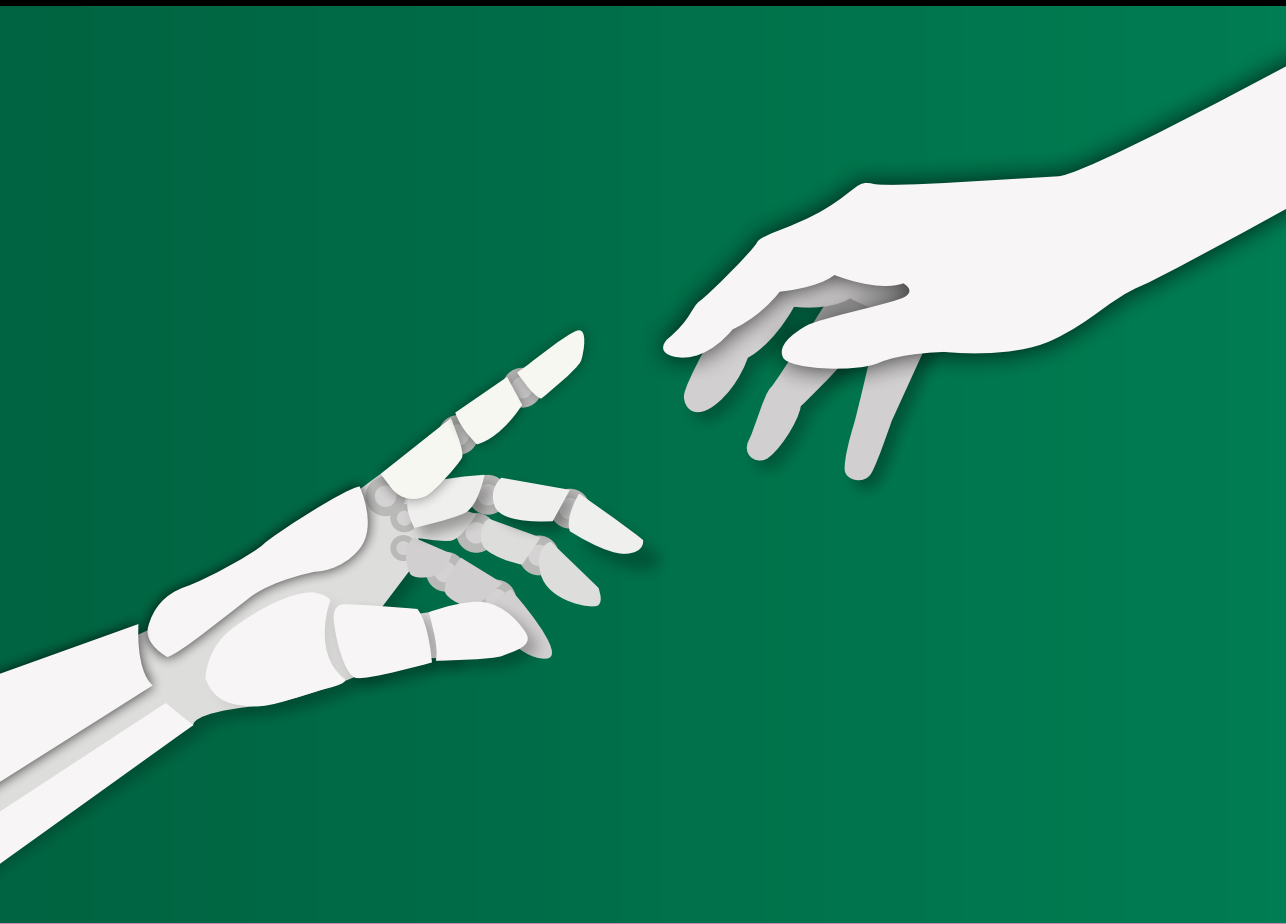
HUMINFRA HANDBOOK

Empowering digital and experimental humanities



Edited by: Gerlof Bouma, Dana Dannélls, Dimitrios Kokkinakis, Elena Volodina
University of Tartu Library

Huminfra handbook

Huminfra handbook

Empowering digital and experimental humanities

edited by
Gerlof Bouma, Dana Dannélls
Dimitrios Kokkinakis & Elena Volodina

HUM
INFRA

We gratefully acknowledge the support of the Swedish Research Council, grant numbers 2021-00176 and 2023-00171 *Huminfra*.

Gerlof Bouma, Dana Dannélls, Dimitrios Kokkinakis
& Elena Volodina (editors)
Huminfra handbook: Empowering digital and experimental humanities
URI: https://hdl.handle.net/10062/117324 (collection; online)
ISBN: 978-99-0853-612-5 (online)
ISBN: 978-91-531-7077-8 (print)

Number 59 in NEALT Proceedings Series
ISSN 1736-6305 (online)
ISSN 1736-8197 (print)

Northern European Association for Language Technology
Online version published by University of Tartu Library

© 2025 The authors (individual papers); the editors (collection)
This collection and the individual papers in it are licensed under a Creative Commons Attribution 4.0 International license. License details: https://creativecommons.org/licenses/by/4.0/

Cover design by Linnéa Tjernström
Document prepared with LaTeX 2$_\varepsilon$ and typeset with LuaTeX by Gerlof Bouma
Paper version printed by Media-Tryck, Lund

Contents

Preface · **ix**

Marianne Gullberg, Coppélie Cocq, Mats Fridlund & Koraljka Golub
1 Huminfra – a Swedish national infrastructure to support research in digital and experimental humanities · **1**

Infrastructure units, labs and centers

Chris Haffenden & Justyna Sikora
2 Doing digital research at KBLab: A practical introduction to using the National Library of Sweden's data lab · **17**

Daniel Brodén, Mats Fridlund, Cecilia Lindhé & Jonathan Westin
3 Interdisciplinary digital project design · **57**

Gijs Aangenendt, Maria Skeppstedt & Karl Berglund
4 Applied NLP for humanities research · **77**

Emma Sköldberg & Kristian Blensenius
5 Empirisk ordforskning som grund för vidareutveckling av *Svensk ordbok utgiven av Svenska Akademien* · **113**

Tools, methods and resources

Maria Skeppstedt, Magnus Ahltorp, Kostiantyn Kucher, Gijs Aangenendt, Matts Lindström & Ylva Söderfeldt
6 The Word Rain visualisation technique applied to digital history: How to visualise, explore and compare texts using semantically structured word clouds · **147**

Jonas Svensson
7 Navigating Swedish Salafism
Large language model-augmented content detection and topic modeling using BERTopic with YouTube metadata · **183**

Dalia Ortiz Pablo, Sushruth Badri, Gijs Aangenendt, Mo von Bychelberg & Matts Lindström
8 A machine learning pipeline for digitalising historical printed materials – from data collection to a searchable database · **207**

Elena Volodina, Arianna Masciolini, Beáta Megyesi, Julia Prentice, Lisa Rudebeck, Gunlög Sundberg & Mats Wirén
9 SweLL with pride: How to put a learner corpus to good use · **251**

Beáta Megyesi & Rex Ruan
10 From text to insight: Uncovering linguistic patterns with SWEGRAM · **307**

Therese Lindström Tiedemann, David Alfter & Elena Volodina
11 A practical guide to the Swedish L2 lexical profile · **355**

Dana Dannélls, Robin Kurtz, Erik Lenas & Viktoria Löfgren
12 Unlocking Swedish historical material through OCR and HTR · **405**

Magnus Ahltorp & Harald Berthelsen
13 Tugg: A transcription tool for language documentation · **433**

Arianna Masciolini, Herbert Lange & Márton András Tóth
14 Exploring parallel corpora with STUnD: A Search Tool for Universal Dependencies · **455**

Daniel Ihrmark, Hanna Carlsson & Fredrik Hanell
15 Low-code web scraping and text analysis with Octoparse and KNIME: An example from the CICuW project · **505**

Dimitrios Kokkinakis & Gerlof Bouma
16 Exploratory Swedish text analysis using notebooks – a *smörgåsbord* of basic corpus linguistic insights · **541**

Preface

This handbook is a result of the first Huminfra Conference (HiC 2024), which was held in January 2024 at the University of Gothenburg and organized in the context of *Huminfra*. As the central website huminfra.se explains, "Huminfra is a Swedish national infrastructure supporting digital and experimental research in the Humanities by providing users with a single entry point for finding existing Swedish materials and research tools, and by developing national method courses."

Perhaps not surprisingly, artificial intelligence (AI) – a branch of computer science devoted to developing systems capable of performing tasks that typically require human intelligence, such as learning, reasoning, and problem-solving – and its subfield natural language processing (NLP) – focused on understanding, interpreting, and generating human language – play increasingly significant roles in approaches relevant to the humanities. Both AI and NLP rely heavily on advances in machine learning and the availability of large-scale digital data. In recent years, these fields have undergone profound transformations, marked by shifting priorities, innovative techniques, and expanding applications. These changes are reflected in the evolving methods and resources shaping contemporary humanities research. Given their broad scope, AI and NLP need to be situated within specific disciplinary contexts to yield meaningful and practical outcomes. When these technologies are applied to and integrated with the humanities – drawing on digital tools, computational methods, and large datasets – they give rise to what is commonly referred to as the digital humanities and, when additionally combined with empirical methods from the behavioural sciences, experimental humanities.

The chapters in this handbook contribute to Huminfra's mission to strengthen and advance the field of digital and experimental humanities in Sweden by showcasing what existing digital tools, research methodologies and innovative infrastructural components can offer to the field through examples and step-by-step guidelines.

Organization of the volume

The volume begins with a chapter introducing the Huminfra infrastructure itself, and describing its context, aims and organization. The rest of the volume is organized in two parts.

The first part, *Infrastructure units, labs and centers*, provides an overview of four units that operate in the area of humanities infrastructure in Sweden:

KBLab at the National Library of Sweden, the Gothenburg Research Infrastructure for Digital Humanities (GRIDH), the Centre for Digital Humanities and Social Sciences (CDHU) at Uppsala University, and the lexicography unit at Språkbanken Text (SBX). They represent coordinated ecosystems of digital tools, services, data, and expertise designed to enable and enhance scientific inquiry, and provide researchers with the means to store, access, analyse, and share data across disciplines and national borders, thereby fostering open science, innovation, and data-driven research practices.

Building on technological developments in AI and NLP, the chapters collected in the second part of this volume introduce a diverse array of *Tools, methods and resources*, accompanied by practical guidelines and, where relevant, code implementations. Many contributions also feature real-world case studies that demonstrate how such approaches can be applied in humanities research. Together, the tools and case studies aim to advance research and methodology across disciplines such as history, linguistics, literature studies, and religious studies. The approaches presented span a wide range of techniques – from topic modelling, optical character recognition (OCR), and handwritten text recognition (HTR) to the use of computational notebooks – and include tools for targeted corpus searches, linguistic analysis, and data visualization. They further encompass methods for processing multiple data modalities, including audio, images, and video, as well as key digital resources vital for experimental and NLP-based research, such as learner corpora.

Taken together, while the primary focus of this volume is on Swedish researchers and the Swedish language, many of the presented approaches are language-independent, and are readily transferable to other domains. Likewise, language-specific approaches may serve as models or inspiration for adaptation to other linguistic and cultural contexts.

A note on the compilation process

HiC 2024 took place at the University of Gothenburg at the beginning of 2024, and was followed by a call for handbook chapter proposals in the spring of that year. The final selection of 16 contributions presented in this volume was made on the basis of both external reviews by peers and internal reviews by the editors, of the full chapters.

Acknowledgments

We thank the chapter authors for their contributions and their collaboration during the editing of this volume. We thank the external reviewers, viz. Magnus Ahltorp, David Alfter, Aleksandrs Berdicevskis, Eva-Marie Bloom

Ström, Lars Borin, Daniel Brodén, Andrew Caines, Mats Fridlund, Koraljka Golub, Marianne Gullberg, Chris Haffenden, Nikolai Ilinykh, Therese Lindström Tiedemann, Eva Pettersson, Maria Skeppstedt, Sara Stymne, Henricus Verhagen, for their invaluable assistance in assuring the quality of this volume.

We thank NEALT editor-in-chief Marcel Bollmann for including this volume in the NEALT Proceedings Series and for his help during the publication process. We are grateful to Marianne Gullberg, Anna Blåder, and the other members of "Huminfra HQ" for their support and advice, and to Linnéa Tjernström for designing the cover.

We gratefully acknowledge the support of the Swedish Research Council, grant numbers 2021-00176 and 2023-00171, *Huminfra*, as well as the support of the Swedish national research infrastructure *Språkbanken* – jointly funded by its 10 partner institutions and the Swedish Research Council (grant numbers 2017-00626 and 2023-00161).

Elena Volodina, Gerlof Bouma,
Dana Dannélls and Dimitrios Kokkinakis
Gothenburg, October 2025

1 Huminfra – a Swedish national infrastructure to support research in digital and experimental humanities

Marianne Gullberg
Lund University

Coppélie Cocq
Umeå University

Mats Fridlund
University of Gothenburg

Koraljka Golub
Linnaeus University

1 *Introduction*

Huminfra is a Swedish national infrastructure supporting research in digital and experimental humanities. Research in the humanities is increasingly interdisciplinary, quantitative, and multi-methodological, drawing on digital resources, tools, and technology. Huminfra's establishment and development is timely given the current emergent focus on research infrastructures within the Swedish research landscape, and especially on digital and experimental infrastructures (cf. the research and innovation bill *Forskning och innovation för framtid, nyfikenhet och nytta, Prop 2024/25:60* 2024).

Huminfra (https://www.huminfra.se/), a distributed national infrastructure, brings together 12 nodes in the domains of digital and experimental humanities. Its scope includes digital materials, tools, and critical interpretative perspectives, as well as experimental, quantitative approaches (sensor-based data, real-time analyses). Huminfra coordinates and provides a single entry-point to available Swedish materials, research tools, and

HUM INFRA Marianne Gullberg, Coppélie Cocq, Mats Fridlund & Koraljka Golub. 2025. Huminfra – a Swedish national infrastructure to support research in digital and experimental humanities. In Gerlof Bouma, Dana Dannélls, Dimitrios Kokkinakis & Elena Volodina (eds.), *Huminfra handbook: Empowering digital and experimental humanities* (NEALT Proceedings Series 59), 1–13. University of Tartu Library. DOI: 10.58009/aere-perennius0170 © The authors, CC BY 4.0

expertise via huminfra.se to facilitate discovery of and access to existing Swedish resources. It also hosts national training events in methods related to the resources. As of 2024 it hosts DARIAH-SE, the Swedish national node in the pan-European research infrastructure consortium Digital Research Infrastructure for the Arts and Humanities (DARIAH ERIC). DARIAH-SE provides Swedish and international scholars with mutual access to resources, tools, and training. From 2025 onwards, Huminfra will also develop a small set of national tools based on user needs. Huminfra strengthens research in areas of specific interest to Sweden (cultural heritage, languages spoken in Sweden, national archives) to promote Swedish humanities and ensure its international competitive edge.

In the following, we describe Huminfra and outline the background for its establishment. The chapter serves as a backdrop to the other contributions in this volume which exemplify resources and expertise found in Huminfra and give practical guidelines to a range of infrastructural components in the humanities.

2 *What is Huminfra and why?*

Some of the biggest challenges in the future for academic research and society in general include issues of social equality, diversity, health, migration, climate change, consciousness, and effects of artificial intelligence (AI). Many of these topics revolve around the impact on or change of human behaviours, communication, culture, education, cognition, and art (cf. United Nations 2015). Research challlenges include: What effect will open AI tools have on society and knowledge production? How do social media affect the way individuals and groups think, produce, perceive, and process information? How do environmental factors interact with individual cognition and socio-economic conditions to affect decision-making and changes in behaviours (e.g., becoming a vegetarian to minimise climate change)? How do we access, interpret, preserve and present cultural heritage data? How and when are linguistic, communicative, and cultural skills learned and transmitted in an age of migration? These challenges are both scientific and societal.

The humanities have traditionally tackled such questions qualitatively, but research in the humanities (including the arts, and social sciences) has become more quantitative, interdisciplinary, and multi-methodological. It is increasingly based on complex data sets, both born digital and newly digitised. An increasing number of researchers in the humanities are no longer strangers to statistics, algorithms and programming. They now use state-of-the-art digital and experimental methods that enable rapid and

collaborative analyses of complex data sets. With enriched and expanded tool kits and interdisciplinary linkages, researchers in the humanities now play a key role in contributing human-centred perspectives on the big scientific and societal challenges and furthering and deepening our understanding of the human condition.

However, although methodological resources are widely available, many researchers in the humanities struggle to find and access the research infrastructures they need (cf., Riksbankens Jubileumsfond 2019; Vetenskapsrådet 2019). Research infrastructures can be defined as organised facilities designed to support research by providing access to equipment, data sets and materials, databases, digital platforms, and networks. They should also provide computational resources and e-scientific/digital tools to process and analyse data and materials. Equally vital is access to the appropriate expertise (e.g., programming skills and methodological know-how) and to training in the relevant techniques including developing collaborative modes of working (Brodén et al. 2024, Foka et al. 2018, Golub et al. 2020, Niehorster et al. 2024). Hitherto, the infrastructural landscape in Sweden has been fragmented, making it difficult for researchers to identify, select, and gain access to the most appropriate and useful resources. The recent rise of many new (national) infrastructures has made the research landscape even richer, but also more difficult to navigate. This state of affairs may lead to missed research opportunities and a risk that Swedish scholarship loses international ground.

To remedy the fragmentation of the landscape, and to support and help internationalise Swedish research in the humanities, Huminfra was established in 2022 as a distributed consortium, funded by the Swedish Research Council (SRC) and by the 12 participating universities and organisations, labelled nodes. The Huminfra nodes have complementary resources and expertise. This includes digital humanities, defined as work focusing on e-scientific/digital materials, tools and critical interpretative skills and perspectives, but also experimental humanities, characterised by experimental and quantitative approaches, the use of sensor-based data, and sophisticated real-time data analyses. Finally, Huminfra supports research in knowledge areas of specific interest to Sweden such as aspects of Swedish cultural heritage, research on languages spoken in Sweden, and national archives and data sets.

During the establishment phase 2022-2024 Huminfra had three aims:

1. to coordinate and provide a single entry-point to available Swedish materials, research tools, and expertise at the consortium nodes via a web-based information platform, https://www.huminfra.se (Figure 1).

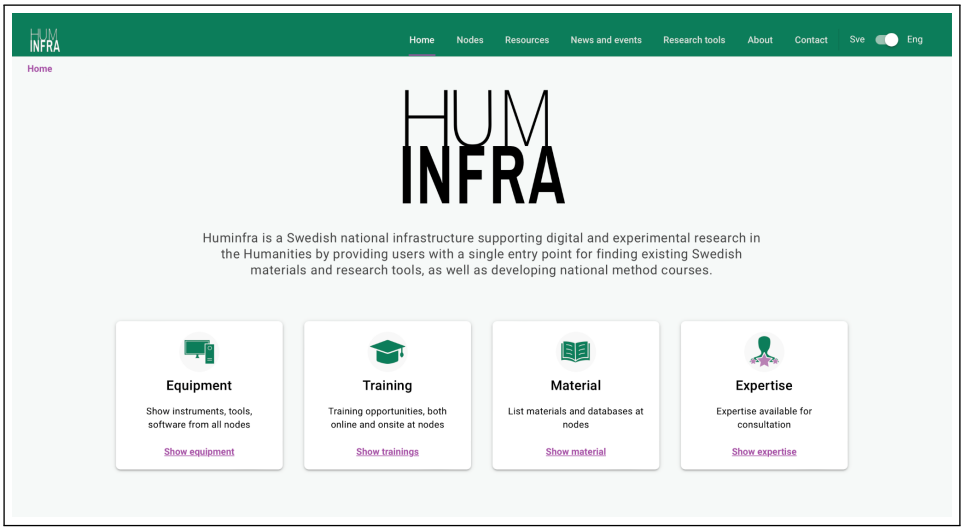


Figure 1: The web-based information platform Huminfra.se

Huminfra.se enables the rapid matching of needs to resources and facilitates access to local resources in accordance with local guidelines and conditions. Furthermore, Huminfra.se also links to other relevant local, national, and international infrastructures (e.g., *Litteraturbanken*, the Swedish National Data Service (SND), etc.).

2. to provide training in methods related to the resources at hand via jointly developed national Huminfra-courses, tutorials, workshops, etc. These training activities along with methodological advice at the local nodes constitute Huminfra's user support. They also serve as a forum for monitoring new needs for tools and training not yet available.

3. to establish and host Swedish membership in the Digital Research Infrastructure for the Arts and Humanities (https://www.dariah.eu/), the European Research Infrastructure Consortium (ERIC) that connects Swedish research to European resources, enhancing collaboration options, tool development and training within the domain of cultural heritage to match the existing networks available in the language domain via the European infrastructure Common Language Resources and Technology Infrastructure (https://www.clarin.eu/) and its Swedish node Språkbanken CLARIN. In this way Huminfra.se points Swedish researchers to European resources, and guides European researchers towards Swedish ones.

In the consolidation phase 2025-2028, Huminfra adds a fourth aim:

4. to jointly develop a small set of national tools arising from a survey of needs led by DARIAH-SE and from the specific needs and critical issues expressed and identified during the national training events. These will be freely available and wherever possible conform to standards and principles of interoperability as well as the implementation of FAIR data principles (findability, accessibility, interoperability, and reuse of digital assets), including those of DARIAH for cultural heritage data and CLARIN for language data. The tool development transforms Huminfra from the meta-infrastructure it was in the establishment phase to a more explicit methods and resource infrastructure in the consolidation phase.

3 *Huminfra for whom?*

Huminfra's user base is potentially all researchers in the humanities and professionals at cultural heritage institutions in Sweden, as well as international humanities scholars looking specifically for Swedish materials and resources. Other areas also draw on Huminfra's resources. Computer science and associated disciplines are finding new applications in the development of technical solutions, e.g., developing interfaces and tools, for research in these domains, not least through the recent emphasis on the use of AI in the humanities and social sciences (e.g., Wallenberg AI, Autonomous Systems and Software Program – Humanity and Society, https://wasp-hs.org/). Requests for expertise on text mining, topic modelling, automated transcription of recordings, making handwritten manuscripts machine readable, automatic dialect identifiers, etc., has increased dramatically and continues to grow. Large Language Models exploded onto the scene in 2022, and the KBLab node at the National Library of Sweden and the Swedish National Archives have both drawn international attention to their work in this area, again attracting new potential users of the technology and the materials housed at these nodes. The user base spans all disciplines and areas of science, as well as other societal stakeholders in the public and private sectors such as health, education, law enforcement , and government agencies. Cultural heritage professionals are another major user base for DARIAH ERIC/SE. The recent national emphasis on digitisation of cultural heritage data means that 3D scanning, visualisation, and corpus related activities will increase considerably in volume. Moreover, the creative industries are expected to grow as a significant new user base over the coming years. Finally, an important outcome of the Huminfra consortium is the Swedish

National Research School in digital humanities: Data, Culture, and Society – Critical Perspectives (DASH). This is the first national Swedish graduate school in digital humanities, which also feeds back into Huminfra's user base by introducing early career researchers to Huminfra as a valuable resource.

4 *What does Huminfra (and DARIAH-SE) add?*

Swedish researchers in the humanities have long worked in a landscape where infrastructural needs at local universities are only partially met, and where the knowledge about the location of resources and how to use them often exists in so-called silos. Knowledge about resources is often shared *within* disciplines, whereas cross-disciplinary knowledge of methods and tools is less widespread.

The added value of Huminfra as a national infrastructure is thus fourfold. First, it meets the growing needs for long-term-provision and maintenance of technology and digital/e-scientific resources for the humanities, including tool development, expertise on management of (especially human) e-data according to the FAIR principles, and adherence to common principles and standards for open science (cf. work by the European Open Science Cloud (EOSC) and the European Strategy Forum on Research Infrastructures (ESFRI)). In the language domain, the consortium CLARIN ERIC, of which Sweden is a member state, develops language resources and tools according to general principles. In other areas of the humanities and the arts and in the domain of cultural heritage data, DARIAH ERIC offers platforms for training materials, tools and data thereby increasing exchange with the international arena. DARIAH-SE provides Swedish researchers in this domain with an entry point for digital content, tools, services, facilities, and expertise.

Second, Huminfra also helps to make the most of human resources within the experimental and digital humanities, in particular through its dedicated research competent staff by developing general tools and methods, providing training and guidance, maintaining and upgrading hard- and software, and acting as liaison officers between users across disciplines in the humanities and beyond, and with stakeholders in education, industry, cultural and creative institutions, and with other data providers.

Third, Huminfra safeguards a sustainable future for this work by pooling resources and coordinating training, thus securing recruitment and reducing vulnerability.

Fourth, it is vital to coordinate this work at a national level in order to optimise resource exploitation and avoid duplication.

In sum, Huminfra makes Swedish humanities more visible and inter-

nationally competitive. For all four added values, Swedish membership in DARIAH ERIC is an important component. Through DARIAH-SE Swedish researchers in digital arts and humanities gain access to digital content, tools, services, facilities, and expertise in Sweden, in Europe, and beyond. In turn, European and international researchers are able to develop new research, tools and services on the basis of uniquely Swedish resources and integrate them with other data sets. This ensures the visibility of Swedish resources and enhances research related to areas of study in Sweden. Through Huminfra, DARIAH-SE has direct access to data, tools, and content providers via the nodes in the infrastructure.

5 *Huminfra and other infrastructures*

Huminfra has drawn inspiration from international models, for instance DIGHUMLAB Denmark (https://dighumlab.org), a national, distributed research infrastructure that integrated and promoted tools, digital resources, communities, and opportunities to Danish researchers. At the international level, Huminfra is linked to and participates in CLARIN ERIC, and as of July 2024 Huminfra hosts DARIAH-SE, the Swedish node of DARIAH ERIC. Moreover, a Scientific Advisory Board composed of international distinguished scholars with long experience in infrastructure development contributes to profiling and strengthening Huminfra's relevance in an international context.

Nationally, Huminfra works closely with a number of existing infrastructures. Traditional key structures for the humanities, such as university libraries, the National Library of Sweden, and the Swedish National Archives, aim to facilitate preservation and access to high-quality research data. The libraries are also important digital data providers and curators, and, to an extent, experts on data management and archiving. The KBLab at the National Library of Sweden, the Swedish National Archives, and the Swedish School of Library and Information Science are members of the Huminfra consortium and they have all made considerable headway towards meeting new research needs (cf. Kungliga Biblioteket 2020). Moreover, although not part of the consortium, the Swedish National Data Service (SND), custodian and curator of digital research data and the national resource for the accessibility, preservation, and reuse of data, is an important conversational partner to Huminfra.

Another relevant structure is the national infrastructure *Språkbanken* which hosts and develops language technology and resources (e.g., Borin et al. 2022; Edlund et al. 2022). It comprises *Språkbanken Text*, *Språkbanken*

Tal, *Språkbanken Sam*, and *Språkbanken CLARIN* (the Swedish membership organisation in CLARIN ERIC), coordinated by the University of Gothenburg. *Språkbanken Tal* and *Språkbanken Text* are members of the Huminfra consortium to ensure smooth communication and interaction between the infrastructures. Their inclusion in Huminfra importantly makes access to expertise on language and texts available to non-language disciplines, for example through joint events. Such complementarity is an important gain of the interdisciplinary composition of Huminfra's nodes.

Huminfra also collaborates with the National Research Infrastructure for Data Visualization (InfraVis) and the Swedish National Infrastructure for Digital Archaeology (SweDigArch). InfraVis provides resources connected to scientific visualisation, covering needs from a wide range of disciplines including the humanities. Several nodes in Huminfra are part of both InfraVis and SweDigArch. The connections between the infrastructures enhance interdisciplinary collaborations and knowledge sharing.

The National Academic Infrastructure for Supercomputing in Sweden (NAISS), based at Linköping University, provides computing resources for large-scale computation operations and storage in all disciplines, including user support from 2023. Its predecessor, Swedish National Infrastructure for Computing (SNIC) hosted only a dozen projects from the humanities, but the potential for growth is obvious and Huminfra will promote such growth through its user base.

In sum, existing infrastructures support very specific facets of the infrastructural needs in the relevant disciplines, and Huminfra helps coordinate and strengthen links between them to facilitate knowledge transfer and access to Swedish scholarship.

6 *The societal relevance of Huminfra*

Huminfra.se, which provides information on resources and research output, heightens the visibility and impact of humanities research both within and outside of Academia. The platform acts as an interface to external stakeholders such as cultural institutions, education, health, industry, and public and creative sectors. Researchers using Huminfra resources have the potential to fundamentally change research on how human activities (communication, culture, education, and cognition) evolve as a result of factors such as migration, diversity, climate change, social equality or health, highlighted in the UN's global sustainability goals. This, in turn, could pave the way for new Swedish innovations in many arenas, such as AI applications for (inclusive) communication in written or spoken form (cf. KBLab's develop-

ment of Swedish Large Language Models), new strategies on how to nudge individuals to act in more environmentally sound ways, tools for learning languages, new ways to execute and distribute performing arts to wider audiences, etc.

Huminfra has a dedicated programme of activities targeting nonacademic stakeholders. Several nodes organise joint MA theses with industry (KTH Royal Institute of Technology, Linnaeus University, Lund University), invite external stakeholders as speakers at workshops (University of Gothenburg, KTH Royal Institute of Technology, Linnaeus University, Lund University), initiate collaborations with the arts (University of Gothenburg, Lund University, Umeå University), etc. Commissioned education is another route for interacting with non-academic users (e.g., Lund University). Huminfra also has a programme for outreach proper, for example via outreach webinars targeting non-academic potential users. Huminfra outreach also includes activities in social media with the aim of initiating new collaborations.

7 *The long-term impact of Huminfra*

Huminfra aims to contribute to the long-term accumulation of national competence in digital and experimental humanities in at least four ways.

1. huminfra.se contributes to visibility and dissemination of information on existing resources, training activities, and provides examples of research output (for academic and popular science purposes). One such effort was Huminfra's first national research conference, HiC 2024, which gathered researchers, developers, and users of research infrastructures in digital and experimental humanities (cf. Volodina et al. 2024). Through such knowledge-sharing activities Huminfra contributes to enhanced opportunities for Swedish researchers to find appropriate tools, data sets, training opportunities, and potential collaborators. The Swedish membership in DARIAH further guarantees greater international contacts and international impact of Swedish scholarship and resources, which feeds back to accumulated national competence.

2. Through the national coordination of training activities, Huminfra empowers Swedish scholarship by contributing to the growth of a new generation of humanities scholars, well trained in the techniques needed to tackle new research questions, with skills to make them independent agents in designing and conducting their own studies and not dependent on 'hired help'.

3. National coordination and development of joint tools also means that knowledge, expertise, and resources will become more distributed and are likely to accumulate not only in a few locations, but be disseminated across the entire national arena.

4. Huminfra is a long-term commitment. Huminfra is likely to generate new collaborative research in many domains through the increased visibility and facilitated access to resources, through the training initiatives, and through collaborative method and tool development. As with all infrastructures, sustained maintenance (and funding) is vital to competitive research to avoid resource destruction, jeopardising research initiatives.

8 *Conclusion*

Huminfra.se gives Swedish researchers in the humanities a single entry-point to highlight and advertise Swedish research in relevant fields. It also points to resources and actors outside the nodes participating in Huminfra, which strengthens the visibility and competitive edge of Swedish research and research resources more generally. Moreover, considering the key role of digital tools and modes of communication in research communication today, and Huminfra's ambition to equip researchers with necessary skills in order to handle those, we foresee that Huminfra will contribute to increase the sharing and dissemination of humanities' research results in society.

Further, the establishment of DARIAH-SE and its connections to the DARIAH ERIC will also ensure maximal international visibility of Swedish humanities research. Importantly, Huminfra's sustained dedication to outreach with activities in social media, outreach conferences, and popular science will highlight research to popular audiences, which will generate further potential for new collaborations with non-academic stakeholders, and innovations.

Huminfra is a first major national effort towards equipping Swedish humanities with the skills and tools needed to meet the future research challenges in a research landscape that is increasingly interdisciplinary. The chapters in this volume provide good examples of the kind of expertise and tools scholars now need. They provide valuable practical guidelines but also more general insights on interdisciplinary humanities research and infrastructures. To ensure the sustained momentum of Huminfra, its long-term impact, and continued growth requires foresight, vision, and perseverance. The Huminfra consortium looks forward to the work ahead.

References

Borin, Lars, Rickard Domeij, Jens Edlund & Markus Forsberg. 2022. *European language equality: D1. 33: Report on the Swedish language*. https://european-language-equality.eu/wp-content/uploads/2022/05/ELE___Deliverable_D1_33__Language_Report_Swedish_.pdf.

Brodén, Daniel, Mats Fridlund, Leif-Jöran Olsson, Magnus P. Ängsal & Patrik Öhberg. 2024. Samförfattande som datadriven tvärvetenskap: Pragmatiska lärdomar från SweTerror-projektet. *Proceedings of the Huminfra Conference* (*HiC 2024*) 205. 148–153. DOI: 10.3384/ecp205020.

Edlund, Jens, Daniel Brodén, Mats Fridlund, Cecilia Lindhé, Leif-Jöran Olsson, Magnus P. Ängsal & Patrik Öhberg. 2022. A multimodal digital humanities study of terrorism in Swedish politics: An interdisciplinary mixed methods project on the configuration of terrorism in parliamentary debates, legislation, and policy networks 1968–2018. In Kohei Arai (ed.), *Proceedings of the 2021 intelligent systems conference* (*IntelliSys*), vol. 2, 435–449. Springer. DOI: 10.1007/978-3-030-82196-8_32.

Foka, Anna, Anna Misharina, Viktor Arvidsson & Stefan Gelfgren. 2018. Beyond humanities qua digital: Spatial and material development for digital research infrastructures in HumlabX. *Digital Scholarship in the Humanities* 33(2). 264–278. DOI: 10.1093/llc/fqx008.

Forskning och innovation för framtid, nyfikenhet och nytta, Prop 2024/25:60. 2024. https://data.riksdagen.se/fil/956CAAE7-7DC5-48E5-B7EA-08BA38553A77.

Golub, Koraljka, Elisabet Göransson, Anna Foka & Isto Huvila. 2020. Digital humanities in Sweden and its infrastructure: Status quo and the sine qua non. *Digital Scholarship in the Humanities* 35(3). 547–556. DOI: 10.1093/llc/fqz042.

Kungliga Biblioteket. 2020. *Kungliga biblioteket och forskningsbiblioteken i forskningens infrastruktur: Underlag till utredningen Organisation, styrning och finansiering av forskningsinfrastruktur, Dir. 2020:52*. https://www.kb.se/download/18.6057ac19175f71f980c14e9/1609144338873/KB-och-forskningsbiblioteken-i-forskningens%20infrastruktur.pdf.

Niehorster, Diederick C., Marianne Gullberg & Marcus Nyström. 2024. Behavioral science labs: How to solve the multi-user problem. *Behavior Research Methods* 56(8). 8238–8258. DOI: 10.3758/s13428-024-02467-4.

Riksbankens Jubileumsfond. 2019. *RJ rapporterar 2019:1. Humanistisk infrastruktur*. https://rj.se/globalassets/rapporter/rj%5C_wrapport%5C_2019-1%5C_190508.pdf.

United Nations. 2015. *Transforming our world: The 2030 agenda for sustainable development. A/RES/70/1.* https://sustainabledevelopment.un.org/post2015/transformingourworld.

Vetenskapsrådet. 2019. *Research overview 2019: Humanities and social sciences*. https://www.vr.se/english/analysis/reports/our-reports/2019-05-10-research-overview-2019-humanities-and-social-sciences.html.

Volodina, Elena, Gerlof Bouma, Markus Forsberg, Dimitrios Kokkinakis, David Alfter, Mats Fridlund, Christian Horn, Lars Ahrenberg & Anna Blåder (eds.). 2024. *Proceedings of the Huminfra Conference (HiC 2024)* (Linköping Electronic Conference Proceedings 205). Linköping University: Linköping University Electronic Press. DOI: 10.3384/ecp205.

List of abbreviations

CLARIN	Common Language Resources and Technology Infrastructure
DARIAH	Digital Research Infrastructure for the Arts and Humanities
DASH	Swedish national research school in digital humanities: Data, culture, and society – critical perspectives
EOSC	European Open Science Cloud
ERIC	European Research Infrastructure Consortium
ESFRI	European Strategy Forum on Research Infrastructures
FAIR	principles Findability, Accessibility, Interoperability, and Reuse of digital assets
InfraVis	National Research Infrastructure for Data Visualization
NAISS	National Academic Infrastructure for Supercomputing in Sweden
SND	Swedish National Data Service
SNIC	Swedish National Infrastructure for Computing
SRC	Swedish Research Council
SweDigArch	Swedish National Infrastructure for Digital Archaeology
WASP-HS	Wallenberg AI, Autonomous Systems and Software Program – Humanity and Society

Corresponding authors

Marianne Gullberg
Center for Languages and Literature
Lund University
marianne.gullberg@ling.lu.se

Coppélie Cocq
Humlab
Umeå University
coppelie.cocq@umu.se

Mats Fridlund
Gothenburg Research
Infrastructure in Digital
Humanities
University of Gothenburg
mats.fridlund@lir.gu.se

Koraljka Golub
Department of Cultural Sciences
Linnæus University
koraljka.golub@lnu.se

Infrastructure units, labs and centers

2 Doing digital research at KBLab: A practical introduction to using the National Library of Sweden's data lab

Chris Haffenden
National Library of Sweden

Justyna Sikora
National Library of Sweden

The emergence of digital heritage data and the rapid development of new AI tools for computational analysis are transforming GLAM institutions, particularly in the design of digital research infrastructure. Researchers in the digital humanities and social sciences increasingly expect to access collections at unprecedented scales. This chapter addresses such expectations by providing a hands-on guide to KBLab, the data lab at the National Library of Sweden (KB). It outlines the lab's resources, including access to KB's digitized collections and AI models like KB-BERT, and showcases innovative development projects like *Bildsök*, which makes visual archives more accessible. The chapter also details the steps to initiate research collaborations and discusses best practices for utilizing KBLab's tools effectively. By bridging technical insights with practical applications, it serves as a comprehensive starting point for conducting large-scale digital research at KB and beyond.

1 *Introduction: Library labs as digital research infrastructure*

The emergence of digital cultural heritage has significantly expanded the scope and possibilities of humanities research. With the initiation of mass digitization programs and new collecting practices for born digital material in the GLAM sector—i.e. **g**alleries, **l**ibraries, **a**rchives and **m**useums—the range and volume of heritage data potentially available to researchers has increased exponentially over the past few decades (Bingham and Byrne 2021). In tandem with this digital expansion, recent developments in Machine Learning and Artificial Intelligence (hereafter ML and AI) have enhanced

HUM INFRA Chris Haffenden & Justyna Sikora. 2025. Doing digital research at KBLab: A practical introduction to using the National Library of Sweden's data lab. In Gerlof Bouma, Dana Dannélls, Dimitrios Kokkinakis & Elena Volodina (eds.), *Huminfra handbook: Empowering digital and experimental humanities* (NEALT Proceedings Series 59), 17–56. University of Tartu Library. DOI: 10.58009/aere-perennius0171 © The authors, CC BY 4.0

the tools available for computational analysis of these materials. Researchers in the digital humanities and computational social sciences now expect access to GLAM collections at unprecedented scales. Rather than focusing upon the individual object, i.e. a book or newspaper, these large-scale approaches routinely focus upon thousands—or even millions—of documents to uncover new patterns or trends (e.g. Underwood 2019, Hurtado Bodell et al. 2024).

Such expectations have presented GLAM institutions with significant infrastructural challenges, not least national and research libraries with rapidly expanding electronic legal deposit collections. How can these organizations ensure that their "collections as data" are accessible to researchers (Padilla 2025)? A widespread response has been the establishment of GLAM Labs, which function as interdisciplinary spaces—both physical and digital—where data scientists and collections specialists can collaborate to address the complexities of digital heritage (Mahey et al. 2019). In particular, the Library Lab has become an established feature of the information and research landscape, with examples ranging from the US Library of Congress' LC Labs to the BNElab at the National Library of Spain. As Ryan Cordell has highlighted, these labs have provided opportunities for libraries to experiment with ML to enrich collections, "making them more useable for scholars, students, and the general public" (Cordell 2020: 1). Library Labs are thus utilizing AI to enhance the library as a digital research infrastructure (Börjeson et al. 2024).

This chapter focuses upon the research possibilities enabled by KBLab, the data lab at the National Library of Sweden (*Kungliga biblioteket*, hereafter KB). Established in 2019, KBLab initially aimed to:

1. offer access to the library's digital collections in structured form for large-scale research; and
2. experiment with AI tools to innovate the library's working practices.

The emergence of transformer-based language models (Devlin et al. 2019) soon expanded the lab's mission, enabling it to contribute to an improved national AI infrastructure for the Swedish language (Haffenden, Fano, et al. 2023: 44–45). By leveraging KB's high-quality, language-specific collections, KBLab has trained novel AI models such as KB-BERT. These models enable the analysis of Swedish text, sound, and image data, significantly enhancing the library's role as a research infrastructure. This ensures a productive symbiosis between the lab's functions: digital heritage enables the development of better Swedish AI tools, which in turn enhance the accessibility and searchability of KB's collections. KBLab thus constitutes an important entry point for researchers seeking to analyze and generate new insights about Swedish heritage data.

The sort of infrastructural work carried out at the lab creates a distinctive position for KBLab in the broader ecology of Swedish support for digital research encapsulated by Huminfra. Whereas the various nodes for digital humanities based at universities tend to work very closely with research projects, often including research engineers for ongoing hands-on support, our emphasis on producing new Open Access AI models for Swedish material necessitates a more arm's length approach towards collaborating with such projects. Insofar as we concentrate on providing access to the collections and using these as the basis for collections-based AI development (Börjeson et al. 2024: 572–576), KBLab most closely resembles the other GLAM lab within Huminfra, the AI lab at the National Archives of Sweden.[1] While the National Archives' lab has prioritized models for handwritten text recognition (HTR) to improve access to archival material, KBLab's focus on text, sound and images reflects the multimodal nature of KB's collections.

This chapter offers a practical introduction to accessing and using KBLab's various resources. We provide an orientation for digital researchers interested in making use of these in their projects, whether it be using the lab as a physical site to access the library's collections or deploying our AI models as research tools to analyze data. While these models for computational research are best utilized by scholars with some prior knowledge of programming and data analysis, our overview also aims to inspire those with more qualitative backgrounds to explore new research opportunities. The first part of the chapter explains what data is available at KBLab and how it can be accessed in the lab environment. The second part describes the various AI models we have released, offering specific use cases and examples of how these might be used to conduct digital research, as well as what existing projects at the lab have engaged with. The final part explains the application process for initiating a project based at the lab. Our aim in producing such a hands-on guide is to give sufficient context and knowledge to understand the possibilities of KBLab. By bridging technical insights with practical applications, the chapter serves as a comprehensive starting point for conducting large-scale digital research at KB and beyond.

2 *What data is available at KBLab? Identifying and preparing research data*

KB's collections range from books and manuscripts to television programmes and radio broadcasts. However, not all of the items in the collections are digitized, and not all that is digitized is equally easily accessible for researchers.

1 See https://huggingface.co/Riksarkivet.

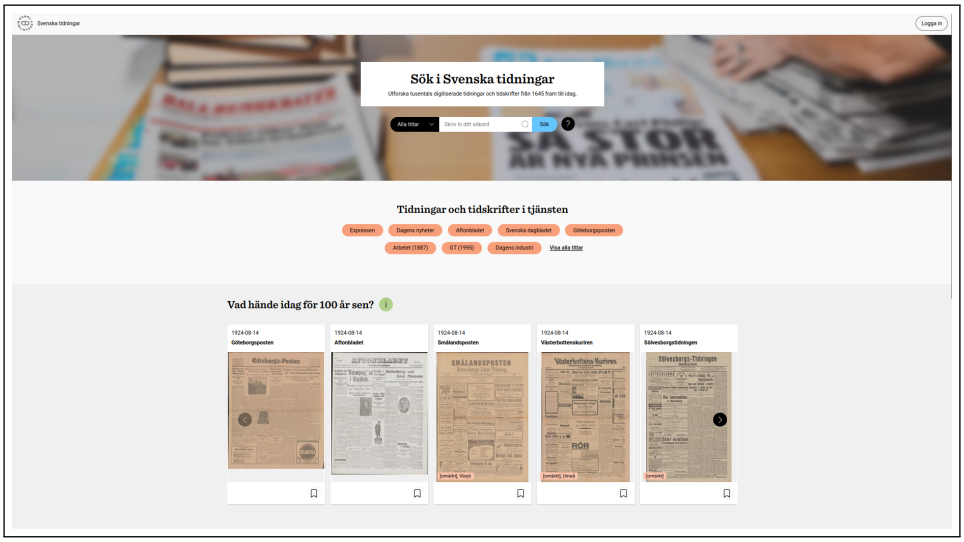


Figure 1: Searching for digitized newspapers with the library's search service

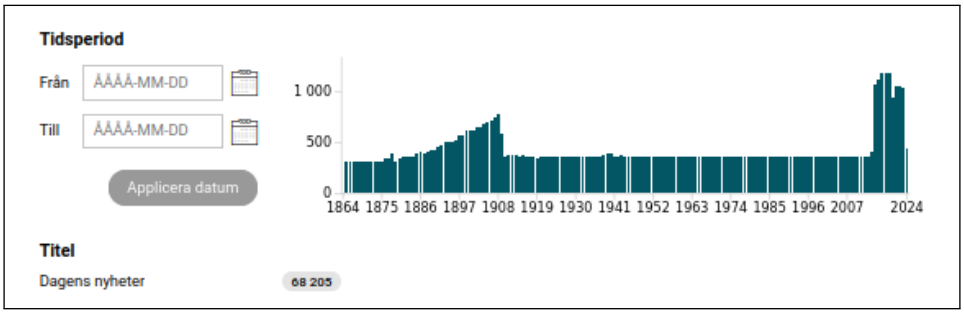


Figure 2: The availability of *Dagens Nyheter*, 1864-2024

This means that it is important to know how to search for the desired materials and to understand how ready the data is for research. Some items might be readily accessible and easy to use, while others may require additional steps or specialised tools for access and analysis. Recognizing these differences is essential, as it enables more efficient navigation through the resources and the selection of appropriate methods for working with the data.

2.1 *How to find the data?*

Let us take the newspaper archive as an example. This collection includes more than 4,376 titles from 1600 to the present, yet not all of the issues of

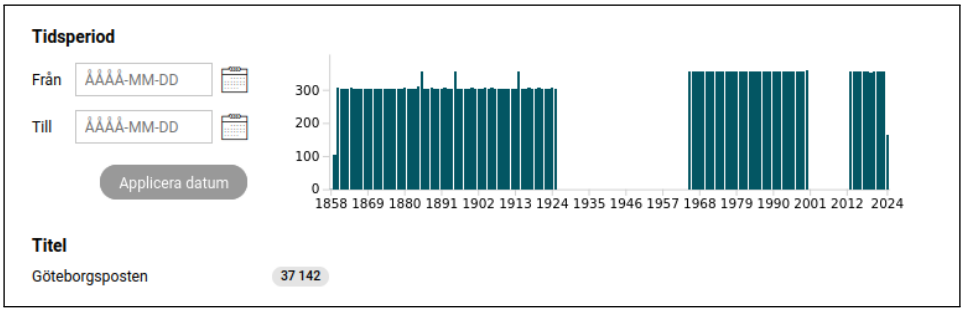


Figure 3: The less complete digital availability of *Göteborgsposten*, 1858-2024

these newspapers have been digitized. What is currently available in digital format is the entire collection of the newspapers up to 1906, along with everything from 2014 onwards, but the coverage for the intervening years is incomplete. The library's newspaper database is continually updated with newer material, though it takes approximately four months to digitize incoming material and make the most recent newspapers available.

A useful tool for checking the availability of the newspaper data is the library's online service: *Tidningar*.[2] This serves as a comprehensive database for all digitized newspapers in KB's collections (see Figure 1). Due to copyright restrictions, it is not possible to see the actual content of newspapers that are less than 100 years old outside of the library itself (though several Swedish university libraries have a designated computer to access this restricted material). However, even in the limited mode available to all users online, this is a helpful tool for checking the coverage of the digitized materials, as well as conducting preliminary searches to determine if the newspapers contain the desired content (i.e. to see if a particular word or term appears over time).

For example, a quick search shows that the newspaper *Dagens Nyheter* is fully available from 1864 to 2024 (see Figure 2); whereas *Göteborgsposten* can be accessed from 1859 to 2024, but the years 1925-1967 and 2001-2012 have yet to be digitized and are thus not digitally available (see Figure 3).

For exploring materials in KB's collection beyond the newspaper archive, the Swedish union catalogue, *Libris*, is an invaluable resource.[3] This contains freely available digital material as well as information about nearly 13 million items, including books, e-books and images from around 600 Swedish libraries (see Figure 4).

2 See https://tidningar.kb.se/.
3 See https://libris.kb.se/.

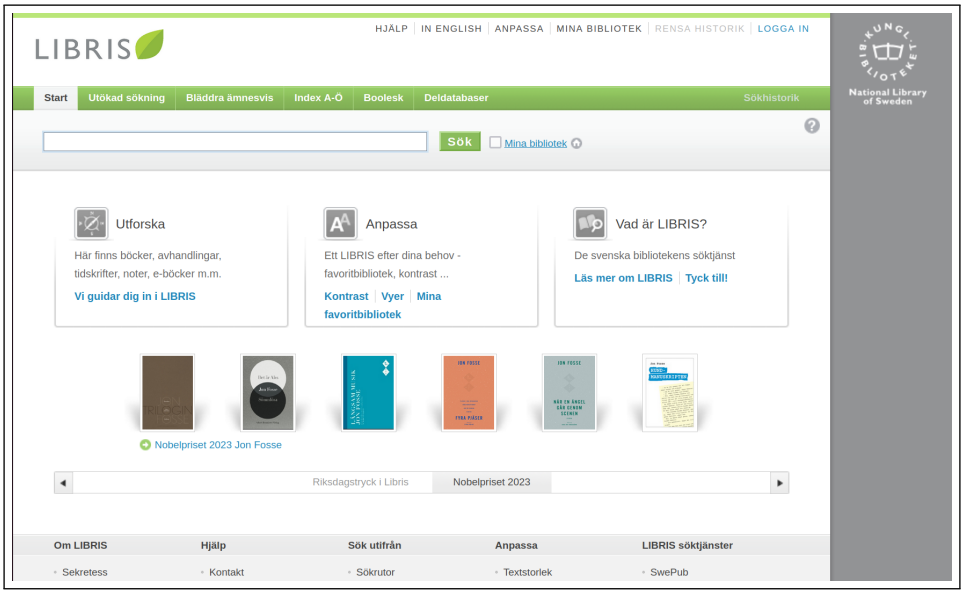


Figure 4: Libris as entrance point for search and discovery

2.2 *Current state of the data and challenges*

Knowing what data is available for research is a crucial first step. But it is equally important to consider the readiness of the data for research, which includes being aware of the format of the material, potential optical character recognition (OCR) errors associated with the digitization process and the extent to which the data has been cleaned, organized and prepared for analysis (Sikora & Haffenden 2024).

Some challenges concerning the quality of the newspaper data and other digitally available sources stem from the particular terms of Swedish legal deposit. Under current regulations, physical copies are prioritized, which means that digital versions of books or other materials do not need to be deposited if a printed version has already been submitted. This is unfortunate, as valuable metadata is lost during the digitization process. Moreover, the quality of digitized materials tends to be lower than their digitally-born equivalents. Direct access to digital versions of materials, when available, would help to mitigate these quality issues.

We can again take newspapers as an example to illustrate this point. Once the newspaper issues are submitted to KB as physical copies, they are forwarded to the digitization centre for scanning. Only after this step, and following the application of OCR models, do the materials become available as digital copies within KB. In practice, this means that much of the

information which is typically taken for granted when reading a physical newspaper is lost during digitization. This includes details such as the reading order of the text, the number of articles on a page, and more specific metadata such as distinguishing between text boxes in the body of an article, identifying which text boxes belong to a specific article, locating the title of an article or captions for images on a page. This loss of structure can complicate tasks like content analysis, archival research or simply navigating the digital format in a way that mirrors the original print experience (Sikora & Haffenden 2024: 62–63).

The final product of the digitization process is raw outputs from the OCR engine, in the form of scanned images and ALTO files. ALTO stands for Analyzed Layout and Text Object and is a metadata format used for encoding information about both layout and text of digitized documents, e.g. newspapers and books, by providing details such as the spatial coordinates of text elements. To enable easier navigation and processing of the materials, KBLab has developed a set of additional files derived from the content of the ALTO files. These are designed to organize and summarize key information from the OCR-outputs, and are useful for researchers who want to work with textual data on a larger scale, since they simplify the process of accessing and analyzing the content. However, it is important to note that these texts are *not* pre-processed and include OCR-errors that may require manual correction (see Figure 5). Consequently, users need at least a basic understanding of programming to be able to retrieve, process and analyze the data.

2.3 *What other data is available beyond newspapers?*

Moving beyond the newspaper data, KB's digital collections also include a number of digitized books, the Swedish Government Official Reports (SOUs) and parliamentary data, among other things. We provide more detail about how to access such data in Section 6 below.

Since material less than 100 years old is still protected by copyright and cannot be made publicly available, KBLab has undertaken various collaborations to make derivatives of this data accessible. One such initiative was a project with Språkbanken Text to create the Kubord and Kubord2 datasets.[4] These include word frequency analyses of the biggest Swedish newspapers, syntactic annotations of the texts at the word level and frequencies of collocations for individual words. The datasets can be downloaded from Språkbanken's webpage or explored graphically via the online tool *Korp*.[5]

4 See spraakbanken.gu.se/resurser/kubord and spraakbanken.gu.se/resurser/kubord2.
5 See https://spraakbanken.gu.se/korp/?mode=kubord.



Figure 5: OCR success and failure on a newspaper page. The OCR software correctly recognized the text in the yellow box: "DAGENS NYHETER". However, the text in the red box—"ÅNGMASKINER och"—was misread as: "tciåmasy\ver oc\v".

Another collaborative effort to enhance the availability of KB's collections involved creating vector models based on the same newspaper data as the Kubord2 datasets. These models can be used to track semantic changes and relationships between words (Forsberg et al. 2023, Bouma et al. 2024). As with the datasets mentioned above, the vector models are freely available online (Språkbanken Text 2024).[6]

6 See https://spraakbanken.gu.se/resurser/kubord-fasttext.



Figure 6: Accessing digitized manuscripts via Manuscripta

2.3.1 *Audio data*

Besides offering access to text data, KBLab has also created an audio dataset entitled RixVox (Rekathati 2023b,a). The underlying material for this dataset consists of Swedish parliamentary debates made available via the Swedish Parliament's open data initiative.[7] The debates were aligned with the corresponding transcripts from written protocols. Additionally, the debates are annotated with various metadata such as gender, age and name. The final product is a dataset consisting of 5,500 hours of Swedish audio material, which is freely available on the data science community platform Hugging Face.[8]

2.3.2 *Manuscripta*

A further source of data is the range of material published in KB's service *Manuscripta*.[9] This is a database of medieval and early modern digitized manuscripts from Swedish libraries, currently including around 800 manuscripts. The images are available in high quality according to IIIF standards and are freely available to download and make use of in research (see Figure 6).

7 See: https://www.riksdagen.se/sv/dokument-och-lagar/riksdagens-oppna-data/.
8 For this audio dataset, see: https://huggingface.co/datasets/KBLab/rixvox.
9 See: https://www.manuscripta.se/.

3 *How to access data via KBLab?*

Members of research projects that have been granted access to KB's digital collections are welcome to visit KBLab in Stockholm to prepare for working with the library's resources. (For more details about how to apply, see Section 6, below). During the onboarding process, we provide researchers with an overview of the infrastructure and logins to our systems. Since it is not possible to move copyrighted material beyond the lab's internal network, we have significant computational capacity in-house to ensure that researchers have sufficient compute resources to carry out their work. This includes a server environment and two NVIDIA DGX A100 servers for more computationally heavy analysis (Börjeson et al. 2024: 570).

The first stage of our onboarding involves signing a formal user agreement, which outlines the terms and conditions for using the Lab and our code of conduct.[10] We then set up an account on a local workstation with the Linux-based Ubuntu system. This provides researchers with the flexibility to create and manage their own software environments according to their own specific needs. On the workstations, our Application Program Interfaces (APIs) can be accessed to explore data, computation can be performed, and both data and results stored. We also have a local GitLab to allow researchers to manage and archive their code.

After the introduction, researchers can start using the library's resources and data, for example via our dedicated graphical user interfaces (GUIs): *Datalab* (datalab.kb.se) and *Betalab* (betalab.kb.se). Both interfaces offer similar functionalities, with one key difference: the data available in the beta version is, for the most part, not under copyright protection, which allows researchers to use it outside of the library and in advance of any visit to KBLab. This provides an opportunity to learn how to navigate the webpage, search for resources and learn how to download the data. Doing this beforehand can save time familiarizing yourself with the interface at the lab. We also offer an overview of the API's and GUIs as part of the onboarding process, along with providing detailed written instructions about how to use these.

3.1 *Datalab*

After logging in to one of the GUIs—*Datalab* or *Betalab*—searches of KB's digital collection can be conducted using the menu to the left or by filtering the packages in the search bar (see Figure 7). A package can constitute, for

10 For details of this code (in Swedish), see: https://www.kb.se/samverkan-och-utveckling/kb-labb/samarbeta-med-kb-labb.html.



Figure 7: Accessing data via the Lab's GUI

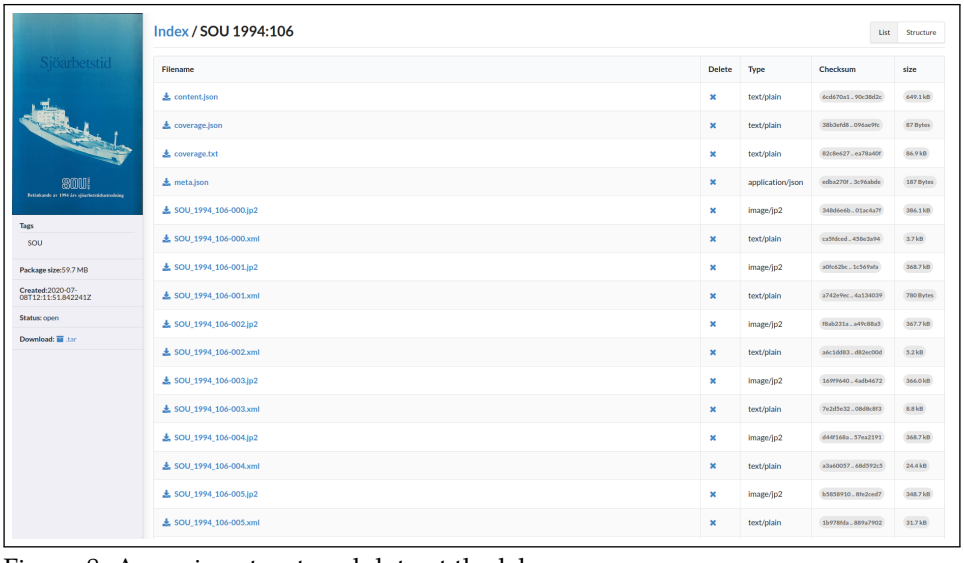


Figure 8: Accessing structured data at the lab

example, a complete newspaper issue, a full book or a SOU. Packages can be filtered based on various tags, such as title, creation date or resource type.

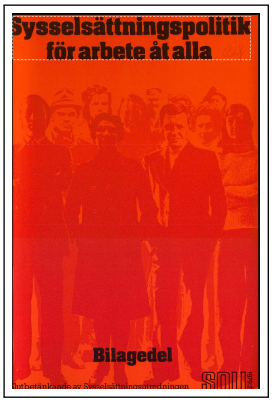
There are two ways to navigate the search: through the "package" or "page" format. Selecting the "page" option displays an individual page from newspapers or books that meet the search criteria. Results are shown as JPG files, and clicking on one of these opens a scanned image of the page. Alternatively, choosing the "package" option combines pages into a single package, such as an entire newspaper issue or a complete SOU document. In this mode only the first page of the package is shown during the search. When using the "package" format, two additional modes are available: "list" and "structure".

The "structure" tab shows all scanned images within a package, providing a visual overview of the entire contents. On the other hand, the "list" tab displays all files that are part of this package, including, but not limited to, the digitized pages in JPG format (see Figure 8). This view allows access to the following files:

`content.json` During digitization, the content of a page is divided into individual boxes containing text or images. The content.json file provides information about all the boxes present on each page within a package. The file format is JSON, where one entry in a dictionary corresponds to one OCR-box. In the example below, metadata about two boxes is shown (see Figure 9b). The "content" key contains information about the text recognised by the OCR engine, while the "box" key provides the coordinates for the box's position on the page. Additionally, there are two other keys: "type", which indicates whether the content was recognized as a picture or text, and "id", a unique identifier for each bounding box. The "id" is a Uniform Resource Identifier (URI), consisting of the following components:

- *https://betalab.kb.se*—the link to the API;
- *sou-1979-27*—the identifier for the entire package;
- *#1-1*—the first number corresponds to the part of the package, and the second number specifies the page; and
- *cblock_0-block_0*—the identifier for the specific box on the page.

Stable URIs are an essential means for researchers using the lab to be able to find their way back to the same point in the collections, and ultimately to demonstrate that their results are reproducible (by providing others with a means of verifying the data).



a.


```
[
  {
    "@id": "https://betalab.kb.se/sou-1979-27#1-1-cblock_0-block_0",
    "@type": "Text",
    "box": [
      "0",
      "11",
      "1745",
      "319"
    ],
    "content": "Syssels\u00e4ttningspolitik\nf\u00f6r arbete \u00e5t alla\n"
  },
  {
    "@id": "https://betalab.kb.se/sou-1979-27#1-1-cblock_1-block_1",
    "@type": "Text",
    "box": [
      "1720",
      "2619",
      "35",
      "100"
    ],
    "content": ":6Z61\n"
```


b.

Figure 9: The cover image of an SOU (a) and its corresponding "content" information in the lab environment (b)


```
title:      "SOU 1979:27"
year:       "1979"
created:    "1979"
```


Figure 10: Lab metadata in JSON format

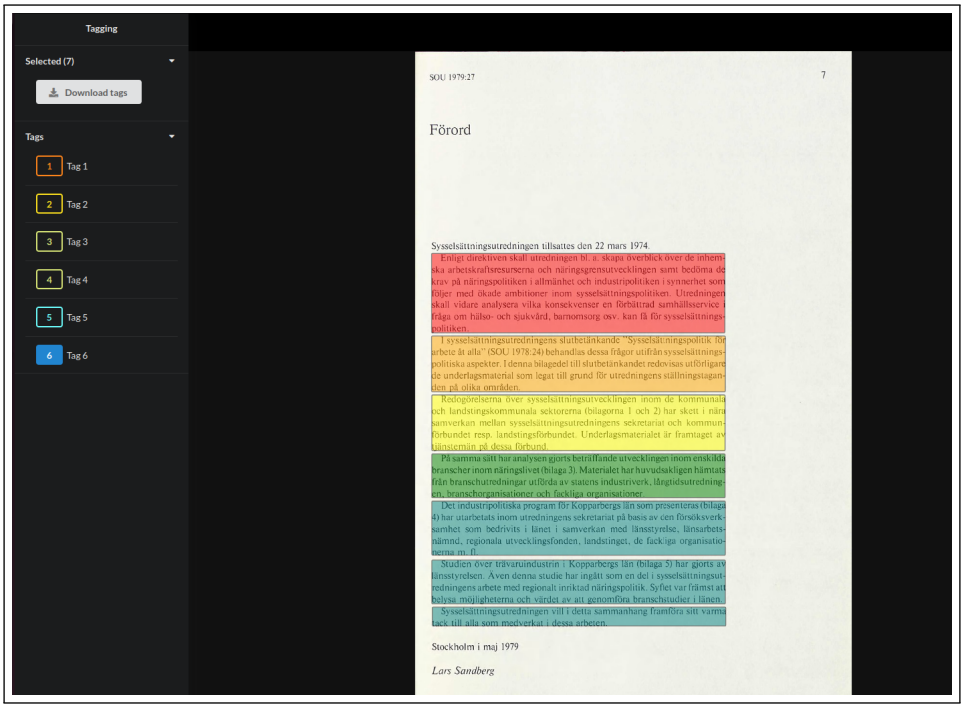


Figure 11: Annotating data via the lab's GUI.

`meta.json` This file consists of metadata about the package, such as the title, publication year and details about the year when the item was created (see Figure 10). The metadata also serves as tags that can be used to filter and search for specific pages and packages.

XML files These are the raw output files created during the OCR process and serve as the basis for creating content.json files.

`structure.json` . This file provides detailed information about the structure of a newspaper. It outlines which text blocks or elements are part of specific pages, offering a clear view of the layout and arrangement of the newspaper's content.

JP2 files These are digitized pages in JPEG 2000 format, representing the scanned pages of the material.

In addition to exploring the data graphically, the GUI also provides a simple annotation tool (see Figure 11). It can be opened by using the "Shift + T" shortcut while in the "structure" view of a package. The tool allows users to annotate six different classes and export the results into a CSV file. The annotation file includes the ID, type, tag, box and content fields. Due to copyright restrictions, annotations of the collections in *Datalab* can only be stored locally on lab computers.

3.2 *KBLab's Python library*

Datalab is an important part of KB's infrastructure, providing a graphical interface for exploring digital data. However, for researchers who want to conduct large-scale analysis, manually calling KBLab's API to download and process large volumes of data programmatically might be more suitable than examining individual objects. Another alternative is to use the python "kblab-client" package, which can be installed with pip.[11]

To familiarize yourself with the API calls, the package can be tested outside of the library on the beta version of the webpage, *Betalab*. Instructions and examples for using the APIs programmatically are available on our GitHub page, where you can access Jupyter notebooks demonstrating how to filter, download and process data at the lab.[12]

11 See: https://pypi.org/project/kblab-client/.
12 See: https://github.com/kb-labb/kblabb-examples/tree/master/api.

4 *Digital tools and AI models*

Curating digital collections and producing datasets for researchers on-site while enhancing access to non-copyrighted materials are central tasks at KBLab. Another key objective is utilizing these collections to develop digital research tools, including collections-based AI models trained on the library's data. Due to practical constraints, such as the limited number of workstations available, only a select number of projects can be hosted at KBLab at any given time. However, by producing and openly releasing AI models, we provide access to the benefits of these resources for a broader audience beyond the library itself. Using Swedish-language data to train high-quality AI models and making them freely available represents an important step toward democratizing AI for smaller languages (Haffenden, Fano, et al. 2023: 44–45). The following section provides an overview of these models and the research they enable.

4.1 *Language models for text analysis*

Our primary focus has been on training and fine-tuning encoder models, such as BERT (Devlin et al. 2019), which produce contextual representations rather than generating text. Unlike encoder-decoder architectures like GPT, which are designed for text generation, encoder models are better suited for tasks such as classification and retrieval. Their smaller size and fewer parameters make them more computationally efficient and cost-effective, which is particularly beneficial for digital researchers working with limited computational resources.

Although the original BERT model was released in 2018, it remains highly relevant today. Subsequent advancements, such as ModernBERT, have introduced improvements in speed, longer context windows and more efficient training methods (Warner et al. 2024). While generative models are optimal for open-ended text generation, our focus at the lab remains on developing smaller, fine-tunable encoder models that require fewer computational resources but still perform effectively in a range of applications.

Different models are better suited to specific tasks, depending on the research needs. The performance of KBLab-trained models, alongside other Swedish language models, can be evaluated through the Superlim leaderboard, a benchmark designed to compare Swedish models across multiple tasks (Berdicevskis et al. 2023). Since KBLab's establishment in 2019, we have trained multiple models in various modalities. The first of these was KB-BERT, a transformer-based language model trained on approximately 15–20 GB of Swedish text, including digitized newspaper archives dating

back to 1945, social media texts, Wikipedia articles and more (Malmsten et al. 2020). The variety of these textual resources helped capture different styles and uses of the language, which led to the creation of a diverse corpus for training a robust Swedish model. KB-BERT's compactness and versatility for applications within Natural Language Processing (NLP) means that this model is still widely used.[13] It can be utilized for a range of NLP tasks such as part-of-speech tagging (POS) and named entity recognition (NER).

Training a high-performing language model requires significant computational resources and large datasets. However, fine-tuning a pre-trained model is considerably less demanding and can be done on a consumer GPU or via an online cloud platform such as Google Colab. Due to copyright regulations, KB's data cannot be uploaded to these platforms, so we provide local GPU access at the lab for researchers who need to fine-tune models for tasks such as classification. For those lacking the resources or expertise to fine-tune models themselves, we also offer several pre-trained, off-the-shelf models ready for immediate use.

Among these is a BERT model for Swedish NER, designed to identify entities such as locations, organizations, and personal names. For example, given the sentence:

> *Göran Persson och Carl Bildt träffades på ett café vid Stureplan den 5e juli 1995.*
> 'Göran Persson and Carl Bildt met at a café by Stureplan on 5 July 1995.'

The NER model can identify the following entities:

- Person: Göra Persson, Carl Bildt
- Location: Stureplan
- Date: 5 July 1995

This capability is particularly useful for anonymization, where personally identifiable information (PII) must be removed or altered to comply with data privacy regulations such as GDPR. Identifying sensitive entities in datasets enables researchers to work with anonymized data while maintaining compliance with GDPR.

4.1.1 *Sentiment Analysis*

Another fine-tuned model available at KBLab is a sentiment analysis model for Swedish, capable of classifying text as positive, negative or neutral.[14] This

13 To access this model, see: https://huggingface.co/KBLab/bert-base-swedish-cased. For user statistics that offer a sense of how widely the model is used, see: https://github.com/kb-labb/huggingface_stats.

14 To access this model, see: huggingface.co/KBLab/robust-swedish-sentiment-multiclass.

model was fine-tuned using a diverse mix of sources, including Trustpilot reviews, news headlines and other texts (Hägglöf 2023). Unlike rule-based sentiment systems, which rely on predefined word lists, transformer-based models consider the semantic context in which words appear, mitigating errors that arise when sentiment is determined solely by individual word associations. Using a BERT model therefore enables a more linguistically sophisticated approach to identifying and analyzing the sentiment of a given text.

Beyond NER and sentiment analysis, fine-tuned models can solve a wide range of classification tasks. For social scientists, classifiers can be valuable for opinion mining: extracting and categorizing public sentiment from social media, surveys and other sources. More generally, digital humanists can use classification models to organize large text corpora into categories to enable further analysis. In literary studies, author attribution models can help determine the authorship of anonymous or disputed texts through stylistic analysis. These diverse applications illustrate the broad research potential and the versatility of fine-tuned NLP models.

4.1.2 *Sentence-BERT and topic modeling*

The original BERT model generates contextual embeddings for individual tokens within a sentence, making it ideal for token-level tasks such as NER, part-of-speech tagging and question answering. However, for some research applications, encoding larger chunks of text at once is more useful. This can be achieved using Sentence-BERT (SBERT), a modification of BERT that generates sentence-level embeddings well-suited for semantic similarity, text clustering and retrieval tasks.

We have trained two Swedish SBERT models at KBLab, both freely available on Hugging Face (Rekathati 2021b). The first KB-SBERT model processes texts up to 256 tokens, while the updated version extends the context window to 384 tokens to enable analysis of longer text chunks at the sentence level.[15] Sentence-BERT excels at detecting paraphrases and retrieving relevant documents in response to queries. While the standard BERT model is optimal for fine-grained word-level analysis, SBERT provides better overall sentence-level understanding.

4.1.3 *Transformer-based topic modeling with BERTopic*

One notable application of Sentence-BERT is BERTopic, a transformer-based approach to topic modeling (Grootendorst 2022). BERTopic clusters text by first generating sentence embeddings, reducing their dimensionality, and

15 To access these models, see: huggingface.co/KBLab/sentence-bert-swedish-cased.

then applying clustering algorithms to group similar topics. This process enables researchers to uncover coherent and meaningful topics in large text corpora. Unlike classical topic modeling methods such as Latent Dirichlet Allocation (LDA), which require extensive text preprocessing—including stop-word removal, stemming and lemmatization—BERTopic can operate effectively with minimal pre-processing, working directly on raw text (Fano & Haffenden 2022). This is possible because SBERT retains the contextual meaning of words within the text rather than treating them as isolated tokens. BERTopic thus provides a more contextually-aware analysis of topics.

BERTopic can be applied in various domains of digital research. It can be used for analyzing historical documents, for instance in helping identify thematic trends and semantic change in newspapers, speeches and letters. It can also be used in literary research as a means of detecting recurring themes across different works. Within the social sciences, BERTopic can be deployed to analyze social media posts and news articles to allow researchers track shifts in public discourse and sentiment over time.

To facilitate research that uses BERTopic, we provide various training resources, including workshops and Google Colab notebooks with detailed instructions and adjustable scripts.[16] These require no specialist software or hardware to use, and can be tested by users with little prior programming experience. The notebooks include detailed instructions and clarifications alongside the code blocks, making them accessible even to those new to coding. The script is also highly adjustable: allowing researchers to swap out datasets, adjust topic parameters, and customize model outputs, making it a flexible and user-friendly introduction to topic modeling.

4.1.4 *Use case: topic modeling at the library*

We have explored applying BERTopic at the library as an alternative method for classifying textual collections (Malmsten et al. 2025). Currently, subject classification is carried out manually by specialist librarians using the Swedish subject headings system, *Svenska ämnesord* (SAO). This controlled vocabulary is designed to provide consistency and clarity in cataloging. However, manual indexing poses various difficulties, as indexers may interpret classification guidelines differently, and the system itself has grown organically over time rather than being optimized for emerging research needs. These classification challenges for librarians can affect how easy it is for users to navigate through the collection.

16 To access this online workshop, see: https://colab.research.google.com/drive/10kB3wfoHSfZE48vEKmznIw-ff36uR8gs?usp=sharing.

To address these challenges, we have experimented with automated methods for subject analysis. Instead of assigning predefined categories, we use transformer-based topic modeling, allowing the classification and ordering to emerge from the contents of the works themselves. Since BERTopic clusters documents based on their textual properties, it can be used to generate data-driven topics rather than relying on pre-established subject headings. Adopting this method could reduce inconsistencies in cataloging and improve discoverability. We discuss this project in greater depth in an Open Access IFLA volume on the use of AI in libraries (Malmsten et al. 2025).

4.2 *Speech recognition models for audio data*

At KBLab, we develop AI models and digital tools not only from textual collections but also by leveraging KB's extensive audio-visual archives. The collections include thousands of hours of legally deposited resources, such as TV and radio broadcasts, which provide an ideal material for training Automatic Speech Recognition (ASR) models.

State-of-the-art ASR models such as Wav2Vec 2.0 (Meta) and Whisper (OpenAI) are particularly valuable for transcribing spoken language into text (Baevski et al. 2020, Radford et al. 2023). These models play a crucial role in making audio-visual collections more accessible, enabling automated transcription for research, cataloging, and search applications. In addition to providing accurate transcriptions, ASR models can be used to develop audio-driven search tools, allowing users to efficiently retrieve specific content from vast audio and video datasets, such as podcasts, radio programs or historical broadcasts.

For speech models to perform reliably across different accents and linguistic variations, they must be trained on a diverse range of audio data. Similar to large language models (LLMs), ASR models trained primarily on English-language data often underperform when applied to smaller languages such as Swedish. Even multilingual models, which are trained on several languages, are typically exposed to limited amounts of Swedish data, making them less effective at capturing the nuances of regional dialects and spoken variations. Recognizing this limitation, we have focused on training speech recognition models specifically for Swedish using local radio broadcasts from KB's archives, ensuring stronger performance across a range of dialects. By prioritizing regional diversity, we aim to improve linguistic inclusivity and enhance the overall democratic accessibility of our ASR models (Haffenden, Fano, et al. 2023: 44-45).

Our Wav2vec2 model called VoxRex is trained on more than 10,000 hours of P4 radio material from the last 20 years (Malmsten et al. 2022). P4 is

Table 1: Comparative WER on Swedish dialects for Meta's XLSR, the monolingual VoxPopuli model, and KBLab's VoxRex models (Malmsten et al. 2022: 2). VoxRex-A, -B, and -C refer to versions trained on different amounts of speech data and varying numbers of training updates. WER refers to the percentage of errors registered by each model, as explained on page 36.

	Model				
Region	XLSR	VoxPopuli	VoxRex-A	VoxRex-B	VoxRex-C
Dalarna	5.62	3.17	2.74	2.74	1.67
Göteborg w. env.	5.64	3.26	2.86	2.82	1.78
Mellansverige	5.62	3.30	2.85	2.83	1.79
Norrland	6.27	3.68	3.12	3.20	1.95
Stockholm w. env.	4.57	2.64	2.26	2.30	1.43
Västergötland	5.53	3.26	2.78	2.81	1.76
Västra sydsverige	7.62	4.25	3.77	3.84	2.10
Västsverige	5.40	3.16	2.65	2.71	1.60
Östergötland	5.65	3.23	2.81	2.77	1.68
Östra sydsverige	6.68	3.80	3.29	3.23	1.94

a Swedish public local radio channel, which consists of around 25 radio stations and therefore provides diverse representation in terms of dialects. The broadcasts contain a large variation of different types of content ranging from music, news and sport reports to people calling in and on-site reporting.

One of the common metrics used to evaluate the performance of ASR models is Word Error Rate (WER), which measures the accuracy of transcriptions by calculating the percentage of words that were incorrectly predicted by the ASR system compared to a reference transcription. As we can see in the table below, the models perform differently on various regional variations of Swedish with Southern and Northern dialects posing more difficulties than, e.g., the Stockholm accent (see Table 1). This proves the need to train the models on more balanced and varied datasets in order to mitigate this effect. In comparison to Meta's speech model XLSR, our Wav2vec 2.0 large VoxRex Swedish model (C) provides an improvement of up to 5.52% on the South-Western dialect ("Västra sydsverige").

KBLab is conducting a project to train updated versions of Wav2Vec 2.0 with larger amounts of data, as well as datasets enriched with additional dialects. Besides releasing Wav2Vec 2.0, the project has also fine-tuned a Swedish Whisper model, KB Whisper, which offers state-of-the-art performance for Swedish speech to text. The work was conducted on the EURO HPC supercomputer Leonardo in Italy.

4.2.1 *Choosing a model for ASR*

Which model should you choose for your research? The choice between Wav2Vec 2.0 and Whisper mostly depends on the specific requirements of the ASR task, given differences in how they handle formatting, accuracy and potential errors like hallucinations.

Wav2Vec 2.0 is designed to output transcriptions that closely mirror the spoken materials, without introducing words or phrases that were not explicitly stated. This makes it ideal for tasks requiring word-for-word accuracy, such as legal transcriptions, technical documentation or academic research. However, Wav2Vec 2.0 typically produces transcripts in lowercase and without punctuation, which may require additional processing to improve readability and format the text appropriately. Additionally, Wav2Vec 2.0 is generally less prone to hallucinations—errors where the model "imagines" content that wasn't actually spoken—making it more reliable for tasks that require verbatim transcription that are faithful to the original speech.

On the other hand, Whisper is renowned for its ability to paraphrase, making it well-suited for tasks such as real-time captioning. Whisper automatically adds punctuation and capitalization, which significantly improves the readability of the transcripts. However, Whisper might sometimes introduce paraphrasing or include words that weren't explicitly spoken, which could be problematic in scenarios where absolute precision is crucial. For researchers working with audio data who might be interested in using either of these models for transcription, we have produced a freely-available tutorial notebook on Google Colab with code and instructions.[17]

4.2.2 *Model-based applications at the library*

Beyond providing a means to democratize AI, collection-based Swedish models can also be used to make the library's collections more accessible to researchers and anyone interested in exploring KB's materials. One initiative aimed at demonstrating this potential involved developing a demo platform called *Mediesök*, which showcases how speech models can make audio materials searchable in new ways (see Figure 12).

Audio materials legally deposited at KB often lack detailed metadata, which is important for understanding the content of these materials. Typically, the only information available is the time and name of the radio broadcast. To allow users to search through the materials without needing to listen to each audio clip individually, an alternative approach involves

17 See: https://colab.research.google.com/drive/1RCP53jqClJz0zDX_VBT_K04BevUKk842?usp=sharing#scrollTo=ib2Owc2w3gHa.

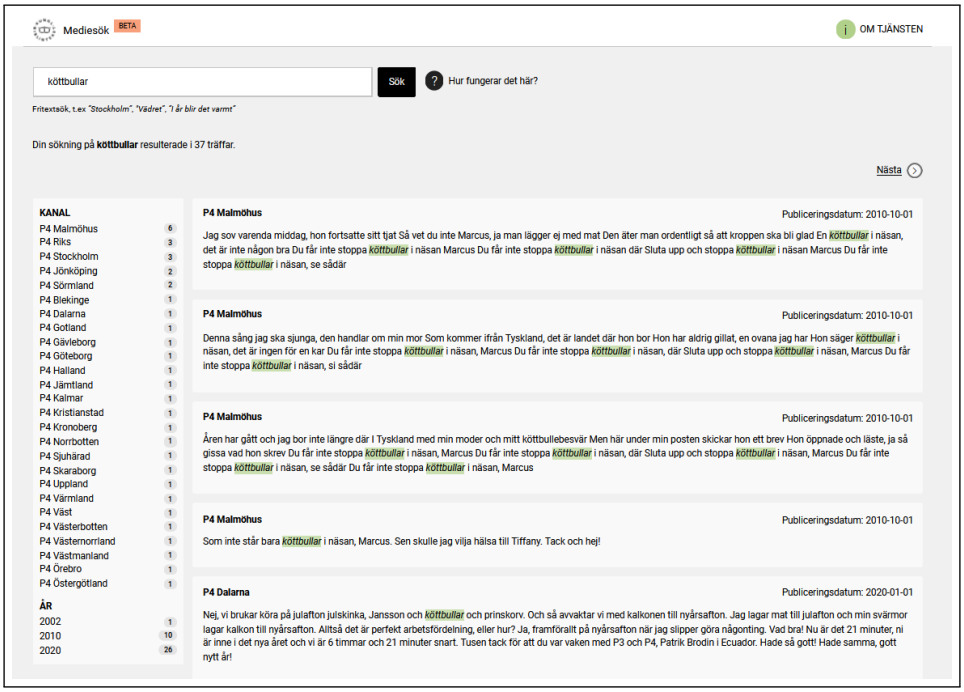


Figure 12: Locating *köttbullar* 'meatballs' in the radio archives. Making audio-visual data amenable to contents-based search via the *Mediesök* demo.

generating textual materials from the audio data via transcription.

To demonstrate this idea, we transcribed 10 hours of P4 radio broadcasts using WhisperX, which is a version of the Whisper model that includes word-level time-stamping, making it possible to easily navigate through the transcriptions and corresponding audio clips. This feature enhances the usability of audio archives, allowing users to effectively search for specific content within the audio materials. *Mediesök* enables researchers to carry out free text searches upon audio material that was previously challenging, if not impossible, to search through, short of listening through each file in real time.

4.3 *Visual models for image analysis*

Besides audio and textual materials, we also work with image collections at KBLab. Similar to our efforts with the audio materials, AI tools can significantly enhance the searchability of visual heritage. One way that we have shown how AI can be integrated into the library's working practices is

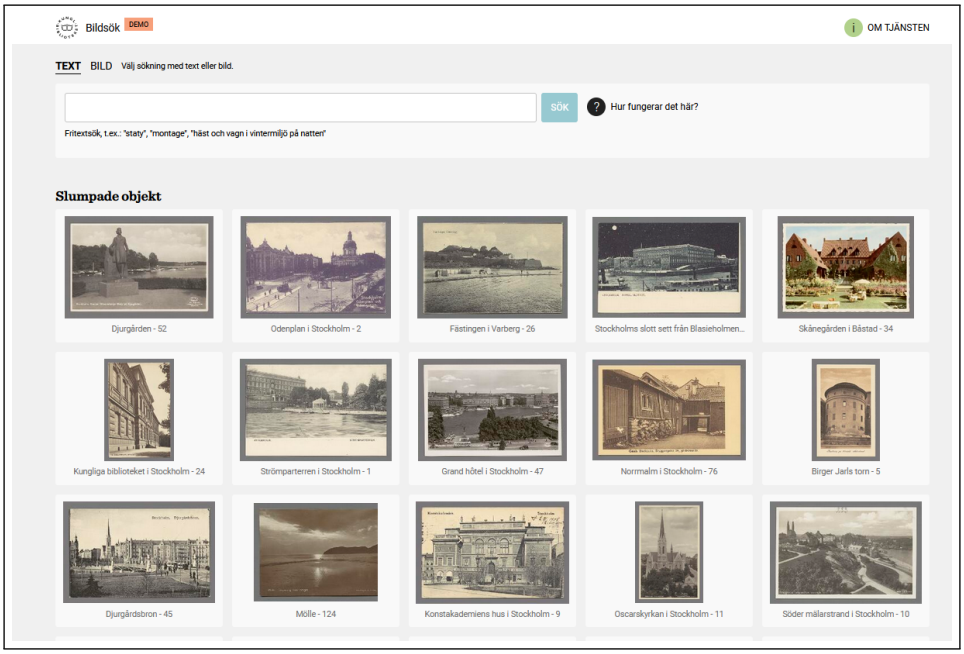


Figure 13: Image search with the help of AI

through the creation of a demo tool called *Bildsök*—i.e. image search.[18] This allows users to explore image collections in new ways: either by uploading an image and searching for similar ones, or by entering free text to find images that match the desired search terms (Haffenden, Rekathati, et al. 2023). It is particularly useful for material such as ephemera that often lacks metadata at the level of the individual object. Deploying an AI model to analyze the visual content of such material sidesteps the absence of metadata to make it searchable according to users' interests.

The model enabling this demo is Swe-CLIP 2M, a model that links images to their textual descriptions. Originally, this model was trained on English data by OpenAI, but its capabilities have since been adapted to other languages, where we helped to produce a Swedish version (Carlsson et al. 2022). The model progressively learns to associate text with images more effectively, enabling it to match the best images to specific text queries. The *Bildsök* demo currently features a collection of more than 17,000 postcards to showcase this functionality (see Figure 13). As an example of how this works, we can observe that searching for a church does indeed return a number of images depicting a sacral building (see Figure 14).

18 See: https://lab.kb.se/bildsok/.

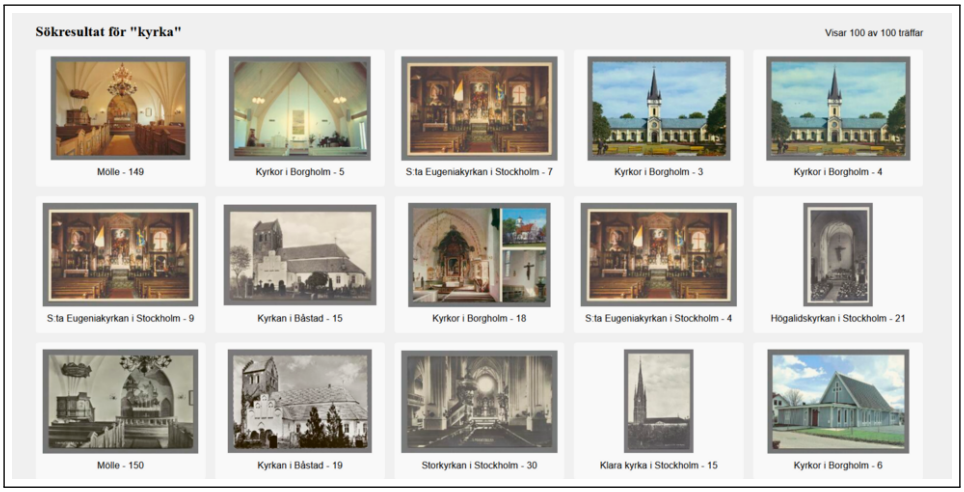


Figure 14: Unlocking image collections with AI-based search

The flexibility of the image search engine lies in its underlying method. When a text query or an image is entered into the search box, it is first converted into a numerical representation or vector. This vector is then compared with the vectors of the items stored in the database. The comparison is made based on cosine similarity—a measure of how closely two vectors align in multidimensional space. The search results are ranked according to their similarity scores, with the most similar items appearing at the top of the list. We have described this demo in more detail in a post on our blog (Haffenden, Rekathati, et al. 2023).

4.4 *Development work and future models*

We are currently engaged in a range of development projects aimed at improving the digital infrastructure for Swedish language processing, leveraging access to the EU's EuroHPC supercomputing infrastructure.[19] A key initiative within this framework is the ongoing effort to train a Swedish Whisper model mentioned above, which is designed to support more diverse and representative speech-to-text applications. By fine-tuning Whisper on a broader range of high-quality training data, we aim to enhance its ability to transcribe spoken Swedish with greater precision and cultural nuance. The new series of ASR models resulting from this work was made freely available via our Hugging Face page in spring 2025, providing researchers and practitioners

19 See: https://www.eurocc-access.eu/why-does-the-royal-library-of-sweden-need-hpc-video/.

with improved tools for audio data processing (Vesterbacka et al. 2025).[20]

Another major focus of our development work is the creation of an updated and improved generation of BERT models for enhanced text analysis. A key constraint of encoder models is their context window, which determines how much textual information a model can process at once before encoding it into numerical representations. At present, KB-BERT has a maximum token length of 512, limiting the amount of text it can process per input. Our goal is to train new models with extended context windows, allowing them to process significantly larger text segments in a single pass. This will be particularly valuable for tasks such as classification, information retrieval and NER, as it enables models to capture and analyze broader linguistic patterns without requiring extensive text chunking or segmentation.

5 *Use cases: What sort of research is possible using KBLab's AI models?*

AI enables large-scale research on the library's collections across diverse academic areas, from social sciences, linguistics and history to literary studies, ecology and beyond. The integration of AI-based technologies into library collection management is transforming how researchers interact with and utilize these resources. By combining humanities perspectives with data science methods, researchers are unlocking new analytical possibilities. One striking example is the "Living with Machines" project, a large-scale collaboration between universities, libraries and research institutes in the UK.[21] This multidisciplinary initiative brought together historians, geographers, data scientists and computational linguists to analyze nineteenth-century newspapers, studying the human impact of the Industrial Revolution. By using newly digitized materials alongside the British Newspaper Archive, the project led to the development of novel tools and assessments of OCR quality, among other outcomes.[22]

A significant aspect of this project was the creation of collections-based historical language models, which were later applied in various studies, such as research on "atypical animacy" (Coll Ardanuy et al. 2020). By utilizing BERT's masking function as a register of linguistic change, researchers examined how the semantic associations of words evolved over time. This approach allowed them to trace shifts in meaning and animacy in an innovative manner (see Figure 15).

20 See: huggingface.co/collections/KBLab/kb-whisper-67af9eafb24da903b63cc4aa.
21 See: https://livingwithmachines.ac.uk/.
22 For further results, see: https://livingwithmachines.ac.uk/achievements/.

Original sentence: And why should one say that the machine does not live?
Masked sentence: And why should one say that the [MASK] does not live?
Predictions with scores: *man* (5.0788), *person* (4.4484), *other* (4.1866), *child* (4.1600), *king* (4.1510), *patient* (4.1249), *one* (4.1141), *stranger* (4.1067), ...

Figure 15: Using a historical language model to register the surprise of semantic innovation (Coll Ardanuy et al. 2020: 4536)

A further instance of the creative deployment of AI within machine learning-based humanities was the project's use of Word2Vec vector models trained on historic newspaper data to track the semantic change of terms related to mechanization, such as "gear", "fellow" and "wheel" (Pedrazzini and McGillivray 2022).

Similar methodologies have been employed at KBLab, where collections-based models have been used in multiple projects. These examples showcase the potential for using AI as the basis for innovative research on GLAM collections, which would otherwise be impossible to conduct at scale. These AI models can facilitate the identification of patterns, extraction of insights and visualization of results, thereby broadening the scope for interdisciplinary collaboration. In the following section we outline instances of such work from projects carried out at the lab.

5.1 *Televising information*

One of the first projects at KBLab, "Televising Information: Audiovisual Communication of Swedish Government Agencies," led by Emil Stjernholm (Lund University), focused on KB's audio-visual collections. The study examined *Anslagstavlan*, a bulletin program produced by Sveriges Television (SVT) since 1972, which government agencies have used to communicate with the public (Stjernholm 2023). By collaborating with the lab, the project applied ASR models to transcribe spoken content from the video recordings, making large-scale analysis of this archival material feasible.

Audio-visual materials in KB's collections are stored as continuous recordings of radio and television broadcasts, which posed a particular challenge. The first step of the project involved manually locating and extracting the relevant *Anslagstavlan* segments from the recordings. Once isolated, these clips were processed using KBLab's Wav2Vec 2.0 large VoxRex model (see Section 4.2), which transcribed the spoken content into text. Manual corrections were then made to enhance accuracy. This approach enabled the researcher to study how Swedish government agencies have communicated with the

public over time in a way that would have been prohibitively labor-intensive using traditional methods.

The result was a 160,000-token corpus derived from 600 episodes of *Anslagstavlan*. Using AntConc, a text analysis tool, the researcher conducted concordance and keyword-in-context analyses, identifying changes in communication strategies and thematic trends over time. The study could thereby offer a new perspective on Sweden's public information practices and their evolution.

This project aligns methodologically with the previously mentioned *Mediesök* demo, as both employ machine learning to transform complex datasets into accessible research materials. Without automated techniques like ASR and text analysis, studying extensive video content like *Anslagstavlan* would have required exhaustive manual effort. By leveraging AI, the project could embark upon large-scale analysis of historically significant audio-visual data.

5.2 *Mining for meaning*

Since 2019, KBLab has hosted "Mining for Meaning: The Dynamics of Public Discourse on Migration," a six-year initiative led by researchers from Linköping University. This project analyzes large text corpora—including parliamentary speeches, web data and newspapers—to examine shifts in the discourse on migration and integration.[23]

Through computational analysis, researchers track the development of sentiment on these topics over time within three domains: the public sphere, the media and politics. By treating newspapers as historical records of collective discourse, as "social sensors", the project applies machine learning techniques to analyze meaning-making processes in society.

One part of the project involved investigating how sentiment towards migration has shifted over time in the national news, using this as a proxy for tracking societal changes in attitudes (Hurtado Bodell et al. 2024). They use a corpus derived from the four largest Swedish newspapers—*Aftonbladet*, *Dagens Nyheter*, *Svenska Dagbladet* and *Expressen*—from the period 1945–2019, which they filtered to create a sub-corpus with "immigration-rich" articles, where each text contains a sufficient percentage of tokens connected to the topic of immigration. Researchers developed a semi-supervised extension to classical topic modeling, called seeded topic modeling, which directs the model's focus using predefined "seed" words, such as "refugees" and "asylum seekers".

By applying machine learning to large-scale newspaper archives, the project enables sociological analysis of historical discourse, providing in-

23 See: https://liu.se/en/research/computational-text-analysis.

sights into how Swedish society has framed migration issues over the decades. It demonstrates how innovative combining sociology and data science approaches to large digital collections can offer new ways of approaching how society makes sense of social changes and trends over time.

5.3 *Welfare state analytics*

Another example of a project conducting large-scale text analysis of digitized collections is "Welfare State Analytics: Text Mining and Modeling Swedish Politics, Media & Culture, 1945–1989" (WeStAc), a collaboration involving researchers from Umeå University, Uppsala University and Aalto University in Finland.[24] The project focused on digitizing and analyzing Swedish texts from 1945 to 1989 to create datasets from Swedish governmental reports, newspapers and literary journals, and thus explore societal and discursive shifts in the domains of politics, media and culture.

Techniques like word embedding and probabilistic methods have been employed to track the evolution of concepts such as emancipation and individualization between 1945 and 1989. One case study analyzed how the notion of "political" evolved over time in post-war Sweden (Norén et al. 2023). This was done by creating and analyzing bigrams and LDA topic models based on a corpus of 27 million tokens derived from articles from the newspapers *Dagens Nyheter* and *Aftonbladet*. The analysis could demonstrate the expansion of the term "political" into broader social and cultural discussions, while noting how core topics like international conflicts, party politics, elections and economic policy remained dominant. This suggests how large-scale approaches that examine the library's collections as data can fruitfully explore change and continuity over time, given a project team with the right skill set (see Section 6 below).

5.4 *Multimodal topic modeling*

Topic modeling has proven to be a valuable tool in digital humanities research, and it could be fruitfully deployed in the GLAM sector too. We have previously described how Sentence-BERT models can be used in BERTopic pipelines to organize and analyze large volumes of text data. Beyond textual analysis, BERTopic also supports multimodal topic modeling, which incorporates both visual and textual signals to improve classification and retrieval processes. This method is based on the CLIP model, used to create the *Bildsök* demo discussed earlier, and further demonstrates how machine learning can enhance search capabilities in visual collections.

24 See: https://www.westac.se/en/.

	Topic	Count	Representation	Visual_Aspect
0	-1	6834	[person, horse, carriage, drawn, church, riding, going, front, street, the]	
1	0	650	[walking, down, street, couple, city, people, building, large, tower, it]	
2	1	570	[church, tower, clock, top, with, front, in, on, of,]	

Figure 16: Topic modeling and thematic analysis of image collections

By integrating multimodal AI models, cultural heritage institutions can significantly enhance the searchability and classification of their collections. Multimodal topic modeling enables automated tagging and categorization of archival materials by considering both images and associated text descriptions. For example, a model like CLIP, which simultaneously processes visual and textual inputs, can detect similarities between a painting's image and its descriptions, fostering richer semantic connections within a database. This approach allows for more nuanced and intuitive explorations of cultural collections (see Figure 16 for AI-based thematic clustering examples).

Combining machine learning techniques further expands the potential applications of topic modeling. For instance, ASR models can transcribe audio and audio-visual materials into text, enabling topic modeling on previously inaccessible material. This facilitates large-scale analysis of radio and TV archives, which contain valuable but underutilized historical insights. By converting spoken content into structured, searchable text, institutions can significantly enhance the accessibility and usability of extensive audio-visual archives.

5.5 *OCR and post-processing improvements*

A persistent challenge in working with digitized data is the quality of OCR outputs. As mentioned earlier, KBLab's data undergoes minimal processing beyond correcting the most common OCR errors. Without systematic post-correction or accounting for potential misspellings, OCR inaccuracies can distort research findings. In the previously described projects, two studies relied on digitized newspaper data to create datasets. To curate these datasets, text blocks containing the term "political" or tokens related to immigration were extracted and used as a basis for corpus construction. However, when working with sources not originally produced in digital form, additional processing steps are required to handle misspelled or fragmented words, making them more difficult to identify and analyze.

This is a common problem with OCR, and researchers have recently developed various models to mitigate these issues. One potential solution involves training a new OCR model specifically for Swedish, which could yield more accurate results. However, while this approach is promising, it is also resource-intensive and costly when applied to large datasets. A more feasible alternative is to improve the quality of existing OCR outputs through post-processing techniques. Methods such as spell-checking, context-aware corrections and machine learning models trained to recognize common OCR errors can significantly enhance the accuracy of datasets, leading to more reliable research outcomes.

One study exploring these possibilities involved fine-tuning a transformer model to perform post-OCR correction on Swedish newspapers from the nineteenth and twentieth centuries (Löfgren & Dannélls 2024). The researchers selected the ByT5 model, which operates at the character level and was pre-trained on the multilingual mc4 dataset. Fine-tuning was conducted using a manually annotated subset of digitized newspapers from KB's collections. The best-performing model achieved a 36% reduction in character error rate, demonstrating the potential for automated OCR correction to improve textual accuracy. This model is intended to be integrated into Språkbanken Text's automatic processing pipeline and is freely available on Hugging Face.[25]

5.6 *Layout analysis*

We have tested layout analysis in a project that experimented with using a Document Image Transformer (DIT) model for filtering the editorial materials from the rest of the contents (Sikora & Haffenden 2024: 64–65). Likewise,

25 To access the model, see: https://huggingface.co/viklofg/swedish-ocr-correction.

similar efforts have been undertaken in another project for dividing the advertisements from the rest of the text blocks by training multiple multimodal models on various mixtures of text, images and metadata (Rekathati 2021a).

Beyond improving OCR quality, the data readiness of the digitized newspapers could also be enhanced by reconstructing the structure of the original publications. Studies using KB's newspaper collections typically operate at the text block level due to the absence of metadata linking blocks to specific articles. This lack of structural information means that researchers cannot easily determine which text blocks belong to which articles, how many articles are contained within a newspaper issue or the correct reading order of the texts. However, recent advances in machine learning provide new opportunities to address these challenges.

One approach has been to apply machine learning techniques to reconstruct the layout of digitized newspapers. In a recent project at the lab, a Document Image Transformer (DIT) model was tested for filtering editorial content from other textual elements (Sikora & Haffenden 2024). Similarly, another project sought to separate advertisements from editorial texts by training multiple multimodal models on a combination of textual, visual, and metadata inputs (Rekathati 2021a). These initiatives demonstrate how computational methods can be used to organize and segment complex historical documents more effectively.

A key prerequisite for these methods is the annotation of training data, which allows machine learning models to learn how to differentiate between various types of text blocks. A useful strategy involves leveraging existing datasets to improve corpus quality, extending to the development of classifiers for filtering specific content within large corpora. For example, researchers can annotate a sub-corpus of gold-standard text blocks related to a specific topic, which can then be used to fine-tune a machine learning model capable of automatically identifying relevant text blocks in broader volumes of data.

By employing supervised learning techniques, researchers can develop classifiers that accurately identify relevant content, even within vast and noisy corpora. This approach not only improves the accuracy of topic-specific datasets but also significantly reduces the time and effort required for manual data curation. Ultimately, such processes can improve the overall quality and usability of digitized resources, making them more accessible and valuable for research.

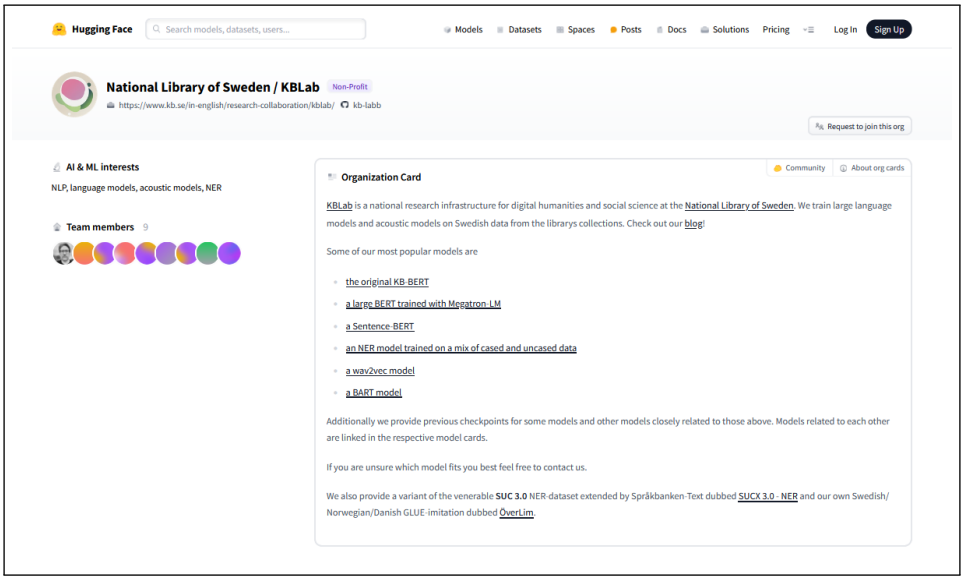


Figure 17: KBLab's models available Open Access via Hugging Face

6 *Initiating research via KBLab: How and where to get started?*

Having outlined the available data, models and research possibilities at KBLab, we now turn to practical steps for initiating a research project. This section provides guidance on different entry points, depending on your research interests and the level of engagement you seek with the lab's resources.

6.1 *Accessing AI models*

Our AI models are freely available via our Hugging Face page (see Figure 17).[26] If you plan to use these models in your research and have questions about their training data or suitability for your specific use case, we encourage you to reach out via email to kblabb@kb.se. Our data scientists can provide orientation and guidance, although we may not always have definitive answers to every inquiry. Documentation and pre-print articles related to the models are linked on our Hugging Face page where available. We also welcome feedback from researchers who have used our models—sharing your experiences helps improve and refine our resources.

26 See: https://huggingface.co/KBLab.

6.2 *Betalab for open data access*

For researchers working with copyright-free material, such as nineteenth-century Swedish newspapers, Swedish Government Official Reports (SOU's) and parliamentary records from the twentieth century, remote access is available via *Betalab* (see Section 3.1). This interface allows you to search, annotate and download data in a structured format. Additionally, *Betalab* serves as a preparatory environment for researchers planning an on-site visit to KBLab, allowing for script development and preliminary analysis before engaging with restricted materials at the lab. Contact us via `kblabb@kb.se` for further dialogue about access.

6.3 *Conduction research on-site*

For researchers requiring access to copyrighted materials, such as twentieth-century Swedish newspapers, KBLab provides a secure on-site environment in Stockholm. Due to copyright restrictions, this cannot be accessed remotely or exported beyond the lab's internal network. Researchers planning to work with this data should consider the following:

Firstly, since copyright restrictions mean access is only available on-site at KBLab's locale in Stockholm, all research needs to be conducted at the lab. It is not possible to move data beyond the lab's internal network, making it comparable to an archaeological site, i.e. amenable to research but essentially immobile.

Secondly, since working with processing structured data presumes a degree of technical competence, large-scale research at the lab requires programming proficiency in Python or R. While our visual search interface is sufficient for some tasks, most research projects working at scale must include team members with data science expertise. If your project lacks these skills, consider collaborating with digital humanities research infrastructures (e.g., GRIDH at Gothenburg University, CDH at Uppsala University) or university-based computer science departments. Our data scientists at the lab do not generally work hands-on in external projects, partly due to resource constraints but also because we are sceptical towards the outsourcing of technical skills that form an integral part of doing digital research (Börjeson et al. 2024: 571).

Thirdly, we provide high-performance computing resources for compiling and processing large datasets, which raises the question of funding. We judge the particular needs of a project on a case-by-case basis depending on its particular demands, but each project generally pays some form of annual overhead fee to use the lab. These costs tends to be included within

project budgets financed by funding institutions like the Swedish Research Council (VR) or Riksbankens Jubileumsfond (RJ). Fees may be waived in cases where the project contributes significant infrastructural value to the library, though such a judgement is a result of KB's process of assessing incoming collaboration requests.

6.4 *Applying for research collaboration*

To initiate a research collaboration at KBLab, you should visit the library's website for further instructions.[27] This process is designed as a dialogue rather than a formal funding application: it is *not* comparable to the type of application that you might submit to VR or RJ and it will suffice to send us a brief outline of your project plans (ca. 1 page). By outlining your research goals and specifying which parts of KB's digital collections you intend to work with, we can assess the feasibility of your proposed project and explore the possibilities for collaboration. Once this dialogue is complete, you will receive a decision about the terms of a collaboration. We recommend beginning this process well in advance of external funding deadlines to ensure smooth planning.

6.5 *Blog for ideas and examples*

For insights into how our AI tools are being used, as well as updates on new models, we invite you to explore the KBLab blog (see Figure 18).[28] This includes case studies on topic modeling with transformers (Hägglöf & Sikora 2023), classical topic modeling approaches (Fano 2021) and the creation of a Swedish speech corpus from parliamentary debates (Rekathati 2023a). We also welcome guest contributions from researchers who have utilized KBLab's resources, such as a recent Master's project featured on the blog (Nachesa 2023). If you are interested in sharing your experiences here, please reach out to us!

7 *Conclusion*

In this chapter, we have provided a hands-on orientation to using KBLab at the National Library of Sweden, highlighting both its role as a research infrastructure and the innovative ways AI is being used to enhance digital

27 See: https://www.kb.se/om-oss/forskningsprojekt-i-samverkan.html.
28 See: https://kb-labb.github.io/.

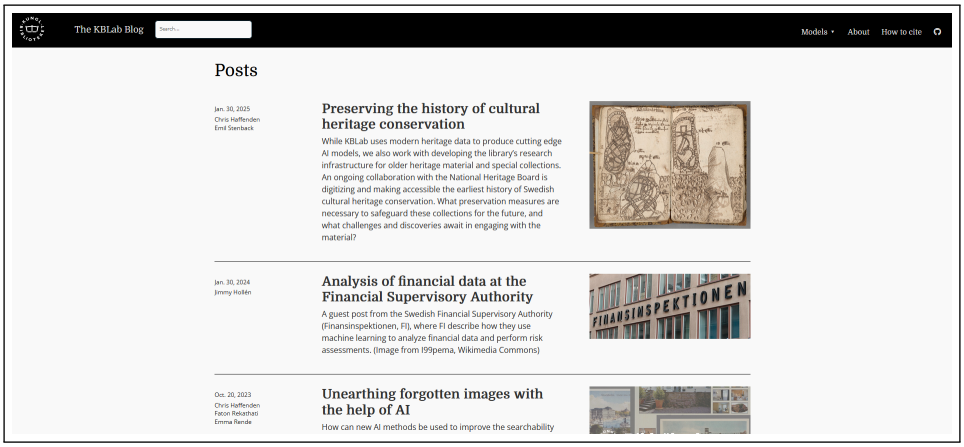


Figure 18: Sources of inspiration from the KBLab blog

heritage accessibility. The rapid development of AI tools has not only improved the searchability of heritage collections but has also created novel opportunities for digital researchers to engage with these collections at scale.

By showcasing how various research projects at KBLab have utilized AI-based methodologies—whether for text analysis, OCR correction or multimodal exploration—we hope to have illustrated the lab's potential for fostering new forms of scholarly inquiry. Beyond demonstrating what is currently possible, this chapter also invites researchers to think creatively about how AI and digital tools might further expand the scope of historical and cultural analysis in the future.

As AI continues to develop, so too will the possibilities for digital research. We encourage scholars from diverse disciplines to explore KBLab's resources and reflect upon how emerging AI tools might shape their own research. If this chapter has sparked new ideas or raised questions, we welcome further dialogue—please feel free to reach out and engage with us.

References

Baevski, Alexei, Henry Zhou, Abdelrahman Mohamed & Michael Auli. 2020. *Wav2vec 2.0: A framework for self-supervised learning of speech representations*. https://arxiv.org/abs/2006.11477v3 (8 April, 2025).

Berdicevskis, Aleksandrs, Gerlof Bouma, Robin Kurtz, Felix Morger, Joey Öhman, Yvonne Adesam, Lars Borin, Dana Dannélls, Markus Forsberg, Tim Isbister, Anna Lindahl, Martin Malmsten, Faton Rekathati, Magnus Sahlgren, Elena Volodina, Love Börjeson, Simon Hengchen & Nina Tah-

masebi. 2023. Superlim: A Swedish language understanding evaluation benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 8137–8153. Singapore. DOI: 10.18653/v1/2023.emnlp-main.506.

Börjeson, Love, Chris Haffenden, Martin Malmsten, Fredrik Klingwall, Emma Rende, Robin Kurtz, Faton Rekathati, Hillevi Hägglöf & Justyna Sikora. 2024. Transfiguring the library as digital research infrastructure: Making KBLab at the National Library of Sweden. *College & Research Libraries* 85(4). 564–582. DOI: 10.5860/crl.85.4.564.

Bouma, Gerlof, Markus Forsberg, Justyna Sikora & Emma Sköldberg. 2024. Konsten att bedriva svensk ordforskning utan att kränka upphovsrätten. *Proceedings of the Huminfra Conference* (*HIC 2024*). 161–167. DOI: 10.3384/ecp205022.

Carlsson, Fredrik, Philipp Eisen, Faton Rekathati & Magnus Sahlgren. 2022. Cross-lingual and multilingual CLIP. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 6848–6854. Marseille, France. https://aclanthology.org/2022.lrec-1.739/.

Coll Ardanuy, Mariona, Federico Nanni, Kaspar Beelen, Kasra Hosseini, Ruth Ahnert, Jon Lawrence, Katherine McDonough, Giorgia Tolfo, Daniel CS Wilson & Barbara McGillivray. 2020. Living machines: A study of atypical animacy. In *Proceedings of the 28th international conference on computational linguistics*, 4534–4545. Barcelona, Spain (Online). DOI: 10.18653/v1/2020.coling-main.400.

Cordell, Ryan. 2020. *Machine learning + libraries: A report on the state of the field*. Report commissioned by LC Labs. Library of Congress. https://labs.loc.gov/static/labs/work/reports/Cordell-LOC-ML-report.pdf.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1* (*long and short papers*), 4171–4186. Minneapolis, Minnesota. DOI: 10.18653/v1/N19-1423.

Fano, Elena. 2021. *Topic models for Statens Offentliga Utredningar*. The KBLab Blog. https://kb-labb.github.io/posts/2021-05-04-topic-models-for-sous/.

Fano, Elena & Chris Haffenden. 2022. *BERTopic for Swedish: Topic modeling made easier via KB-BERT*. The KBLab Blog. https://kb-labb.github.io/posts/2022-06-14-bertopic/.

Forsberg, Markus, Justyna Sikora & Emma Sköldberg. 2023. *Words unboxed: discovering new words with Kubord*. The KBLab Blog. https://kb-labb.github.io/posts/2023-08-29-kubord/ (8 April, 2025).

Grootendorst, Maarten. 2022. *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. https://arxiv.org/abs/2203.05794v1 (8 April, 2025).

Haffenden, Chris, Elena Fano, Martin Malmsten & Love Börjeson. 2023. Making and using AI in the library: Creating a BERT model at the National Library of Sweden. *College & Research Libraries* 84(1). DOI: 10.5860/crl.84.1.30.

Haffenden, Chris, Faton Rekathati & Emma Rende. 2023. *Unearthing forgotten images with the help of AI*. The KBLab Blog. https://kb-labb.github.io/posts/2023-10-20-unearthing-forgotten-images-with-the-help-of-ai/.

Hägglöf, Hillevi. 2023. *A robust, multi-Label sentiment classifier for Swedish*. The KBLab Blog. https://kb-labb.github.io/posts/2023-06-16-a-robust-multi-label-sentiment-classifier-for-swedish/.

Hägglöf, Hillevi & Justyna Sikora. 2023. *Scientific discourse with BERTopic*. The KBLab Blog. https://kb-labb.github.io/posts/2023-03-17-scientific-discourse-with-bertopic/.

Hurtado Bodell, Miriam, Måns Magnusson & Marc Keuschnigg. 2024. Seeded topic models in digital archives: Analyzing interpretations of immigration in swedish newspapers, 1945–2019. *Sociological Methods & Research*. 00491241241268453. DOI: 10.1177/00491241241268453.

Löfgren, Viktoria & Dana Dannélls. 2024. Post-OCR correction of digitized Swedish newspapers with ByT5. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, 237–242. St. Julians, Malta. https://aclanthology.org/2024.latechclfl-1.23/.

Mahey, Mahendra, Aisha Al-Abdulla, Sarah Ames, Paula Bray, Gustavo Candela, Sally Chambers, Derven Caleb, Milena Dobreva-McPherson, Katrine Gasser, Stefan Karner, Kristy Kokegei, Ditte Laursen, Abigail Potter, Armin Straube, Sophie-Carolin Wagner & Lotte Wilms. 2019. *Open a GLAM lab*. Digital Cultural Heritage Innovation Labs, Book Sprint, Doha, Qatar, 23-27 September 2019. DOI: 10.21428/16ac48ec.f54af6ae.

Malmsten, Martin, Love Börjeson & Chris Haffenden. 2020. *Playing with words at the National Library of Sweden: Making a Swedish BERT*. https://arxiv.org/abs/2007.01658v1 (8 April, 2025).

Malmsten, Martin, Chris Haffenden & Love Börjeson. 2022. *Hearing voices at the National Library – a speech corpus and acoustic model for the Swedish language*. https://arxiv.org/abs/2205.03026v2 (8 April, 2025).

Malmsten, Martin, Viktoria Lundborg, Elena Fano, Chris Haffenden, Fredrik Klingwall, Robin Kurtz, Niklas Lindström, Faton Rekathati & Love Börjeson. 2025. Without heading? Automatic creation of a linked subject

system. In *New horizons in artificial intelligence in libraries*. Edmund Balnaves, Leda Bultrini, Andrew Cox & Raymond Uzwyshyn (eds.). Berlin, Boston: De Gruyter Saur. 179–198. DOI: 10.1515/9783111336435-014.

Nachesa, Maya. 2023. *For how long is a person recognisable by their voice?* The KBLab Blog. https://kb-labb.github.io/posts/2023-07-04-for-how-long-is-a-person-recognisable-by-their-voice/.

Norén, Fredrik, Johan Jarlbrink, Alexandra Borg, Erik Edoff & Måns Magnusson. 2023. The transformation of 'the political' in post-war Sweden. In *Digitised newspapers: A new Eldorado for historians? Reflections on tools, methods and epistemology*. Estelle Bunout, Maud Ehrmann & Frédéric Clavert (eds.). Berlin: De Gruyter Oldenbourg. 411–436. DOI: 10.1515/9783110729214-019.

Padilla, Thomas. 2025. □Humanities data in the library: Integrity, form, access□. *D-Lib Magazine* 2016(23). https://www.dlib.org/dlib/march16/padilla/03padilla.html (8 April, 2025).

Radford, Alec, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey & Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th international conference on machine learning* (ICML'23). Honolulu, Hawaii, USA.

Rekathati, Faton. 2021a. *A multimodal approach to advertisement classification in digitized newspapers*. The KBLab Blog. https://kb-labb.github.io/posts/2021-03-28-ad-classification/.

Rekathati, Faton. 2021b. *Introducing a Swedish sentence transformer*. The KBLab Blog. https://kb-labb.github.io/posts/2021-08-23-a-swedish-sentence-transformer/.

Rekathati, Faton. 2023a. *Finding speeches in the Riksdag's debates*. The KBLab Blog. https://kb-labb.github.io/posts/2023-02-15-finding-speeches-in-the-riksdags-debates/.

Rekathati, Faton. 2023b. *RixVox: A Swedish speech corpus with 5500 hours of speech from parliamentary debates*. The KBLab Blog. https://kb-labb.github.io/posts/2023-03-09-rixvox-a-swedish-speech-corpus/.

Sikora, Justyna & Chris Haffenden. 2024. AI, data curation and the data readiness of heritage collections: Exploring the Swedish newspaper archive at KBLab. *Proceedings of the Huminfra Conference* (*HiC 2024*). 60–67. DOI: 10.3384/ecp205009.

Språkbanken Text. 2024. *Kubord-fasttext* [*Data set*]. DOI: https://doi.org/10.23695/sp99-9h02.

Underwood, Ted. 2019. *Distant horizons: Digital evidence and literary change*. Chicago: University of Chicago Press.

Vesterbacka, Leonora, Faton Rekathati, Robin Kurtz, Justyna Sikora & Agnes Toftgård. 2025. *Welcome KB-Whisper, a new fine-tuned Swedish Whisper*

model! The KBLab Blog. https://kb-labb.github.io/posts/2025-03-07-welcome-KB-Whisper/.

Warner, Benjamin, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard & Iacopo Poli. 2024. *Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference*. https://arxiv.org/abs/2412.13663v2 (8 April, 2025).

List of abbreviations

AI	Artificial Intelligence
ALTO	Analyzed Layout and Text Object (metadata schema)
ASR	Automatic Speech Recognition (the basis for speech-to-text applications)
BERT	Bidirectional Encoder Representations from Transformers (a language model)
CLIP	Contrastive Learning-Image Pretraining (a multimodal model)
CPU	Central Processing Unit (a standard processing chip for a computer)
GLAM	Galleries, Libraries, Archives and Museums
GPT	Generative Pre-trained Transformer (the type of language model underpinning ChatGPT)
GPU	Graphics Processing Unit (a processing chip used to train AI models)
GUI	Graphical User Interface
IFLA	International Federation of Library Associations and institutions
KB	*Kungliga biblioteket* 'the National Library of Sweden'
LLM	Large Language Model
ML	Machine Learning
NER	Named Entity Recognition
NLP	Natural Language Processing
OCR	Optical Character Recognition (process of converting image of text to machine-readable text)
PII	Personally Identifiable Information
RJ	*Riksbankens Jubileumsfond*
SBERT	Sentence-BERT language model

SOU	*Statens offentliga utredningar* 'Swedish Government Official Reports'
VR	*Vetenskapsrådet* 'Swedish Research Council'
WER	Word Error Rate (metric for evaluating speech-to-text performance)

Corresponding authors

Chris Haffenden
KBLab
National Library of Sweden
chris.haffenden@kb.se

Justyna Sikora
KBLab
National Library of Sweden
justyna.sikora@kb.se

3 Interdisciplinary digital project design

Daniel Brodén
University of Gothenburg

Mats Fridlund
University of Gothenburg

Cecilia Lindhé
University of Gothenburg

Jonathan Westin
University of Gothenburg

While discussions in digital humanities increasingly emphasise the importance of reflecting on collaborative workflows for interdisciplinary research, attention to specific practical expertise remains lacking. This paper introduces the concept of *interdisciplinary digital project design* to highlight a professional practice that integrates collaboration between traditional Humanities and Social Science (HSS) researchers and technical experts in developing research projects, digital resources and more. We begin by addressing the need for *protocols* to support workflow-oriented approaches to interdisciplinary collaboration, while underscoring the role of embodied expertise in facilitating teamwork. Furthermore, we argue that *judgement* – a critical yet often overlooked element – is an integral aspect of the professionalism involved. The discussion is grounded in descriptions of our contribution to five digital HSS projects, each offering a different perspective on the integrative professionalism involved. The paper concludes by discussing ways to further advance the conceptual understanding of interdisciplinary digital project design, with particular attention to the expertise that underpins this practice.

1 *Introduction*

It is well-known that while Digital Humanities (DH) projects combine humanistic and technical expertise in various ways, interdisciplinary collaboration can be challenging due to differing disciplinary rationales. Computer scientists often focus on pushing methodological development, whereas researchers in the Humanities and Social Sciences (HSS) primarily aim to apply disciplinary methods to digital scholarship. However, it has also been argued that DH has cultivated the emergence of a "third culture" as originally described by C.P. Snow, where individuals from fundamentally different disciplinary traditions make significant strides toward collaborating

HUM INFRA Daniel Brodén, Mats Fridlund, Cecilia Lindhé & Jonathan Westin. 2025. Interdisciplinary digital project design. In Gerlof Bouma, Dana Dannélls, Dimitrios Kokkinakis & Elena Volodina (eds.), *Huminfra handbook: Empowering digital and experimental humanities* (NEALT Proceedings Series 59), 57–76. University of Tartu Library. DOI: 10.58009/aere-perennius0172 ©The authors, CC BY 4.0

more organically (Snow 1959, Hunter 2014, Ingvarsson 2021). Moreover, there is a growing discussion within DH about the need to develop more integrative interdisciplinary workflows. For instance, the concept of "agile hermeneutics" describes a collaborative process where HSS scholars and data analysts engage in a continuous, dialogical and reflexive relationship (Rockwell & Sinclair 2016; c.f. Fickers et al. 2022).

Nevertheless, there remains a palpable lack of blueprints and considerations regarding the kind of expertise needed to develop interdisciplinary DH projects. As have been observed in a recent study, "new projects and initiatives expend a lot of energy in their start-up period trying to establish collaborative values and project management strategies, often reinventing the wheel in the process" (Ahnert et al. 2023: 4).

In response to Ted Underwood's (2014) call for greater articulation and reflection on the partly tacit working methods within DH and building on an earlier short conference paper (Brodén, Fridlund, et al. 2024), the present chapter highlights our conceptual work at the Gothenburg Research Infrastructure in Digital Humanities (GRIDH, formerly Centre for Digital Humanities, CDH), at the University of Gothenburg. We focus on what we have chosen to call the practice of *interdisciplinary digital project design*, an evolving approach to outlining and integrating collaboration between technical expertise and "traditional" HSS researchers for projects involving multidisciplinary teamwork. By "projects", we refer to research projects in the strict sense, but also project applications and the development of digital resources for researchers and cultural heritage institutions, such as tools, databases, corpora and archives, that call for interdisciplinary approaches to meet current field standards.

1.1 *Design – planning and doing*

The discussion in this chapter draws on our combined experiences as senior staff, project managers and project members at GRIDH, part of the Faculty of Humanities at the University of Gothenburg. Since 2015, our infrastructure has initiated and developed data-rich research projects, research applications and digital resources in collaboration with researchers from the Humanities and other faculties at the University as well as heritage professionals at Swedish GLAM (galleries, libraries, archives and museums) institutions. Additionally, our involvement as researchers in various DH projects and our work with national research infrastructures, such as Huminfra, InfraVis and Språkbanken and Swe-Clarin – the Swedish node of CLARIN (Common Language Resources and Technology Infrastructure) – have further informed and shaped the insights presented in this chapter.

In our efforts to mediate and translate across interdisciplinary fields, we have come to see our expertise, in part, as a form of "design" capability that encompasses both the realms of "planning" and "doing". Etymologically, the word *design* originates from the Latin word *designere*, which means, among other things, to 'mark out', 'point out' or 'devise'. The latter Italian verb *designare* acquired specific aesthetic connotations, such as to draw and to paint. These meanings were adopted into both French and English, with the English term *design* further developing to encompass ideas of "contriving", "outlining" and more artistic forms of creative practice. Today, the term design is associated with a wide range of professional disciplines, including architecture, engineering, crafts, industrial design, fashion and art (Merriam-Webster n.d.). Moreover, the concept of *design thinking* refers to a set of procedures used in design and project organisation in a general sense as well as to the body of knowledge developed around design ability (Rowe 1987, Cross 2011).

The term design may seem somewhat grandiose, but we find it useful for delineating a specific form of practice that involves both planning and doing. More specifically, we propose the notion of *interdisciplinary digital project design* to highlight an emerging form of expertise in DH within the context of what has been described as the fractured trading zones of DH (Svensson 2012). This expertise shares similarities with what (Hunter 2014: 27) refers to as "bridge people", individuals who are either fully competent in the "two cultures" or, at least, "bilingual" in the sense of being able to communicate effectively across the disciplinary boundaries (Berry & Fagerjord 2017). This chapter focuses primarily on the latter category, offering a broader perspective on the potential forms of bridging discussed. However, this form of expertise extends beyond just mediating between technical and humanistic fields. While it does not necessarily involve project management in the sense of "organising" others, it invariably engages with the spheres of "planning" and "doing" that are central to the practice of design.

To clarify, this chapter does not aim to discuss specific skills or best practices for project design. Instead, we focus on highlighting and conceptualising some core aspects of a particular interdisciplinary-oriented professionalism. This professionalism is centred on outlining and integrating collaboration between technical experts and HSS researchers, an area in need of being put into focus in discussions about workflows for interdisciplinary teamwork in DH. We argue that this perspective is important for better understanding the processes involved in collaborative work within the field and the expertise often necessary to achieve integrative interdisciplinarity.

1.2 *Disposition*

We begin our discussion by situating the concept of interdisciplinary digital project design within the context of theoretically oriented writings on integrative mixed methods approaches in DH. Engaging with Sarah Oberbichler et al.'s (2022) discussion of integrative approaches to multidisciplinary teamwork centred around historical data, we emphasise the importance of structuring collaborative workflows and adhering to some kind of *protocol*, procedures designed to enable consistency and quality in the collaborative efforts. However, we also contend that mixed-methods approaches should not be considered solely in abstract terms but must be grounded in embodied expertise. To this end, we draw on philosopher of education Gert Biesta's (2015) writing about *judgement* as an important but often overlooked part of professionalism. Specifically, we argue that the exercise of judgement is a critical part of interdisciplinary digital project design. Judgment is necessary to address collective needs, guide workflows and decide on "what needs to be done" at different points in a project.

Having outlined a conceptual framework for understanding the expertise involved, we ground the discussion by briefly describing our contribution to five digital projects, each offering a different perspective on the integrative professionalism involved. *First*, we comment on the project "The Role of Music and Sound in Audiovisually Mediated Conspiracy Theoretical Discourse" as an example of GRIDH's general approach to supporting researchers in the preparation of research project applications. *Second*, we discuss GRIDH's involvement in the political science project "The Benefits of Conflict" to illustrate how we contribute to the development of digital resources in external projects, providing more than just technical expertise. *Third*, we outline the iterative process involved in producing data for the project "Digital documentation of inscriptions in the Saint Sophia Cathedral in Kyiv". *Fourth*, we describe the early design phase of the large mixed-methods project "Terrorism in Swedish Politics", focusing on the effort to assemble a highly interdisciplinary research team. *Fifth*, we highlight the extensive interdisciplinary collaboration in the project "Tracing the Carvers on the Rocks", which integrates humanistic and AI expertise, emphasising the role of not only human but also machine agencies in the research process.

Taken together, these case studies provide an opportunity to explore the integrative professionalism involved from multiple perspectives, highlighting how it represents a form of expertise that is both distinct and highly context-dependent. We conclude with reflections on how to further think about the conceptualisation of interdisciplinary digital project design and the professionalism involved.

2 *Integrating interdisciplinary research*

In DH, there is a growing interest in critically reflecting on "what is happening" and "what should happen" at the intersections of data-driven and interpretative methods (Ahnert et al. 2023). However, much of the discussion has so far focused on the future potential of computational methods and productive approaches, rather than the nature of interdisciplinary syntheses and mixed methods (Brodén, Ingvarsson, et al. 2024; c.f. Shan 2021). While systematic theoretical reflection on the relationship between different interdisciplinary and methodological syntheses remain limited, there is an ongoing discussion in DH about the importance of context-sensitive approaches grounded in domain knowledge, emphasising the need to engage with the original contexts of archival data to produce both robust and nuanced results. Katherine Bode (2018), for example, critiques the reliance on data models that inadequately represent how historical texts generated meaning in their original contexts, while Jo Guldi (2023) warns against naïve assumptions about the relationship between data and the documentary record.

2.1 *Workflows and protocols*

In this context, Sarah Oberbichler et al. (2022) make an important distinction between "multidisciplinary collaboration" and "integrated interdisciplinary research" in DH. While the former seeks "to build something in between the disciplines so they share more than just the problem", the latter involves "people from different scientific fields come together, collaborate, and study a common question or problem with the goal of reaching common conclusions" (Oberbichler et al. 2022: 226). It is worth noting the long-running and often contradictory debates within academia about the specific meaning of terms such as "multidisciplinarity" and "interdisciplinarity". However, in our context, the focus lies less on the terminology and more on the character and quality of collaborative processes. Oberbichler et al. (2022: 226) emphasise that integrated interdisciplinary research "requires going deeper than just saying something about a phenomenon from different perspectives". This entails understanding how different fields operate and identifying which approaches can be productively used for problem-solving. They further argue that successful collaboration and communication require recognising and addressing the differences and commonalities between disciplines: "Merging of applications, tasks, and traditions, involving mixed method approaches as well as increased interaction between the disciplines, has been identified as a possible common objective" (Oberbichler et al. 2022: 230).

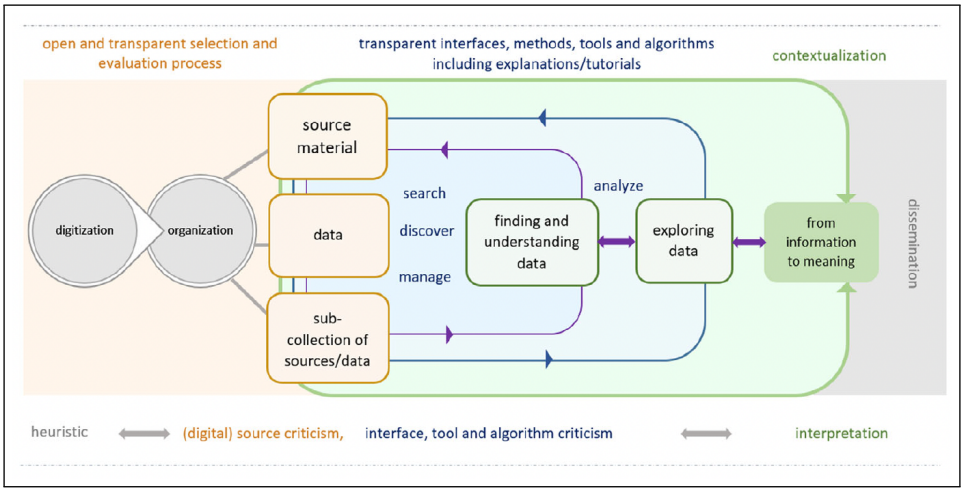


Figure 1: Interdisciplinary digital hermeneutics workflow by Oberbichler et al. (2022: 231).

To achieve an integrated interdisciplinary research process, Oberbichler et al. (2022: 231) propose a workflow (Figure 1) that highlights the importance of iterative collaborative steps between HSS researchers and data analysts. This approach facilitates a gradual deepening of understanding, exploration and analysis of the data, methods and tools in focus. In our own work developing project applications and digital resources at GRIDH, as well as within the frameworks of the research infrastructures Huminfra, InfraVis and Swe-Clarin, we have found that an integrative workflow centred on the data, largely aligned with the model proposed by Oberbichler et al., works well for cultivating dialogue between the involved parties, enabling deeper insights into the data and each other's concerns. We do not necessarily propose this model as an "all-purpose" solution. For instance, one might argue that the category of digitisation should also be included in the iterative steps of the workflow (Figure 1), as this is far from a static or fixed process (Smith & Whearty 2023). Rather, our point is to emphasise that following, or at least considering, some form of proven workflow or protocol is important in the practice of interdisciplinary digital project design. Such workflows provide a collaborative framework that helps establish common ground and a shared understanding of the various research perspectives and tasks involved.

Before proceeding, we however need to address the often overlooked issue of integrative expertise. Oberbichler et al. suggest that their model could serve as a starting point for planning future projects, describing it as

"a concept rather than a concrete architecture" (Oberbichler et al. 2022: 231). While they emphasise protocols and interdisciplinary skills – referred to as an "art of in-betweenness" – these elements are treated as "free-floating" rather than "embodied" and anchored to specific functions or roles within the model. This perspective invites a return to Andrea Hunter's (2014) concept of bridge people, who, in the context of Oberbichler et al.'s discussion, can be understood as individuals proficient in the art of in-betweenness. From the standpoint of interdisciplinary digital project design, it seems reasonable to argue that involving bridge people in every project could save significant time and energy by facilitating cross-disciplinary communication. However, it is also worth noting the somewhat static nature of the bridge people concept, which implies that these individuals naturally manifest competence in both disciplinary cultures and are capable of unpacking all critical issues. To move beyond this static view, we emphasise the need to approach both the application of protocols and the art of in-betweenness as dynamic processes, as a constant "doing" rather than a fixed "being".

2.2 *Professionalism and judgement*

To better understand the application of protocol and the art of in-betweenness in mediating different disciplinary rationales and facilitating collaboration in DH projects, we can turn briefly to the educational philosophy of renowned scholar Gert Biesta. Biesta critiques the dominant tendency in educational research to instrumentalise teaching as merely the "application of protocols", He highlights the misconception that protocols and guidelines can tell teachers "what they should do on the assumption that particular forms of research can provide clear and unambiguous knowledge about "what works"" (Biesta 2015: 81). Instead, Biesta emphasises the critical role of judgement, arguing that teachers as individual professionals contribute something essential to the educational situation. He asserts that all desirable education requires instant and spontaneous judgement about "what to do", both in general terms and in specific situations.

Applied to our context, we argue that interdisciplinary digital project design cannot be fully organised through protocol alone or based on the assumption that bridge people will naturally mediate convergence between the two cultures and facilitate integrative workflows. Instead, the process inherently relies on the enactment of a specific form of professionalism. Drawing on Biesta's conception of teaching as an act of judgement and the enactment of acquired professionalism, we suggest that the application of protocol and the art of in-betweenness similarly depend on active acts of judgement by individuals. As Biesta reminds us, it is helpful to revisit a

"classic" definition of professionalism as an approach tied to specialised and exclusive knowledge and skills (Freidson 1994). In this light, and akin to Biesta's observations on education, bridge people are continually "confronted with situations that, in some respects, are always new and hence call for judgement" (Biesta 2015: 81).

We should keep in mind that professionalism is tied to relationships of knowledge-based authority and trust. While authority can, of course, be questioned, particularly in the relationship between professionals and their "clients", it nevertheless serves as a foundation for justifying the judgements made. In the context of an integrated interdisciplinary research process, such judgments may involve determining the balance between different steps in a workflow, identifying aligned goals and deciding what needs to be done (and not done) at certain stages. The art of in-betweenness, therefore, requires an acknowledgement of the role and responsibilities of the practitioner. As Biesta observes, the essence of professional practices lies not only in addressing the needs of their clients, but also in playing "a crucial role in the definition of those needs" (Biesta 2015: 82). This highlights the active nature of professionalism in shaping the processes and outcomes of interdisciplinary collaboration.

3 *In practice – case studies*

We now turn to five brief case studies that illustrate various aspects of what we term interdisciplinary digital project design. These examples focus primarily on the enactment of protocol and judgement during the early stages of different projects, offering different perspectives on the integrative professionalism involved.

3.1 *Identifying research possibilities: Conspiracy soundtracks*

The "Role of Music and Sound in Audiovisually Mediated Conspiracy Theoretical Discourse" project (henceforth referred to as the Conspiracy Soundtracks project) illustrates how GRIDH contributes expertise to the development of research applications. Established in 2015 as the Centre for Digital Humanities (CDH), GRIDH was initially tasked with supporting research project applications at the Faculty of Humanities (later on, GRIDH expanded to a university-wide infrastructure). In this case, Musicology scholar Tobias Pontara (Department of Cultural Sciences, University of Gothenburg), the eventual principal investigator of the project, approached us for input on a project application to study the role of soundtrack elements – sound effects,

music, voice and silence – in conspiracy films distributed online. The "DH dimension" of the project was initially limited to analysing these elements qualitatively to uncover "hidden" discursive patterns.

During the initial meeting about the application between Pontara, a GRIDH research engineer (Victor Wåhlstrand Skärström) and two of the chapter co-authors (Brodén and Fridlund), we found a potential to extend the project with a quantitative dimension. This involved employing media visualisation techniques to display recurring sound effects and musical patterns within and across films, that is a form of "distant reading" of the material's aural elements. While this approach was merely tentative at the time, our discussion informed the methodological framework of the eventually approved application. Input from Brodén, drawing on a background in film studies and an interest in the works of digital media scholar Lev Manovich, proved useful, as Manovich's (2020) studies on visualising audiovisual data became part of the project's theoretical foundation. This example highlights how GRIDH's contributions extend beyond simply providing technical input for a project application to include translating and refining the client's research ideas into a developed project design.

The Conspiracy Soundtracks project also demonstrates how protocol and judgment frequently intertwine. As indicated by the development process, it is often challenging to distinguish between actions driven by protocol and those requiring judgment. This "impurity" in the process reflects the inherent complexity of interdisciplinary collaboration. For instance, our input on the methodological relevance of Manovich's work raises questions about whether it represents the art of in-betweenness or constitutes domain-specific expertise outside the scope of the specific professionalism discussed in this chapter.

3.2 *"Big picture" meets domain expertise: INTRAPARTY*

For another example of how we collaborate with researchers, this time in the development of digital resources, we turn to "The Benefits of Conflict" (INTRAPARTY) project, a European Research Council (ERC) Starting Grant project in Political Science focusing on ideological factions within European political parties. INTRAPARTY uses, among other things, text data of motions at party congresses in Germany, the Netherlands and the United Kingdom. Initially, the principal investigator Ann-Kristin Kölln (Dept of Political Science, University of Gothenburg) approached GRIDH for potential support in text mining these motions, but our discussions expanded into a broader collaboration.

During our (Brodén, Fridlund and research engineer Johan Åhlfeldt) first meetings with Kölln, it became clear that the text mining aspect of the project raised substantial issues that needed to be addressed, including issues related to digitisation, OCR quality, and annotation. This highlights a recurring aspect of interdisciplinary DH projects: addressing not only how to translate a research question into a DH context but also identifying the broader steps necessary for the research process. In this case, the complexity of the workflow exceeded initial expectations. For instance, our discussions revealed the need for an OCR quality review and the implementation of a separate annotation process, as the required text and metadata could not be easily extracted automatically in a satisfactory way. Recognising the importance of involving technical expertise early on, subsequent meetings focused on database issues. Eventually, GRIDH's role evolved into a contract-based development of two digital resources: a platform for storing and annotating party motions (titled "Factions") and a Python application ("Moxy", Motion Extraction with YAML Configurations) for extracting structured information from datasets for text analysis.

While the technical development of these resources was led by GRIDH research engineers (main developer David Alfter, with frontend contributions by Kristin Åkerlund), we remained involved in coordinating meetings and providing input on various issues. Among other things, having hands-on experience of annotation, we contributed to discussions on document handling and OCR challenges. Additionally, drawing on our work with parliamentary data in the SweTerror project (see below), we were able to provide input on how to approach and to structure the party motion data for immediate and future research needs. Similar to the Conspiracy Soundtracks project, our contributions extended beyond bridging roles to include some domain-specific insights. It is worth noting, however, that the INTRAPARTY project has not been an interdisciplinary endeavor in a more integrated sense, as the resources developed by GRIDH were primarily tools for use by the Political Science team. Nonetheless, this case illustrates how interdisciplinary digital project design encompasses something more than facilitating straightforward collaboration between HSS and technical expertise. The interdisciplinary element can take various forms, encompassing contributions shaped by interdisciplinary experience and, in some cases, aspects of our own research.

3.3 *Digitisation as a dialogical process: The inscriptions of Saint Sophia*

In 2023, GRIDH (Westin) was approached about putting together an expedition to Ukraine to digitise (or, in our terminology, "collect data from") the

inscriptions of Saint Sophia Cathedral in Kyiv. The first step involved assembling an interdisciplinary team of domain experts to define specifications for the data collection, drawing on an approach to digitisation and data collection that GRIDH has developed over time. For instance, in 2022, GRIDH led the project "Methods for Digital Diagnosis of Threatened Cultural Heritage", funded by the Swedish National Heritage Board. Using the digital documentation of remains from the Swedish Antarctica Expedition (1901–1903) on the Antarctic Peninsula as a case study (part of the CHAQ2020 expedition), the project explored the need for cross-disciplinary documentation practices. Through workshops involving experts from Archaeology, Glaciology, Geology, Photography, Architecture, Conservation, Biology, Zoology and Physics, a set of principles and best practices was developed to guide future digitisation efforts (Westin & Almevik 2024).

The project emphasized that terms like "documentation" and "digitisation" can imply an unrealistic sense of completeness, ignoring the inherently selective and interpretive nature of these processes. Technologies and workflows shape what aspects of material are preserved or discarded, influencing future analyses and workflows (Westin 2021). Consequently, the team adopted the term "data collection" to better reflect the selective nature of acquiring digital material (cf. Dahlström 2010). This framing emphasises the need to critically assess what data to collect, the requirements of the disciplines it will serve, and the quality of the resulting data. This critical perspective also informs the workflow proposed by Oberbichler et al. (Figure 1), introducing an analytical feedback loop that connects the initial data collection with the subsequent analytical processes. This iterative approach ensures that the data collected is guided not only by digital methods but also by domain-specific knowledge. For instance, the Antarctic project incorporated expertise from over eight disciplines, each contributing distinct methods and analytical frameworks.

Applying these principles to The Inscriptions of Saint Sophia project, interdisciplinary dialogue helped identify several aspects of the inscriptions that would otherwise have been overlooked. These included spatial context, depth information for inscriptions invisible to the naked eye, capturing entire surfaces rather than isolated inscriptions, and employing layered, overlapping technologies to support diverse workflows of source evaluation. The process evolved through iterative improvements to both the data collection and the tools and interfaces being developed. After an initial, limited data collection in Ukraine, which tested various techniques and technologies, a series of workshops were held in Sweden and Ukraine. These workshops brought together heritage workers and researchers performing the data collection, research engineers developing tools and workflows, and domain

experts. The workshops evaluated the initial results, allowed domain experts to suggest refinements and facilitated collaboration with research engineers to set goals for the tools under development. This iterative process ensured that the collected data better aligned with the needs of the interdisciplinary team while advancing the tools to meet project goals.

3.4 *Interdisciplinary team and research questions: SweTerror*

The "Terrorism in Swedish Politics" project (SweTerror, 2021–2026; Edlund et al. 2022) originated from a 2020 DIGARV (Digitisation and Accessibility to Cultural Heritage Collections) call for data-rich HSS research using digitised materials from Swedish GLAM institutions. SweTerror brings together an interdisciplinary team from DH, Terrorism Studies, History of Ideas, Linguistics, Political Science, and Speech and Language Technology to conduct a comprehensive mixed-methods study of political discourse on terrorism, using text transcripts and audio recordings of Swedish parliamentary debates (1968–2018). In its assessment, the Swedish Research Council emphasised the project's "remarkably interdisciplinary team".[1]

The interdisciplinary design of SweTerror began with core project partners (Brodén, Fridlund and Speech Technology scholar Jens Edlund, Division of Speech, Music and Hearing, KTH Royal Institute of Technology) collaboratively defining the project's scope in relation to the parliamentary data, partly in line with the data-centered workflow in Figure 1. However, as in other case studies, the enactment of protocol and judgment frequently overlapped. For instance, assembling an interdisciplinary team of HSS and Language Technology researchers required relying on judgment to determine the expertise necessary to address the research questions, rather than following any protocol.

A significant part of the project development involved negotiating research questions through iterative dialogue among team members. These questions emerged as a compromise between the analytical interests of HSS researchers and the technical feasibility of data analysis proposed by Speech and Language Technology researchers. As bridge people, we (Brodén and Fridlund) were able to mediate between these disciplinary perspectives. The aim was to prioritise problem-solving over disciplinary boundaries, creating a collegial atmosphere and building trust in collective decision-making. This approach influenced decisions about trade-offs, the use of domain-specific concepts and the precise framing of research questions.

1 Cited from the review panel's statement for "Terrorism in Swedish politics (SweTerror): a multimodal study of the configuration of terrorism in parliamentary debates, legislation, and policy networks in Sweden 1968–2018", Swedish Research Council, ref. nr. 2020-05052.

After the project's approval, further adjustments were required to align collaborative tasks with the evolving research questions. For instance, co-writing conference papers depended on the specific data that analysts could extract at different stages of curation. This necessitated ongoing interdisciplinary dialogue as well as judgments about the level of data refinement needed to meet the methodological standards of various disciplines. In some cases, this also involved recalibrating the research questions to align with the available data, ensuring valid conclusions about parliamentary discourse on terrorism.

3.5 *Multidisciplinary structures and computational collaborations: Carvers on the rocks*

The "Tracing the Carvers on the Rocks" project (2021–2026) also originated from the 2020 DIGARV call (see above) and aims to differentiate among carving techniques in Nordic Bronze Age rock art (1700–500 BC). It brings together scholars from Archaeology, Structural Geology, Cultural Heritage experts and DH researchers. The project builds on GRIDH's collaboration with the Swedish Rock Art Research Archive (SHFA), initiated when GRIDH sought to apply AI and image analysis but lacked relevant data. Fortunately, SHFA was keen to start a collaboration and had already begun considering similar questions.

Bronze Age rock art has been studied in relation to ideology, religion, long-distance trade, warfare, landscapes and social organisation, but the techniques used to create the art remain underexplored. The Tracing the Carvers project seeks to deepen our understanding of Bronze Age carving techniques by training AI algorithms to detect and analyse motifs, stylistic variations, superimpositions and the structures of granite rock and grain. The project also focuses on the tools, techniques and bedrock selection employed by the carvers. While the interdisciplinary approach has produced new insights (Liebl et al. 2023), the GRIDH team member (Lindhé) also encountered challenges in fostering integration across disciplines, not least for the two PhD candidates in the project. Questions arose about how to create an environment that accommodates diverse intellectual traditions while maintaining strong disciplinary foundations. Sustaining interdisciplinarity rather than siloed work highlighted the need for dedicated expertise and time to address these issues. To provide a reflexive perspective on these efforts, the project would also have been enhanced by establishing a protocol for workflow-oriented techniques at the outset.

Ideally, DH constitutes an "organic" collaboration between the Humanities and Computer Sciences. Willard McCarthy (2012) asserts that this

meeting place ought to be a "level ground" of reciprocal collaboration. However, in Tracing the Carvers, humanities scholars were sometimes limited by the way computational possibilities were presented. This underscored the importance of humanities researchers articulating their needs clearly to influence computational processes effectively. A robust disciplinary foundation is thus essential for advancing arguments to both the data scientist and, by extension, the capacity of the machine. The project also demonstrated the value of individuals with in-betweenness skills to bridge disciplinary divides and foster dialogue within the project team.

Given its broad scope, the project underscores the importance of individuals who can bridge different disciplines *as well as* individuals who have expertise in multiple areas. Moreover, it is imperative to think of "collaboration" in terms of not only joint efforts by individuals from various disciplines, but also between humans and machines in the sense that "an understanding of machines as collaborators in knowledge production, and an awareness of the impacts of materiality on such production, becomes a disciplinary as well as philosophical concern" (Griffin & Hayler 2018: 28). Building on this perspective, the team came to recognise how materials, algorithms and computational tools in a way serve as active "participants" in their workflows (Lindhé et al. 2016).

4 *Conclusions*

In this chapter we have outlined core elements of an interdisciplinary-oriented expertise that warrants greater attention in discussions about developing projects involving multidisciplinary teamwork. We proposed the concept of "interdisciplinary digital project design" to emphasise a professionalism focused on integrating collaboration between HSS researchers and technical experts when creating project applications, digital resources and more that meet the current standards in the field. We argued that this expertise involves adherence to protocols and established procedures, such as those discussed by Oberbichler et al. (2022), to ensure consistency and quality in collaborative efforts. At the same time, as argued by (Biesta 2015), it relies on the exercise of judgment of individuals deciding "what needs to be done and how" in specific contexts.

To illustrate the practice of interdisciplinary digital project design, we discussed five projects involving GRIDH, each showcasing different aspects of integrative professionalism, or the art of in-betweenness, and the interplay between protocol and judgment. The Conspiracy Soundtracks project demonstrated GRIDH's approach to supporting DH project applications

at the Faculty of Humanities, highlighting the complexity of facilitating interdisciplinary communication beyond a narrow sense. The INTRAPARTY project underscored the importance of translating HSS research questions into DH contexts, seeing the "big picture" and identifying key steps in the research process but also offering insights based on prior work with similar data. In the Saint Sophia project, we emphasized how digitisation serves as a focal point for iterative collaboration between domain experts, researchers, and engineers, ensuring that data collection processes and tools align with disciplinary needs. The SweTerror project illustrated the importance of building trust in judgments, especially when balancing domain-specific interests among team members. Finally, the Tracing the Carvers project highlighted the complexities of maintaining interdisciplinary collaboration and the need to integrate both human and machine agencies into research processes.

While we have addressed some core elements of interdisciplinary digital project design through these examples, the understanding of the professionalism and the practice involved still needs to be further elaborated. For instance, since the notion of bridge people concerns both those who are either fully competent in the two cultures or only bilingual, more elaboration is required on how this relates to the workflows and professionalism discussed in this chapter. We have also used the term "technical expertise" in a rather monolithic way, whereas in practice this category is often inhabited by people with different interdisciplinary experiences and sensibilities. Just as we as Humanities researchers through our interdisciplinary expertise are more or less bilingual in the two cultures, the same applies to data analysts, who can be partly or fully "fluent" in HSS research concerns. Thus, there is, among other things, a need to deepen the discussion about the distribution of the professionalism that we have tried to describe among different individuals and institutions in the field of DH.

Moreover, we need to put further consideration into the notion of "research team", since technical expertise sometimes contributes to DH projects primarily in the role of "computational facilitators" rather than researchers. The continuous possibility that, for instance, deadlines are not met or technical experts would leave for another position not only makes DH projects vulnerable in general, but also calls for conceiving interdisciplinary teamwork as something potentially less fixed, adding another forward looking factor and element of flexibility to the practice of interdisciplinary digital research design.

Finally, we must address the authority tied to the expertise discussed in this chapter. While our authority as professionals can and should be questioned, it nevertheless provides a basis for justifying judgments. The art of in-betweenness requires affirmation that it not only responds to the needs

of clients, but also helps define those needs. Drawing on Biesta (2015: 84), research design entails an inherent "risk": clients may reject our expertise, challenging its necessity. Reducing this risk by presenting research design as purely protocol-driven would undermine its essence, reducing it to an instrumental practice. Accepting that our authority may always remain in flux is a risk we must be prepared to embrace.

Hands-on Practical Guidelines

In addition to emphasising the value of further developing the concept of interdisciplinary digital project design, we would like to offer some general reflections on how to approach the associated professionalism in practice, aligning with the broader structure of this handbook. While it is not possible to provide detailed practical guidelines in the same way as other chapters focused on digital tools and resources, we aim to share a few key takeaways from our experience in cultivating the professionalism discussed in this chapter:

- **Cultivate interdisciplinarity**: actively seek to broaden your knowledge and engage with perspectives outside your discipline.
- **Embrace collaboration**: look for opportunities to work with scholars from other fields and disciplines.
- **Step beyond your "academic comfort zone"**: challenge yourself to engage with unfamiliar methodologies, theories, and frameworks.
- **Publish across disciplines**: share your research in venues beyond your primary field to engage with different academic traditions.
- **Foster open and humble dialogue**: approach conversations with scholars from other fields with humility, respect and a willingness to learn.
- **Think beyond technical aspects**: when dealing with digitisation and computation, focus not only on technical processes but also on the research questions the data can support.
- **Adopt a critical professionalism**: consider your bridging role between disciplines as a form of critical, reflective professionalism.
- **Follow protocol, but know when to break it**: adhere to established workflows where appropriate, but remain flexible and adapt when circumstances demand it.

- **Balance trust and skepticism in your judgment**: trust your expertise, but remain open to questioning your assumptions and decisions.
- **Be forward-looking and flexible**: anticipate unexpected challenges, including broken deadlines and deviations from plans and adapt accordingly.
- **Engage with interdisciplinary discourse**: participate in ongoing discussions about interdisciplinary collaboration and mixed methods in DH.

Acknowledgements

This chapter extends a short paper presented at the Huminfra Conference 2024 (HiC 2024; Brodén, Fridlund, et al. 2024). Importantly, it draws on collaborative work conducted with our colleagues at GRIDH and the projects referenced throughout: Conspiracy Soundtracks, INTRAPARTY, The Inscriptions of Saint Sophia, SweTerror, and Tracing the Carvers on the Rocks. The chapter is also informed by our involvement in the national research infrastructures Huminfra (funded by VR, contract nos. 2021-00176, 2023-00171), InfraVis (funded by VR), and Swe-Clarin (funded by VR, contract no. 2017–00626).

References

Ahnert, Ruth, Emma Griffin, Mia Ridge & Giorgia Tolfo. 2023. *Collaborative historical research in the age of big data*. Cambridge: Cambridge University Press. DOI: 10.1017/9781009175548.

Berry, David & Anders Fagerjord. 2017. *Digital humanities: Knowledge and critique in a digital age*. London: Polity. DOI: 10.5555/3161393.

Biesta, Gert. 2015. What is education for? On good education, teacher judgement and educational professionalism. *European Journal of Education* 50(1). 75–87. DOI: 10.1111/ejed.12109.

Bode, Katherine. 2018. *A world of fiction. digital collections and the future of literary history*. Ann Arbor, MI: University of Michigan Press. https://www.jstor.org/stable/j.ctvdtpj1d.

Brodén, Daniel, Mats Fridlund & Cecilia Lindhé. 2024. Designing digitally-driven integrative interdisciplinarity: Professionalism between protocol and judgement. In *Proceedings of the Huminfra conference (HiC2024)*. DOI: 10.3384/ecp205017.

Brodén, Daniel, Jonas Ingvarsson, Lina Samuelsson & Victor Wåhlstrand Skärström. 2024. Visualization as defamiliarization: Mixed methods approaches to historical book reviews. *Journal of Computational Literary Studies* 3(1). 1–26. DOI: 10.48694/jcls.3926.

Cross, Nigel. 2011. *Design thinking: Understanding how designers think and work*. Oxford: Berg. DOI: 10.5040/9781474293884.

Dahlström, Mats. 2010. Critical editing and critical digitisation. In Willem Peursen, Ernst Thoutenhoofd & Adriaan Weel (eds.), *Text comparison and digital creativity*, 79–97. Leiden: Brill. DOI: 10.1163/ej.9789004188655.i-328.29.

Edlund, Jens, Daniel Brodén, Mats Fridlund, Cecilia Lindhé, Leif-Jöran Olsson, Magnus P. Ängsal & Patrik Öhberg. 2022. A multimodal digital humanities study of terrorism in Swedish politics: An interdisciplinary mixed methods project on the configuration of terrorism in parliamentary debates, legislation, and policy networks 1968–2018. In Kohei Arai (ed.), *Intelligent systems and applications*, 435–449. Cham: Springer International Publishing. DOI: 10.1007/978-3-030-82196-8_32.

Fickers, Andreas, Juliane Tatarinov & Tim van der Heijden. 2022. Digital history and hermeneutics – between theory and practice: An introduction. In *Digital history and hermeneutics: Between theory and practice*. Andreas Fickers & Juliane Tatarinov (eds.). Berlin, Boston: De Gruyter Oldenbourg. 1–20. DOI: 10.1515/9783110723991-001.

Freidson, Eliot. 1994. *Professionalism reborn: Theory, prophecy, and policy*. Chicago: University of Chicago Press.

Griffin, Gabriele. & Matt Hayler. 2018. Collaboration in digital humanities research: Persisting silences. *Digital Humanities Quarterly* 12(1).

Guldi, Jo. 2023. *The dangerous art of text mining: A methodology for digital history*. Cambridge: Cambridge University Press. DOI: 10.1017/9781009263016.

Hunter, Andrea. 2014. Digital humanities as third culture. *MedieKultur* 57. 18–33.

Ingvarsson, Jonas. 2021. *Towards a digital epistemology: Aesthetics and modes of thought in early modernity and the present age*. Cham: Palgrave Macmillan. DOI: 10.1007/978-3-030-78724-0.

Liebl, Carina, Mark Peternell, Johan Ling, Cecilia Lindhé, Christian Horn, Ellen Meijer & Julian Moyano. 2023. Tracing carvers on the rocks. *Adoranten* 54. 113–125.

Lindhé, Cecilia, Ann-Catrine Eriksson, Jim Robertsson & Mattis Lindmark. 2016. Curating Mary digitally: Digital methodologies and representations of medieval material culture. In Matt Hayler & Gabriele Griffin (eds.), *Research methods for creating and curating data in the digital humanities*. Edinburgh: Edinburgh University Press. DOI: 10.1515/9781474409667-008.

Manovich, Lev. 2020. *Cultural analytics*. Boston, Mass: MIT Press. DOI: 10.7551/mitpress/11214.001.0001.

McCarthy, Willard. 2012. Collaborative research in the humanities. In Marilyn Deegan & Willard McCarthy (eds.), *Collaborative research in the digital humanities*, 1–10. Farnham, Surrey.

Merriam-Webster. N.d. Design. In *The merriam-webster.com dictionary*. https://www.merriam-webster.com/dictionary/design, consulted 9 May 2025.

Oberbichler, Sarah, Emanuela Boroş, Antoine Doucet, Jani Marjanen, Eva Pfanzelter, Juha Rautiainen, Hannu Toivonen & Mikko Tolonen. 2022. Integrated interdisciplinary workflows for research on historical newspapers: Perspectives from humanities scholars, computer scientists, and librarians. *Journal of the Association for Information Science and Technology* 73(2). 225–239. DOI: 10.1002/asi.24565.

Rockwell, Geoffrey. & Stéfan Sinclair. 2016. *Hermeneutica: Computer-assisted interpretation in the humanities*. Boston, Mass: MIT Press. DOI: 10.7551/mitpress/9522.001.0001.

Rowe, Peter. 1987. *Design thinking*. Cambridge, Mass: MIT Press.

Shan, Yafeng. 2021. Philosophical foundations of mixed methods research. *Philosophy Compass* 17(1). DOI: 10.1111/phc3.12804.

Smith, Astrid. & Bridget. Whearty. 2023. All the work you do not see: Labor, digitizers and the foundations of digital humanities. In Matthew. Gold & Lauren. Klein (eds.), *Debates in the digital humanities*. Minneapolis: University of Minnesota Press.

Snow, C.P. 1959. *The two cultures and the scientific revolution*. Cambridge University Press.

Svensson, Patrik. 2012. The digital humanities as a humanities project. *Arts and Humanities in Higher Education* 11(1-2). 42–60. DOI: 10.1177/1474022211427367.

Underwood, Ted. 2014. Theorising research practices that we forgot to theorize twenty years ago. *Representations* 127(1). 64–72. DOI: 10.1525/rep.2014.127.1.64.

Westin, Jonathan. 2021. Arosenius translated: Digitisation as a rephrasing of meaning. *The Journal of Nordic Museology* 13(1). 30–55. DOI: 10.5617/nm.8823.

Westin, Jonathan & Gunnar Almevik. 2024. Digitising sensitive heritage monuments in Antarctica. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLVIII-2/W4-2024. 461–467. DOI: 10.5194/isprs-archives-XLVIII-2-W4-2024-461-2024.

Corresponding authors

Daniel Brodén
Gothenburg Research Infrastructure in Digital Humanities
University of Gothenburg
daniel.broden@lir.gu.se

Mats Fridlund
Gothenburg Research Infrastructure in Digital Humanities
University of Gothenburg
mats.fridlund@lir.gu.se

Cecilia Lindhé
Gothenburg Research Infrastructure in Digital Humanities
University of Gothenburg
cecilia.lindhe@lir.gu.se

Jonathan Westin
Gothenburg Research Infrastructure in Digital Humanities
University of Gothenburg
jonathan.westin@lir.gu.se

4 Applied NLP for humanities research

Gijs Aangenendt
Uppsala University

Maria Skeppstedt
Stockholm University

Karl Berglund
Uppsala University

Natural language processing (NLP) has become a field of interest for many researchers within the humanities. However, framing humanities research questions as NLP problems and identifying suitable methods can be a difficult task. Taking previous and ongoing projects from the Centre for Digital Humanities and Social Sciences at Uppsala University (CDHU) as a point of departure, this chapter presents concrete use cases of how humanities research questions can be approached using various NLP methods and tools, from ready-to-use text analysis tools to programming libraries that require basic familiarity with Python. Two case studies from the field of history and literature will be introduced to illuminate how texts can be processed for humanities research purposes. With this chapter, we hope to give the reader the means to directly explore NLP methods for their research as well as encourage further learning.

1 *Introduction*

With the increasing availability of texts in digital form, natural language processing (NLP) has become a field of interest for many researchers within the humanities. Researchers turn to NLP to extract information from collections of texts that are too large to manually analyze. This promise of "distant reading", to use the term coined by Franco Moretti, is one of the main attractions of NLP for the humanities (Moretti 2000).

Despite its potential, utilizing concepts and methods from the technical-linguistic field of NLP can be challenging. Framing humanities research questions as NLP problems, identifying suitable methods, evaluating their performance, and finally building arguments based on the output is not straightforward due to fundamental differences between the domains (Mcgillivray et al. 2020). When starting to engage with NLP, scholars often need to learn

HUM INFRA Gijs Aangenendt, Maria Skeppstedt & Karl Berglund. 2025. Applied NLP for humanities research. In Gerlof Bouma, Dana Dannélls, Dimitrios Kokkinakis & Elena Volodina (eds.), *Huminfra handbook: Empowering digital and experimental humanities* (NEALT Proceedings Series 59), 77–111. University of Tartu Library. DOI: 10.58009/aere-perennius0173
© The authors, CC BY 4.0

several new things at once, including technical skills (e.g., basic programming and data wrangling), statistical and mathematical concepts (e.g, "what is standard deviation?", "what is a vector?"), and – not least, but often forgotten – linguistic concepts (e.g., "what is a lemma and why does it matter?"). Resources for learning and applying NLP, however, are often technical or linguistic in nature and may not depart from a humanities point of view or be tailored to the kinds of sources commonly used in the humanities. Published research from the humanities that utilizes NLP methods may have to prioritize presenting arguments over showcasing the technical process in detail due to space constraints. As a result, learning from these studies and gaining insight into how the texts were processed becomes more difficult.

This chapter offers humanities scholars curious about the possibilities of NLP in their research a first introduction to the field. It begins with a broad overview of NLP in the humanities, including commonly used NLP methods as well as available resources and tools. Following this, two case studies present concrete examples of how research questions, from the fields of history and literature specifically, can be approached with the help of NLP. The history case study employs a ready-to-use text analysis tool, while the literature case study utilizes programming libraries that require basic familiarity with Python. With this chapter, we hope to give readers with varying levels of technical expertise the ability to experiment with NLP in their own research. Notebooks and code examples connected to this chapter are available in a GitHub repository.[1]

2 *NLP and the humanities*

Within NLP, language is analyzed in various ways, from how it sounds and is spoken (phonetics and phonology), how words and sentences are structured (morphology and syntax) to the meaning of words, word combinations and expressions (semantics) (McGillivray 2022). These approaches to language have been adopted to support humanities-based research in a variety of tasks, ranging from tracking semantic change of words to extracting topics from a collection of texts. In this chapter, we adopt a broad definition of NLP that encompasses any attempt to analyze natural language with the help of computers (Schofield 2022).

When humanities researchers use NLP, they do not always directly refer to it as such. Instead they may use terms like "distant reading" (Moretti 2013, Jänicke et al. 2015, Drucker 2017) or "text mining" (Jockers & Underwood 2015, Thompson et al. 2016), which are more closely tied to the humanities

1 https://github.com/CDHUppsala/Applied-NLP-for-humanities-research

field and its practices. There are also subfields such as "digital history" (Fridlund et al. 2020) or "computational literary studies" (Bode 2023) in which NLP methodologies are adapted to support humanities research tasks. This indicates that NLP is often not a goal in itself, but rather a means to an end. Humanities scholars use computational methods from the field of NLP in the hope of gaining new insights into a body of texts, supplementing knowledge gathered by means of traditional humanistic close reading. NLP analyses can provide meaning for the humanities in multiple ways: by detecting patterns and trends in large collections of text, but also identifying outliers that break the pattern. Humanities NLP work often combines quantitative analysis of trends with qualitative analysis of outliers, which may help explain the trends better (So 2017, Underwood 2019).

There is a wide range of NLP methods that can be used for humanities research purposes, making it difficult to choose which method is most suitable. Selecting methods can depend on the research question, the phase of the research process, and the material at hand. Choosing the method also depends on the level of transparency and explainability needed, which is especially relevant when it comes to communicating research results to peers. Results from word frequency-based methods are easier to interpret, validate, and explain than results derived from complex machine learning algorithms. A guiding principle – and here we follow de Bolla et al. (2019) – could be to keep the method *as simple as possible*, not least for the sake of transparency. If no one in your specific humanities field understands what you have done, the research will likely have less impact than if the methods deployed are more intuitively graspable.

When looking for relevant methods and tools from the field of NLP, or any other field for that matter, it is important to be mindful of potential differences compared to the humanities domain. Within NLP, researchers often aim at improving the state-of-the-art and advancing the field, trying to develop models that perform better than its predecessors on gold standard corpora. These corpora have been (at least to some degree) consistently annotated by two or more annotators to ensure reliability in evaluating model perfomances (Artstein & Poesio 2008). The performance of models is typically measured by precision, recall, and the F1-score, the harmonic mean of precision and recall. These evaluation metrics make it easier to compare different models.

Humanities researchers, on the other hand, are interested in whether the NLP method contributes to answering their research questions. The choice of method might therefore not primarily depend on the state-of-the-art and reported F1-scores. Evaluating the performance of the method may be done by examining the output qualitatively. Not in the least considering that

incorrect predictions can still be valuable and provide new directions for further qualitative analysis (Rettberg 2022). Additional factors for deciding which method to use can be 1) the explainability of the method - how well is the model able to show what features were used to make predictions, 2) how well the model is able to state the confidence with which it makes its predictions, and 3) how well the method performs on other corpora (or how easily it can be adapted to other corpora) than the one used for developing the method.

Whether methods can be applied to other types of corpora is especially relevant as humanities researchers tend to make use of project-specific corpora. Most NLP models are trained on specific datasets from a certain language – often English – as well as a particular period and domain, for instance Wikipedia pages or contemporary news articles. Consequently, these models do not necessarily perform well on other corpora (Plank 2016, Bamman 2017), including those typically used within the humanities with messy, historical, domain-specific, or odd and experimental language. Tasks considered 'solved' in NLP may therefore still be difficult to implement and require additional research to develop the best approach for the specific time period, domain, and language of the materials (Mcgillivray et al. 2020).

Finally, it is worth mentioning the importance of data acquisition. Depending on the types of sources relevant to the project, this may involve scanning documents and applying optical character recognition (OCR) or handwritten text recognition (HTR), web scraping, or transcribing audio files using speech-to-text techniques. The quality of the data gathering influences the subsequent application of NLP methods. When digitizing text materials, achieving high-quality OCR or HTR is crucial, as the performance of NLP tasks can drop when accuracy falls below 80 or even 90 percent (van Strien et al. 2020). Historical spelling, fonts, and scripts, as well as poor print or scan quality, can complicate achieving high-quality textual data. For born-digital text materials, problems often concern access, copyright, and data cleaning of various sorts.

3 *NLP methods in humanities research*

Once a corpus, or collection of texts, has been acquired, there are, broadly speaking, two main approaches for how one could study its content. In scenarios where the researcher knows in advance what they want to find or extract from the corpus, a *supervised* approach can be taken. In this case, the intended research approach must somehow be translated or transferred into an NLP solution. This translation may involve creating a list of search

terms to retrieve relevant texts or implementing a *rule-based system* – for example a classic expert system (Russell et al. 1995), where explicit logic or programming rules determine how the texts are processed. *Supervised machine learning models*, in contrast, learn by example. Here, the research approach is transferred through annotated example data, which the model then uses to learn to automatically produce similar annotations on previously unseen texts. Finally, more recent developments include *prompt engineering* (Hu et al. 2024). Similar to rule-based systems, the research approach is transferred through explicit instructions. However, instead of rules, these instructions are formulated in natural language to be interpreted by a large language model (and in practice annotated examples can also be provided as part of the instructions).

When a researcher does not have a predefined approach for how to study the texts, a more exploratory *unsupervised* approach may be taken. The aim is then to extract interesting information based on the content of the texts themselves. This could, for instance, be information retrieved with the help of corpus linguistics methods, such as identifying the most frequent words or examining differences in word usage between texts. Another option is to use NLP methods that, in an unsupervised fashion, extract categories from the text itself, e.g., based on word co-occurrence patterns. Unsupervised NLP methods require no rules or annotated data as input from the researcher. Instead, the bulk of the work consists of preprocessing the texts, with the aim of preserving the meaningful content, followed by interpreting the output. In practice, supervised and unsupervised approaches are often combined.

In this section, we present a number of NLP methods that are commonly utilized for humanities research purposes. As examples, we draw from a pool of past and ongoing projects conducted in collaboration with the Centre for Digital Humanities and Social Sciences at Uppsala University (CDHU). These projects form a set of concrete and realistic example cases of how NLP and corpus linguistics are used by researchers in practice. The examples are either drawn from longer research projects or from a pool of shorter pilots, which have been part of the centre's pilot project program. By briefly describing these examples, we aim to introduce and exemplify some of the key concepts required to understand the field of NLP, and also to provide a practical perspective on how NLP can be used in humanities research.

3.1 *Word frequencies*

Many projects in our example pool employ methods from corpus linguistics to count word frequencies in the studied text collections. As mentioned above, two main approaches can be identified: 1) a supervised approach,

if the words of interest are known beforehand, or 2) an unsupervised exploratory approach for the automatic extraction of potentially interesting words. It could, of course, also be relevant to apply both of these approaches to the same text collection. The three projects discussed below all study word frequency change in text collections that span longer periods of time. It is, however, also common to study word frequencies on synchronic text collections or in cases where the temporal dimension is not relevant for the aims of the study, for example, when comparing texts across genres or authors.

The first example concerns a pilot project on the transformation of language in art criticism. Word frequency methods were applied on a collection of texts on Beethoven's piano sonatas from English newspapers and periodicals dating from the 1850s onward. The aim was to analyze the use of descriptive and evaluative terms over time (Dubremetz 2023). Since the researchers already knew which words they were interested in, a predefined set of keywords was used, containing terms like ”passionate”, ”romantic”, ”beautiful” and ”expressive”. A timeline based on the frequencies (normalized for the size of the text collection also changing over time) visualized the variation and patterns in the usage of these terms over time.

A second example comes from a pilot project that studied the career of the term 'Alt-Right' in Twitter posts from 2008 to 2020 (Ekeman 2023, Piqueras 2023). Using word frequency methods, this pilot aimed to determine when the term began to be used frequently on Twitter, how its usages varied over time, and how the content of the tweets containing the term evolved over time. Similar to the previous example, the content of the Twitter posts was analyzed through the temporal frequency changes of a set of predefined keywords, referring to names, ideologies, movements, events, and activities.

In addition, the most important words in tweets containing the term “Alt-Right”, after lemmatization and stop word removal, were automatically extracted with the word importance measure *TF-IDF* to find out how the content of the tweets varied. TF-IDF stands for “term frequency, inverse document frequency”, and it normalizes the term frequency by the number of documents in which the term appears (where document can mean, e.g., a text or time span, depending on the relevant unit for comparison). The intuition behind TF-IDF is that words appearing in many documents or time spans are often less informative (Liu 2011: 189). These words are therefore down-weighted, while the more informative words that appear in fewer documents receive a higher TF-IDF score.

The third example examined relative word frequency differences across four decades (1950s to 1980s) in the journal Allergia, published by the Swedish Asthma and Allergy Association (Söderfeldt et al. 2019). By study-

ing changes in word frequency over time, the researchers aimed to identify vocabulary shifts in the journal that could reflect transformations within the field of allergy. In the first part of the study, the relative frequency of a set of predefined keywords related to contact allergy (e.g., laundry/dish-washing soap, shampoo), including compound words (machine dishwasher soap), was measured across the four decades studied. After comparing the results with a close reading of the periodical, the researchers discovered that the predefined keywords did not capture the discourse on contact allergies, missing terms related to specific materials and substances (e.g., glue, rubber, turpentine, sodium lauryl). Recognizing this limitation, a second approach was devised where the statistical measure *log-likelihood* was used to extract over-represented words for each decade relative to their frequency in the entire corpus. Instead of relying on a predefined list of keywords, this method allowed interesting words to be extracted from the corpus itself.

3.2 *Word contexts*

A possible next step in investigating the content of a collection of texts, is to not limit the study to the words themselves, but to also investigate the linguistic or textual contexts in which they appear. The idea behind this approach is that the meaning of words can be inferred from their contexts.

An extensive study from our example pool explored the conceptual career of the word "climate" over the period 1800–2010, using two methods to study word contexts (Boyden et al. 2022). The aim of the study was to show that the concept of climate as a global entity – that is, in the meaning the concept has today, as "the climate" – only emerged in the later decades of the twentieth century. In terms of text materials, the study used the Corpus of American Historical English (COHA), a corpus covering roughly 400 millions words from a variety of historical sources: fiction books, magazines, and newspapers.[2]

The first method for studying word contexts concerns collocational analysis from the field of corpus linguistics. *Collocates* are words that have a syntagmatic relationship with the keyword, occuring within a specified window of *n* number of words to the left and right of the keyword. These words co-occur more frequently together with the keyword and can have different part of speech (e.g., noun, verb, adjective) as the keyword (Rapp 2002). In the study, the researchers tracked how collocates of "climate", i.e., words frequently used together with "climate" in a sentence, varied over time, with a particular focus on qualifying adjectives. This showed how in the first period studied (1810-1850) "tropical" and "temperate" were the

2 https://www.english-corpora.org/coha/

most frequent qualifiers, while in the latter half of the twentieth century (1960-2000), "political" and "global" become more prominent.

The second method applied in the study concerns *word embeddings*. Word embeddings capture semantic similarity between words, defined as the extent to which they occur in similar contexts (i.e., paradigmatic similarity). Words with high semantic similarity tend to co-occur with similar words, and generally belong to the same part of speech. They can often be switched out for one another without significantly altering the grammatical structure of a sentence (Rapp 2002). For an introductory non-technical discussion of word embeddings, see McGillivray (2022). In the climate study, word embeddings are used to calculate similarity distances between "climate" and other words. For instance, the study shows how the two words "climate" and "environment" appeared in very different contexts in the beginning of the nineteenth century, but that the similarity of these contexts increased over time, and today the two words are used in very similar contexts.

Word embeddings were also used to study word frequency changes in patient organization periodicals and medical professional journals in the project Acting Out Disease: How Patient Organizations Shaped Medicine (ActDisease).[3] Unlike the climate study, which focused on distances between and clusters surrounding specific words, the ActDisease project uses word embeddings to create multiple *clusters of semantically similar words* and to analyze how the content of these clusters changes over time. For instance, the clusters show how prominent names, medical professions important to the field, or frequently mentioned medical problems varied over time within the corpus. It is also possible to compare different time periods on a semantic level. For example, instead of just observing that the word "driver" was frequently used during a certain period, one can observe that an entire cluster of trafic-related words was prominent during that time (Skeppstedt et al. 2025; Chapter 6 in this handbook).

3.3 *Text categorization*

Text categorization, or classification, can be used to automatically assign texts to organized categories, which is useful when you have acquired a collection of texts that is too large to manually organize into categories. Text classifiers can help to investigate the distribution of categories in a collection of texts or extract texts belonging to specific categories for further (manual or automatic) analysis. Again, two different approaches can be taken for the task of text categorization. If the relevant categories are known beforehand, a supervised approach should be used (Marsland 2009: 7–11). Instead, if

3 https://www.actdisease.org/

the aim of the study is to discover categories based on the content of the texts themselves, an *unsupervised* approach is more appropriate (Marsland 2009: 195–200). For supervised text categorization, the predefined categories may be standard categories, as in the case of sentiment analysis ("negative", "neutral", "positive"), or specific categories tailored to the research question.

A concrete example comes from a pilot project on the commodification of wild berries. In this project, a transformer-based Finnish BERT model was fine-tuned to carry out a binary classification of nineteenth century Finnish newspaper articles about berries and berry picking. The two categories distinguished whether the articles discussed the use of the wild berries as an economic resource or not (Vats 2022, La Mela & Vats 2023). To fine-tune the BERT model, trained on general Finnish language data, and teach it to perform this particular classification task, 415 articles were manually annotated according to these predefined categories.

Another example from the ActDisease project involves classifying texts from patient organization periodicals into different genres, such as academic, news, advertisement, and guidance (Danilova & Söderfeldt 2025). Automatic genre classification makes it possible to perform more focused studies on parts of the text collection, in this case the patient organization periodical. For instance, it could be used to filter out genres not central to the study – such as advertisements – or to extract texts belonging to a specific genre of interest – e.g., the advertisements if these form the focus of the study. In order for the model to learn the features of each of the genres, a set of manually classified examples was required.

In contrast, when relevant categories are not known in advance, unsupervised text categorization offers yet another way to gain insight into large corpora without a predefined approach. A commonly used method for unsupervised text categorization, in particular within humanities research, is topic modeling (Du 2019). Topic modeling is used to discover hidden themes or topics in large corpora (Blei 2012). Automatically extracted topics are typically represented by a list of associated words that form a topic, as well as a list of texts where this topic occurs. There are various topic modeling techniques. Some rely on co-occurrence patterns within the texts (Boyd-Graber et al. 2017), while others, such as BERTopic, use neural network-based representations of text units to generate topics (Grootendorst 2022).

An (unpublished) example of topic modeling concerns a pilot project on survey data related to eHealth. Open-ended survey questions form a prototypical case where unsupervised classification is suitable; the interesting categories are not known beforehand, but it is reasonable to assume that topics can be found in the data, since some respondents will likely have provided similar answers (Baumer et al. 2017). There are also cases in our

example pool where topic modeling has been applied to corpora spanning multiple years or decades. In such cases, it is possible not only to identify potentially relevant text categories in the corpus, but also to create a timeline visualization showing how the occurrence of these topics changes over time (Söderfeldt et al. 2025, Skeppstedt et al. 2024).

3.4 *Extracting entities from text*

The final NLP task discussed here is *named entity recognition and classification* (NER). The aim of NER is to automatically detect spans of text that refer to names or other specific semantic types (Jurafsky & Martin 2008: 759–768). Commonly extracted entity types include personal names, locations, geopolitical entities, dates, times, and organizations. These extracted entities can be used as a basis for network analysis (Tamper et al. 2019), spatial and geographic text analysis (McDonough et al. 2019), or for improving information retrieval in databases (Brandsen et al. 2022).

In addition to the above mentioned standard entity types, study-specific entities are often defined based on the particular research needs. This was the case in a pilot project on Swedish excavation reports. To improve access to information on houses in the reports, five custom entity categories were created related to house type and construction details. A general Swedish BERT language model, pretrained by the National Library of Sweden, was fine-tuned to detect these entities in the reports. Since this project relied on supervised machine learning, a manually annotated dataset was created to use as training data, consisting of 1 000 examples of entities that were to be extracted (Maen 2024).

Another approach to extracting entities was used in a pilot project on digitized eighteenth century handwritten bills of sale. In this case, prompt engineering with the large language model Mixtral was employed to identify and extract entities such as buyer, seller, sum, village, and region (Badri 2024). These entities were then stored as structured data for further analysis. Unlike the project on excavation reports, this approach did not require a manually annotated dataset for training, although annotated examples of bills of sale were included as examples when instructing the model.

4 *NLP resources and tools for the humanities*

Getting into NLP does not mean you need to start from scratch. A wide range of existing resources and tools are available, offering suitable starting points for exploring NLP in the context of humanities research. These resources vary in terms of applicability and technical skills required to use them.

The availability of NLP resources depends on the language in question. Some resources are language-independent, meaning they can be applied to any, or at least many, languages, while others are specifically designed for one or a limited number of languages. For low-resource languages, there may not be any pretrained models, annotated corpora, or larger textual datasets available, which raises the bar for starting to experiment with NLP. Furthermore, tools and models developed for high-resource languages are not always easily adapted to low-resource languages, especially when there are substantial differences in writing system (Tasovac et al. 2023, Dombrowski & Burns 2023).

Tools exist that do not require any programming knowledge. Voyant Tools[4], a web-based application for text analysis, and AntConc[5], a text-analysis toolkit, are examples of programs that allow users to perform computational text analysis through a graphical user interface. While these tools provide suitable starting points for experimenting with NLP and corpus linguistics, they do have certain limitations (Arnold & Tilton 2019). As a user you are limited to the functionalities the developers have decided to include in the tool, which may not be sufficient for answering particular research questions or working with certain types of text materials (Jockers & Underwood 2015). In the case of Voyant Tools, for instance, texts need to be uploaded to a server for analysis, which cannot always be done with sensitive or copyrighted data.[6] Additionally, in terms of transparency, it can be unclear how exactly the texts have been processed behind the interface. Lastly, these types of tools also tend to work better with small or medium-sized corpora rather than large text collections.

Resources are also available for those with knowledge of programming. Python libraries such as the Natural Language Toolkit (NLTK)[7], spaCy[8], and Gensim[9] are commonly used to perform NLP tasks such as those discussed in the previous section. Compared to ready-made tools, these libraries offer a higher degree of flexibility and customization, more transparency and control over the processing steps, and the ability to work with larger text collections. However, they require a higher level of technical expertise, including programming skills and the ability to troubleshoot when packages do not work straightaway. Interactive coding notebooks, such as Jupyter

4 https://voyant-tools.org/
5 https://www.laurenceanthony.net/software/antconc/
6 Voyant Server is a standalone version of Voyant Tools which can be installed and run locally on a computer https://github.com/voyanttools/VoyantServer
7 https://www.nltk.org/
8 https://spacy.io/
9 https://radimrehurek.com/gensim/

Notebook[10], Google Colab[11] and Kaggle[12] take away some of these challenges. In these notebooks, code is presented in executable blocks often alongside explanations of what each block does. Using and modifying existing notebooks reduces the need to code from scratch, but still requires familiarity with programming, to load data correctly, adjust parameters, and install the necessary libraries.

There are also programming resources designed specifically for humanities researchers and their research questions, often in the form of interactive notebooks. The website Programming Historian offers beginner-friendly tutorials for humanists interested in learning about digital tools and techniques.[13] The website includes a variety of tutorials with step-by-step instructions on NLP topics such as word embeddings, topic modeling, and sentiment analysis. Another resource in this category is the course Introduction to Cultural Analytics and Python designed by Melanie Walsh.[14] The course has a broader scope than just NLP. Alongside a chapter on text analysis, it includes introductory chapters on the basics of Python, data collection and analysis, and network analysis.

In the Swedish context, the Huminfra nodes provide access to many useful resources on NLP and other topics relevant to the digital humanities. The National Library of Sweden publishes language models trained on Swedish data that can be used to perform NLP tasks, for example models that can be fine-tuned for entity extraction as exemplified in the pilot study on Swedish excavation reports.[15] Språkbanken offers digital tools for textual analysis and access to Swedish language resources.[16] For example word embedding models trained on Swedish historical newspaper material (Hengchen & Tahmasebi 2021) and a variety of research tools such as Mink, with which texts can be lemmatized and tagged, e.g., with part-of-speech.[17] Another resource is Efselab (Efficient Sequence Labeling, Östling 2018)[18], which we have used in the case studies for lemmatization and part-of-speech tagging.

10 https://jupyter.org/
11 https://colab.google/notebooks/
12 https://www.kaggle.com/
13 https://programminghistorian.org/en/
14 https://melaniewalsh.github.io/Intro-Cultural-Analytics/welcome.html
15 https://huggingface.co/KBLab
16 https://spraakbanken.gu.se/
17 https://spraakbanken.gu.se/mink/
18 https://github.com/robertostling/efselab

5 *Case studies*

In this section, two case studies are discussed to present concrete examples of how methods from NLP and corpus linguistics can be used to answer research questions in the field of history and literature. The case studies illustrate how textual data can be collected, processed, and analyzed in various ways, using both a ready-to-use text analysis tool and programming libraries. Considering the abundance of available resources and tutorials for working with English texts, we will provide examples using Swedish corpora. The basic principles discussed in this section can, however, be applied to other languages of interest. The history case study uses parliamentary motions retrieved from the open data portal of the Swedish parliament (*Riksdagens öppna data*), while the literature case study draws on novels retrieved from The Swedish Literature Bank (*Litteraturbanken*).

5.1 *Case study 1: History*

Parliamentary records are often used by historians interested in applying NLP methods. From a practical perspective, their advantages for computational exploration include accessibility, a relatively uniform structure and format, and coverage across long time periods. From a research standpoint, parliamentary records are an interesting resource to explore as they reflect the political, economic and societal developments of a given era (Ohlsson et al. 2022). This case study explores parliamentary motions. Motions are proposals for decisions submitted by one or more members of parliament. They can be introduced either in response to a specific proposal from the government or during the general private member's motion period when motions can be submitted on virtually any topic.[19]

To explore and analyze the motions, we use the text analysis toolkit AntConc (Anthony 2024). AntConc offers various tools for text analysis and corpus exploration. The program is user-friendly and requires no programming experience, although preprocessing the texts in advance can help get more meaningful results out of some of the tools. Laurence Anthony, the developer, has created a series of video tutorials on YouTube that show how to get started with AntConc and use the tools included in the program.[20] For the purpose of this case study, we will look more closely into the following tools: the Word list, Key Word in Context (KWIC), File, Collocate, and Keyword list tool.

19 https://www.riksdagen.se/en/news/the-general-private-members-motions-period/

20 Search for 'AntConc (Version 4.2) Tutorials' https://www.youtube.com/playlist?list=PLiRIDpYmiC0SjJeT2FuysOkLa45HG_tIu

For historians, it is useful to be able to zoom in and out of a corpus, looking at one text at a time as well as getting a broader overview of the corpus content, moving between so-called 'distant reading' and 'close reading'. AntConc supports this approach. By dividing the text collection into subcorpora it is also possible to study temporal change. For example, Jarlbrink & Norén (2023) use AntConc to study the evolving concept of propaganda in parliamentary debates from 1871 to 2019, combining close reading with the analysis of n-grams, co-occurrences, and keywords. Moving between distant and close reading helps overcome some of the limitations of large-scale text analysis. Calculated word frequencies and word co-occurrences do not automatically return significant or relevant terms that indicate change. Domain knowledge of the materials and their historical context, as well as returning to the original texts, is essential for interpreting the results and drawing meaningful conclusions (Guldi 2023).

The texts used for the AntConc experiments were retrieved from the open data portal of the Swedish parliament (Riksdagens öppna data).[21] Motions put forward by members of parliament between 1971 and 1979 were downloaded as plain text files along with a metadata file containing the date of the motion, its title, and initiator(s). Party affiliation is not included in the metadata and can therefore not easily be utilized for analysis. In total, the corpus consists of 20 972 motions.

Besides the data portal of Swedish parliament, motions can also be accessed and explored in a web interface developed by Gothenburg Research Infrastructure in Digital Humanities (GRIDH) in collaboration with Chalmers University of Technology.[22] Additionally, the project Swedish Riksdag 1867–2022: An Ecosystem of Linked Open Data (SWERIK) is in the process of making parliamentary records, including motions, more accessible for quantitative and qualitative research.[23] Proceedings from parliamentary debates curated by SWERIK can be explored through an interface that supports keyword searches, filtering by year and party affiliation, and the study of collocates and n-grams.[24] The curated corpus of motions are at the time of writing this chapter only available on their GitHub repository as XML files.[25] Users with programming experience may prefer to retrieve motions

21 In the data portal, a wide range of parliamentary documents are available, including proceedings from debates, motions, bills, and public investigations (SOU). https://www.riksdagen.se/sv/dokument-och-lagar/riksdagens-oppna-data/

22 https://riksdagsmotioner.dh.gu.se/

23 https://swerik-project.github.io/

24 https://riksdagsdebatter.se/

25 https://github.com/swerik-project/riksdagen-motions

from this repository, as the SWERIK files have improved OCR and party affiliation included in the metadata.

This case study demonstrate the use of AntConc for historical research in two scenarios: one where we already know what we want to explore in the corpus and one where we adopt an exploratory approach and aim to discover patterns within the corpus itself. For the first scenario, we are interested in examining motions discussing the topic of nuclear power. Nuclear power as an alternative source of energy became an item on the Swedish political agenda during the 1970s as a result of the global oil crisis.

Loading files into AntConc can be done in the corpus manager (navigate to "File" and click on "Open Corpus Manager"). To load the motions, choose "Raw File(s)" as corpus source, give the corpus an appropriate name, and add the files or directory containing the motions that you would like to include in the corpus. For this case study, we will use the standard settings. Press create. Once the process is finished, a database file (.db) is created, which can be reopened at any time from the corpus manager. AntConc supports common file formats, including text (.txt), word (.doc and .docx), and pdf files (.pdf). In the corpus manager, you can see how many files the corpus contains, the number of tokens (in this case the number of words), and the number of token types. In our corpus, there are 20 972 motions, 15 433 308 tokens, and 299 311 token types.

With the corpus loaded into AntConc, we can start exploring its content. The **Word list tool** lets you count the frequencies of words in the corpus. This can be done in two ways. By pressing start without entering any words in the search bar, you get an overview of the frequency of all the words in the corpus. It is also possible to count frequencies of a specific term or multiple terms. The asterisk symbol (*) before or after the word can be used to find variations of that word. Two vertical bars (||) between the words function as OR, allowing you to search for two or more words. Other available wildcard symbols include the question mark (?) for any single character or the plus symbol (+) for one or more characters. These wildcards can also be used in some of the other tools described below.

In addition to frequency, the range of the words in the corpus is also indicated in the Word list tool, that is the number of documents in which the word appears. Range helps determine whether a word is used occasionally across many texts or frequently in only a few. This is important when making arguments about the content of a corpus: if a word occurs a hundred times in one but very rarely in other texts, it says a lot about the content of that particular text but it may not be representative of the corpus as a whole.

To get an overview of the frequency of the word nuclear power and its variations, we can search for "kärnkraft*". In this case, the definite form of

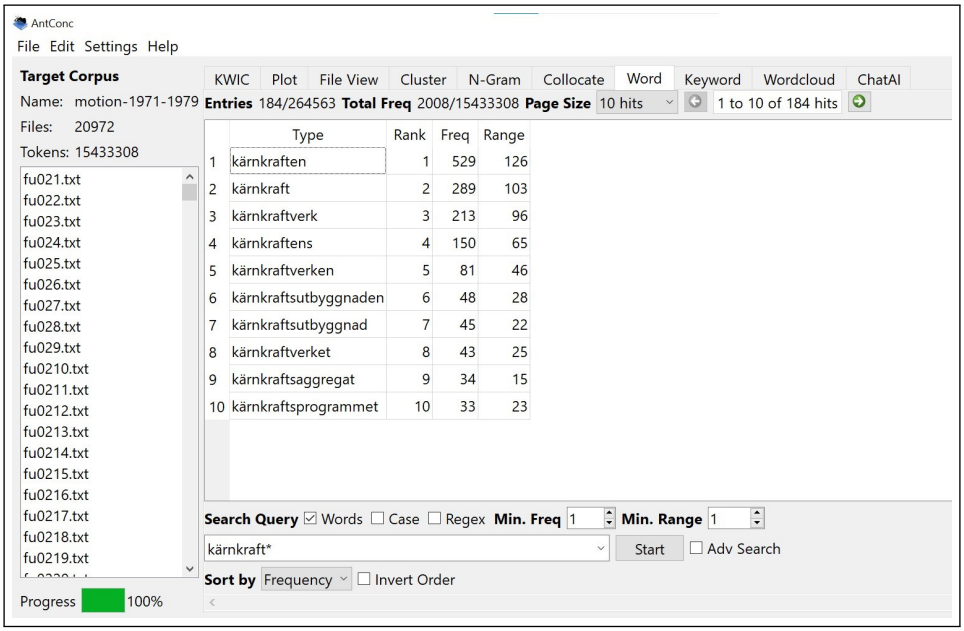


Figure 1: Ten most frequent words of nuclear power and its variations.

nuclear power (kärnkraften) is the most common in the corpus, appearing 529 times (Freq) in 126 motions (Range), which corresponds to 0.6% of the total number of motions in the corpus (Figure 1).

If you find an interesting word that you would like to inspect more closely, you can click on it, and the program redirects you to the **KWIC tool**, also known as the concordance tool. Here, all occurences of that word are listed with a context window of *n* number of tokens on each side of the word. The hits can be displayed in different orderings (see Sort Options), for example, based on the context to the right or the left of the word. This way patterns in the language use can be identified. It is also possible to export the results as a tab separated text file (File > Save Current Tab Results) or csv file (File > Save Current Tab Database Tables) for further analysis in another program.

If we want to learn more about how the motions discuss the construction or expansion of nuclear power, we can, for example, click on "kärnkraftutbyggnaden". Now, the 48 hits for this word are displayed in the KWIC tool, sorted by the right-hand context of the word (Figure 2).

Again, if you find an interesting instance of a word in the KWIC tool that you would like to examine more closely, it is possible to navigate to the content of a single file. Clicking on a hit will redirect you to the **File tool**. Here, the full-text of the motion will be displayed with the hit(s) highlighted.

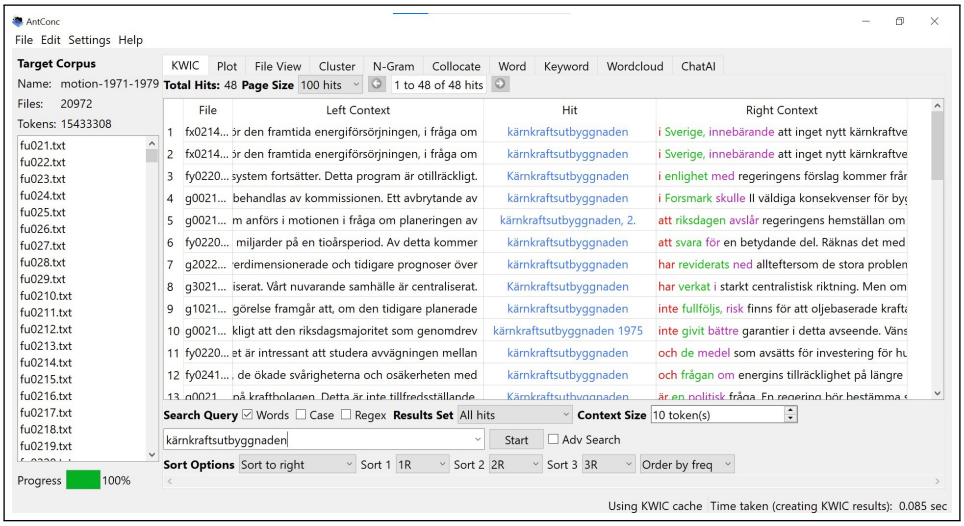


Figure 2: Hits for kärnkraftutbygganden in KWIC tool.

It is also possible to access the content of any of the motions directly by clicking on the name on the left side of the interface. If we click on the first hit for "kärnkraftuppbyggnaden" we can see the entire text of the motion, with each hit highlighted in blue (Figure 3).

So far we have used the texts directly as they were retrieved from the data portal. However, it is also an option to preprocess the text to increase the usefulness of some of the tools in AntConc. For example, when counting word frequencies, it can be helpful if the definite and indefinite forms of a word or various conjugations of a verb are counted as the same. Preprocessing steps are done to make the texts more understandable for machines, keep words that carry the most information, and in some cases also to reduce the size of the dataset. It often takes multiple iterations to get the preprocessing 'right' and how it exactly should be done is determined by the research questions that are to be explored. During the preprocessing a lot of decisions are made about what words to keep and what words to consider as stop words or noise. The choices made during this stage can greatly impact the final output, making preprocessing an integral part of the research process (Schofield 2022). Tahmasebi & Hengchen (2019) provide a detailed overview of the various steps a preprocessing pipeline may contain for this kind of study.

The corpus of motions was preprocessed in the following way. After collecting the motions as text files, the first preprocessing step was to clean the texts, in this case removing metadata text chunks, not part of the actual motions. We identified six such repetitive text chunks, which contained

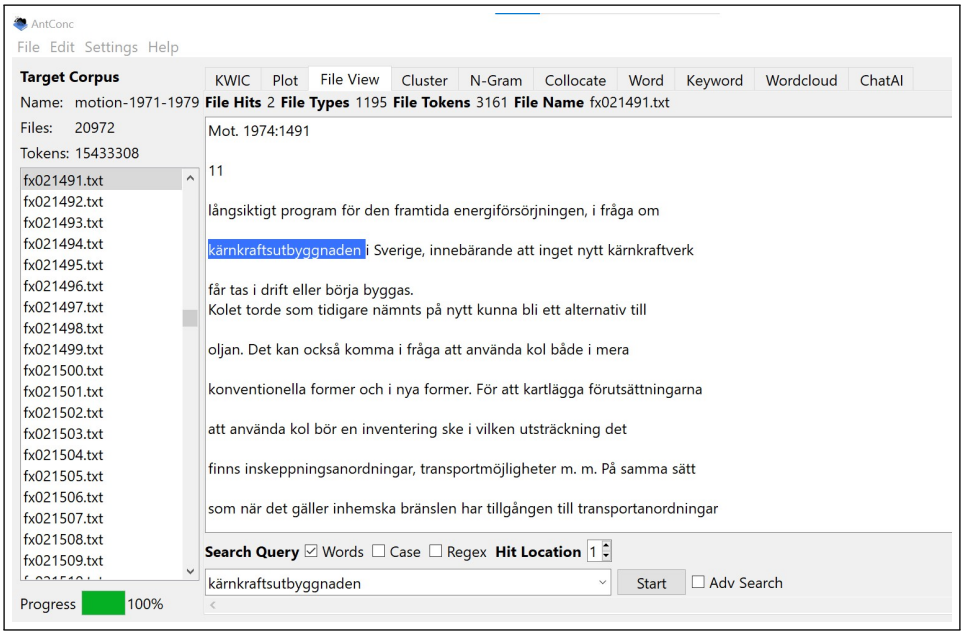


Figure 3: First hit for kärnkraftutbygganden in File tool.

information regarding the quality of the OCR as well as the different printing houses that had been responsible for printing the motions. The second step consisted of tokenizing and lemmatizing the texts. Tokenization refers to the process of splitting up the texts into tokens, smaller linguistic units, while lemmatization is the process of transforming each word into its base or dictionary form (Tahmasebi & Hengchen 2019). For this, we used the previously mentioned Swedish tokenizer/lemmatizer Efselab (Östling 2018). As third and final preprocessing step, stopwords were removed from the texts. We used the Swedish stopword list included in NLTK, a Python programming library for NLP, and expanded it with a list of manually compiled stop words specific to the corpus in question. Whether to remove stop words or not depends on the research question. When focusing on content, words such as "the", "a", "we", "your", are often regarded as uninteresting, but for other explorations there can be reasons to keep them. The code used to preprocess the corpus is available in the GitHub repository.[26]

For newcomers to NLP and programming, using libraries such as Efselab or spaCy to preprocess the texts might not be a possibility in the beginning. There are tools available that can perform lemmatization without the need

26 https://github.com/CDHUppsala/Applied-NLP-for-humanities-research

for programming, for example TagAnt[27] and the previously mentioned Mink. These tools have, however, not been explored and evaluated for this case study. AntConc also includes a functionality to filter the words displayed in some of its tools. This can be done under Settings, Global Settings, Tool Filters. The tool filters can be used in two ways. First, it can be used as a curated keyword list to only display words that you are interested in.[28] Secondly, it is possible to use the tool filter functionality as a stop word list. Instead of removing the words from the actual motions, the words are filtered out from the results displayed in the tools. The list can be iteratively extended.

For this case study we have worked with two corpora. One corpus consists of the 'raw' text files without any preprocessing, adding the files as they were retrieved from the open data portal. This corpus is useful for close reading the texts in the KWIC and file tool. The second corpus contained all words in lemmatized form and had stop words removed, which can be used for looking for co-occurrences and finding keywords that characterize a subsection of the corpus as will be shown below.

The **Collocate tool** gives additional options to study a keyword and its frequently surrounding words. The tool shows how a keyword co-occurs with other words in the corpus within a certain distance. The window span can be adjusted on either side of the word. Random distribution is used as a base line to see if words appear more often or more seldom together. This method is sensitive to statistical outliers. It may therefore be a good idea to set a minimum threshold for terms to be included in the calculations (freq) and in how many of the documents they should appear (range). From the collocate tool, it is possible to navigate to the KWIC tool by clicking on one co-occurring word. Depending on the size of the corpus, it might take a while before the results are generated.

In Figure 4 we can see what words tend to co-occur with the lemma nuclear power ("kärnkraft") within a window span of five tokens to the left and the right side of the search term. From the results, we get a sense of the different contexts in wich nuclear power is discussed, from investing (uppbyggnad, satsning) to decommissioning (avveckla) nuclear power, and the possibility of having a referendum on the topic (folkomröstning). Dividing the corpus into multiple time periods and rerunning the collocate analysis might reveal that in the beginning of the period more motions dicussed the building up of nuclear power while at the end of the period the decommissioning of nuclear power was discussed more. From the collocates

27 https://laurenceanthony.net/software/tagant/

28 The rank of the keywords does not change so it might be necessary to adjust the page size or use the arrows to navigate to the next results to see all the words included in the list.

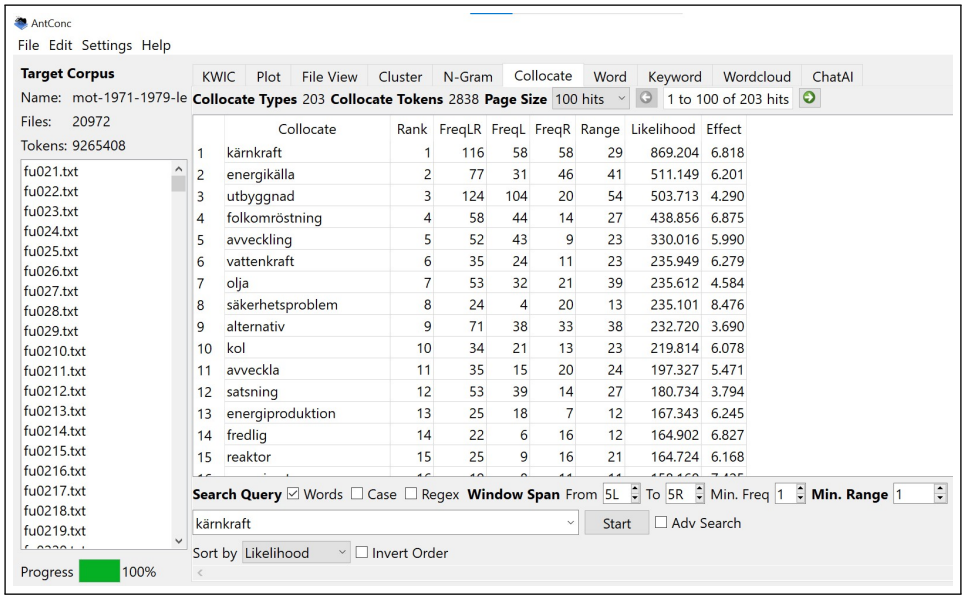
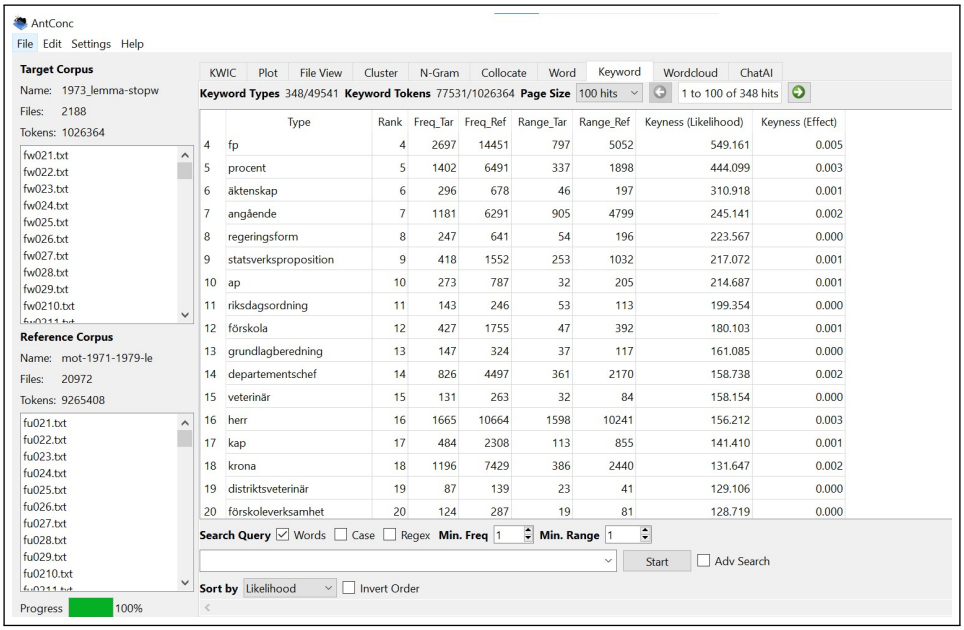


Figure 4: Collocate hits for kärnkraft.

alone it is not possible to deduce the exact context in which the words were used. Therefore, it is important to inspect the actual sentences and texts in which they appear, which we can reach through the KWIC and File tool, which have been introduced above.

Finally, studying the content of the corpus in a more exploratory fashion is also possible with AntConc. As previously mentioned, the Word list tool can be used to get an overview of the most frequent words in the corpus. This can be done by pressing start without entering any search term. Additionally, the **Keyword list tool** can be used to compare a subsection of the corpus (target corpus) against the entire corpus (reference corpus), as was done in the study on the patient organization periodical Allergia discussed previously (Söderfeldt et al. 2019). To use this tool, two corpora need to be loaded into AntConc: the subset of the corpus you would like to analyze and the entire original corpus. Once both corpora are loaded, open the subset as target corpus and the entire original corpus as reference corpus in the corpus manager. Then navigate to the Keyword tool and press Start. Based on a likelihood measure, this tool can show which words appear unusually frequently or infrequently in the subset in comparison to the entire corpus. It shows what words are characteristic for the subset. To inspect the negative keywords, words that appear unusually infrequently, navigate to Settings, Tool Settings, Keyword and select Show negative keywords.



	Type	Rank	Freq_Tar	Freq_Ref	Range_Tar	Range_Ref	Keyness (Likelihood)	Keyness (Effect)
4	fp	4	2697	14451	797	5052	549.161	0.005
5	procent	5	1402	6491	337	1898	444.099	0.003
6	äktenskap	6	296	678	46	197	310.918	0.001
7	angående	7	1181	6291	905	4799	245.141	0.002
8	regeringsform	8	247	641	54	196	223.567	0.000
9	statsverksproposition	9	418	1552	253	1032	217.072	0.001
10	ap	10	273	787	32	205	214.687	0.001
11	riksdagsordning	11	143	246	53	113	199.354	0.000
12	förskola	12	427	1755	47	392	180.103	0.001
13	grundlagberedning	13	147	324	37	117	161.085	0.000
14	departementschef	14	826	4497	361	2170	158.738	0.002
15	veterinär	15	131	263	32	84	158.154	0.000
16	herr	16	1665	10664	1598	10241	156.212	0.003
17	kap	17	484	2308	113	855	141.410	0.001
18	krona	18	1196	7429	386	2440	131.647	0.002
19	distriktsveterinär	19	87	139	23	41	129.106	0.000
20	förskoleverksamhet	20	124	287	19	81	128.719	0.000

Figure 5: Keywords for 1973 compared to the entire corpus.

As an example, we can find the over-represented words for the year 1973 compared to the entire corpus 1971-1979. The output in Figure 5 displays the frequency and range of each word in the target corpus and the reference corpus. The output suggests that an additional preprocessing of the texts may be necessary, for example party abbreviations such as "fp" for Folkpartiet (the predecessor of the Liberal party) or honorifics (herr) could be included in the list of stop words if these are not of interest for the research question. Other words may reveal characteristics of motions submitted during the year 1973, such as words about preschools (förskola, förskoleverksamhet) or the veterinarian (veterinär) that could warrant further exploration.

Ready-to-use tools like AntConc are valuable resources for historians to start experimenting with NLP and corpus linguistics methods without the need to know or learn programming. As shown in this case study, AntConc supports the study of interesting words that are known beforehand but also exploratory approaches for finding interesting words based on the content of the corpus. Such tools do have its limitations compared to programming libraries, as the user is limited to the included functionality and, in the case of AntConc, needs to manually load different versions of the corpus separately in order to analyze them.

5.2 *Case study 2: Literary style*

Our second case study concerns literature, and more specifically literary style in Swedish 20th century prose fiction. "Literary style" is a complex notion within literary studies that can refer to a blend of things on different levels, including syntactic, grammatical, and thematic patterns of various kinds. Yet, certain more formalized aspects of literary style are well-suited to trace by computational means. In fact, the possibilities within computational literary studies has led to a new-found scholarly interest in literary style more broadly (Allison et al. 2013, Herrmann et al. 2015). It is this vein of "instrumentalized" or "operationalized" literary stylistics that the following case study belongs to.

Basically, we are interested in research questions concerning the relationship between literary style – broadly understood as described above – and author position in the literary field. A foregrounding hypothesis is that style relates to levels of public impact and literary prestige. To put it in a very straightforward way: the bigger the audience, the less experimental the writing style. But we also have a second and somewhat contradictory hypothesis that although this might be true in most cases, there are likely both exceptions and other parts of what we here label as style – how literature is written rather than its themes or narrative structures – at work.

To investigate this, we constructed a corpus of 50 novels written by 21 Swedish authors and published between 1910 and 1967. This selection emanates from a much larger study (Berglund & Svedjedal 2025), where principles for corpus selection are discussed thoroughly. For this toy example case study, it is sufficient to say that it includes a broad range of prose fiction that spans from niche and experimental modernist writing with high prestige in the literary field to popular genre fiction with little prestige. Also of importance is that all works in the corpus are written with modern Swedish spelling.

Generally, however, the composition of the corpus is key, and forms the basis for which research questions can be asked. How representative is a certain material in relation to an era or a genre? Are you carrying out an analysis that covers everything you want to investigate (for instance: all novels by Selma Lagerlöf), or have you made a selection? How reliable is your material? For a computational literary analysis to be credible, it is important to be able to answer such questions. The goal is not to create the perfect corpus but to explain and describe each corpus along with its strengths and limitations.

As of materials, we depart from freely available digitized Swedish fiction in Litteraturbanken.[29] From the site we downloaded novels as plain text

files (.txt), and then manually deleted everything apart from the literary text (that is, colophon information about the book such as title, author, and publishing house, prefaces and afterwords that do not belong to the novel text, footnotes and other kinds of editorial comments, etc.). Since the text quality in Litteraturbanken is controlled and excellent, no further data curation of the text material was needed in this case. In total, the curated corpus contains 3.97 million tokens.

In parallel, we made a metadata file with basic information about each of the 50 novels. This includes author name, title of novel, and year of publication. To be able to track how literary prestige relates to style, we also made a literary sociological distinction between three groups of authors: 1) canonized authors with high literary prestige but small readerships in their time (a total of 20 novels in the corpus); 2) canonized authors with high literary prestige *and* large readerships in their time (26 novels); and 3) popular genre fiction authors with low literary prestige but large readerships in their time (4 novels). Literary prestige and levels of canonization were based mainly on major literary prizes and descriptions in literary historiography. Sizes of readerships were based mainly on the number of editions and copies sold (for details, see Berglund & Svedjedal 2025). The different sizes of the three groups are an effect of the focus of Litteraturbanken on canonical Swedish literature. To be able to track this more seriously, other sources would be needed as a complement.

These 50 novels can be imported into AntConc and analyzed in various ways. However, the kinds of stylistic features that we are interested in here are easier to detect by means of programming. This holds true, not least when it comes to visualizations of findings, where AntConc falls short. For all visualizations below, we use standard python libraries *seaborn* and *matplotlib*. You need to learn their specific syntax to be able to plot pretty graphs, but when you do, the possibilities for variation are endless. The Notebook scripts that we have used in this example are available in the GitHub repository.[30]

As a starting point, we ran the corpus through Efselab/Swepipe (Östling 2018). Included in this pipeline and of interest for our analyses are word segmentation, paragraph and sentence segmentation, part-of-speech tagging, and lemmatization. As an output you get a tab-separated conll-file, which includes all of the information above. For example, the first sentence in Selma Lagerlöf's *Körkarlen* (1912) reads: "Det var en stackars liten slumsyster, som höll på att dö." In the output from Efselab/Swepipe, this is transformed as shown in Table 1 (structured as a table for readability).

29 https://litteraturbanken.se/

30 https://github.com/CDHUppsala/Applied-NLP-for-humanities-research

Table 1: Example output from Efselab. (U-POS is universal part of speech.)

Token	Part of speech	U-POS	Lemma
Det	PN|NEU|SIN|DEF|SUB/OBJ	PRON	det
var	VB|PRT|AKT	AUX	vara
en	DT|UTR|SIN|IND	DET	en
stackars	JJ|POS|UTR/NEU|SIN/ PLU|IND/DEF|NOM	ADJ	stackars
liten	JJ|POS|UTR|SIN|IND|NOM	ADJ	liten
slumsyster	NN|UTR|SIN|IND|NOM	NOUN	slumsyster
,	MID	PUNCT	,
som	HP| – | – | –	PRON	som
höll	VB|PRT|AKT	VERB	hålla
på	PL	ADP	på
att	IE	PART	att
dö	VB|INF|AKT	VERB	dö
.	MAD	PUNCT	.
PARAGRAPH _BREAK			PARAGRAPH _BREAK

Although probably confusing at first sight, at least for a humanities scholar, everything that we need for further processing and analysis can be found here. The original text is in the first column, two levels of part-of-speech tagging in the two following columns, and the lemmatized token in the fourth column. End of sentence is indicated by a major delimiter ("MAD" in the second column) followed by a blank row.

So, with all foundational NLP markup at hand: what to search for in terms of style? Let's start with the most low-hanging fruit: punctuation. How authors make use of things like commas, semicolons, and question marks can actually tell us quite a bit about the kind of prose at hand, not least if authors are compared to each other. Since the length of novels vary a lot, absolute numbers are not interesting here. What we need are relative frequencies. To get this is fairly straightforward: we run a script that counts all tokens of a certain kind of punctuation, say semicolon (";"), and then we divide the number of hits with the total number of tokens for the novel in question. If we make a bar plot out of these relative frequencies for semicolon, and highlight the three sociological categories of author through a legend, we get the following graph (Figure 6).

What Figure 6 shows is that semicolons are very unevenly distributed in the corpus. While some authors use it recurrently, around half of the novels

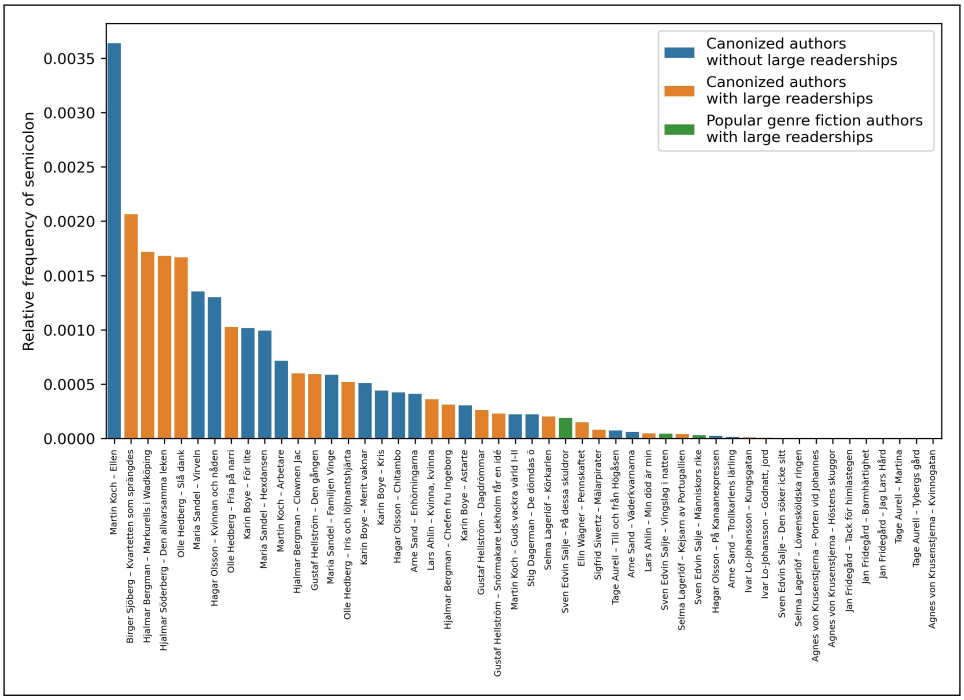


Figure 6: Bar plot of semicolon use.

in the corpus never or very rarely contain any semicolons at all. This typical stylistic feature, then, seems to be tied to novels by certain authors: Martin Kock, Birger Sjöberg, Hjalmar Bergman. Although the popular genre fiction novels in the corpus all have low levels of semicolon, it is not possible to make a link between literary prestige and semicolon use. For instance, Nobel prize laureate Selma Lagerlöf seldom uses semicolons. The same goes for the 1940s modernist Tage Aurell.

Similar calculations can be done with all kinds of punctuation, and they will indicate different aspects of style: high levels of commas indicate lots of subordinate clauses and enumerations, frequent question marks indicate dialogue, lots of exclamation marks dialogue and/or an expressive tone, etc.

But let us move on to two other metrics that are indicative of literary style: length of sentences and length of paragraphs. We use the occurrence of one or several blank lines as an indication of a paragraph break. For sentence lengths, we make use of the major delimiter markup in the Efselab/Swepipe output. For sentences, the most straightforward way is to write a script that divides the number of tokens in each novel with its number of tokens tagged as major delimiters ("MAD") in the second column. For paragraph

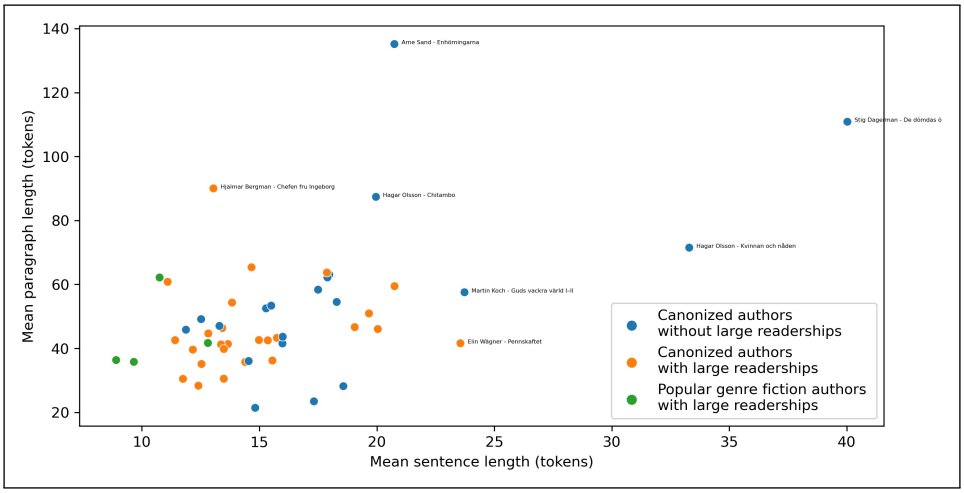


Figure 7: Scatter plot of mean sentence and paragraph length.

breaks, we have added the tag "PARAGRAPH_BREAK" to the output from Efselab/Swepipe to indicate a blank row in the original text. We can, thereby, similarly divide the number of tokens in each novel with its number of paragraph breaks.[31] If we make a scatter plot of how the 50 novels in our corpus are distributed according to these two metrics, we get the result displayed in Figure 7.

What immediately stands out in Figure 7 is that a majority of the novels in the corpus cluster in the lower left corner. Most novels investigated have a mean sentence length of between 10 and 20 tokens, and a mean paragraph length between 30 and 60 tokens. But what we also see in the graph is a smaller group of outlier novels, where both sentences and paragraphs are much longer. These outliers are experimenting with language in different ways: the modernist prose of Stig Dagerman and Hagar Olsson stands out in particular with their long and complex sentences, while the long paragraphs in Arne Sand's *Enhörningarna* (1965) can be explained by the novel's self-reflexive and diary-like internal monologue and its more or less total lack of dialogue.

The most extreme outliers are all written by the group of canonized authors without large readerships, which is at least an indication that these kinds of stylistic experiments is not something that the common reader tends to appreciate. Too bold claims should not be made out of this small selection, but this tendency shows in Figure 7. Novels by popular genre fiction authors

31 You could also use spaCy for this – its sentence tokenizer is even more easy to use than Efselab/Swepipe, see Berglund & Svedjedal (2025: 180–189).

Table 2: Mean sentence and paragraph length per author category

Author group	Novels in corpus	Sentence length (tokens)	Paragraph length (tokens)
Canonized authors without large readerships	20	18.8	57.1
Canonized authors with large readerships	26	15.0	46.1
Popular genre fiction authors with large readerships	4	10.5	44.0

with large readerships (the green dots) are clustered towards the lower left corner. Novels written by canonized authors with large readerships (the orange dots) do so as well, but slightly more in the upper right direction. Novels by canonized authors without large readership (blue dots) are the most scattered, and none of them have very short sentences on average. If calculated on a group level – by taking mean scores of the mean values for all novels respectively – the patterns stand clear, as shown in Table 2.

Also part of speech distribution has something to say about literary style. High amounts of verbs indicate plot-driven narratives with plenty of dialogue, whereas high amounts of adjectives indicate narratives where depictions and settings are important. In general, popular genre fiction is to a high extent plot-driven where canonical fiction are more diverse (Berglund et al. 2019, Berglund & Dahllöf 2021).

To get figures on parts of speech, we depart from the third column of the Efselab/Swepipe output. This information builds on the part of speech tags in Universal Dependencies (U-POS), with 17 classes in total.[32] For our analysis, we are interested in adjectives (tagged as "ADJ") and verbs ("VERB"). To calculate the share of these parts of speeches, we proceed in a similar vein as previously: we write a script that checks every token in every novel, and count all tokens that are tagged as adjectives and verbs respectively. We then divide the number of adjectives/verbs with the total number of tokens for every novel. From this, we get two new metrics: verb share and adjective share.

32 See https://universaldependencies.org/u/pos/ Efselab/Swepipe also provides a more thorough part of speech markup in the second column of its output.

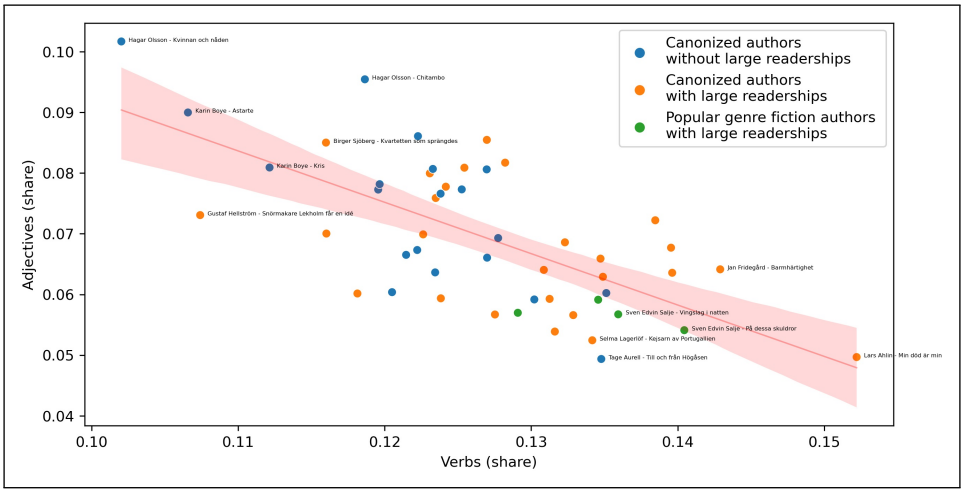


Figure 8: Scatter plot of verb and adjective share.

If we make a scatter plot of verb and adjective share for the 50 novels in our corpus (Figure 8) we see that there is a negative correlation between these two metrics: in general, the more adjectives in a novel, the fewer verbs, and vice versa. The strength of this correlation can be calculated by means of a linear regression analysis – this is what the red regression line in Figure 8 indicates. If we use *lingress* from the standard python library *scipy.stats*, we get that the correlation between share of verbs and share of adjectives for the novels in our corpus is $R = -0.668$, which in statistical terms suggest a fairly strong negative correlation.

This outcome makes sense and is no surprise. What is more interesting, though, is to take a closer look at the outliers. In the upper left corner (lots of adjectives, few verbs) of Figure 8, we find modernist writer Hagar Olsson and Karin Boye's more experimental novels. In the lower right corner (lots of verbs, few adjectives), we find the novels of the only popular genre fiction author in the corpus – Sven Edvin Salje – but also novels by canonized authors such as Selma Lagerlöf and Lars Ahlin. If we calculate the mean shares on a group level, there are apparent differences regarding part of speech use: canonized authors without large readerships are the most adjective heavy, while the popular genre fiction authors use the most verbs (see Table 3). What is important to also keep in mind is that canonized authors without large readerships are the most diverse in their literary style.

We could go on and calculate a number of other metrics concerning literary style (see further Berglund & Svedjedal (2025)), but for this toy example case study we hope to have proven our point: that basic NLP markup

Table 3: Verb and adjective share per author category

Author group	Novels in corpus	Verbs (mean perc.)	Adjectives (mean perc.)
Canonized authors without large readerships	20	12.2	7.4
Canonized authors with large readerships	26	12.9	6.8
Popular genre fiction authors with large readerships	4	13.5	5.7

(in this case derived from Efselab/Swepipe) along with easily graspable visualizations (here from matplotlib/seaborn) can be used for all kinds of insights into literary history and literary style.

6 *Conclusion*

This chapter has aimed to give humanities scholars who are curious about the possibilities of NLP a first introduction to the field. By discussing some of the challenges involved in getting started with NLP, as well as presenting practical examples of how NLP methods can be utilized within the humanities, we hope to have offered the reader an opportunity to begin experimenting with NLP in their own work.

The field of NLP offers a wide variety of computational methods for large-scale text analysis. The two case studies presented in this chapter demonstrate that these methods can be integrated in humanities research by any scholar regardless of their technical background, either through ready-to-use text analysis tools or programming libraries. More advanced NLP applications based on machine learning may fall outside the scope of many humanities scholars and require interdisciplinary collaboration with NLP experts and data scientists. However, having an understanding of the key NLP concepts, available methods, and how they differ from traditional humanities approaches – which this chapter has aimed to provide – can help facilitate such collaborations in future projects.

Notebooks

Code and notebooks connected to the case studies can be found on: https://github.com/CDHUppsala/Applied-NLP-for-humanities-research.

Acknowledgments

Part of the work here was conducted within the ActDisease project. ActDisease is funded by the European Union (ERC ActDisease, ERC-2021-STG 101040999). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

References

Allison, Sarah, Marissa Gemma, Ryan Heuser, Franco Moretti, Amir Tevel & Irena Yamboliev. 2013. Style at the scale of the sentence. *Pamphlets of the Stanford Literary Lab* 5. https://litlab.stanford.edu/LiteraryLabPamphlet5.pdf.

Anthony, Laurence. 2024. *AntConc* (*Version 4.3.1*) [*Computer Software*]. Tokyo, Japan: Waseda University. https://laurenceanthony.net/software/antconc.

Arnold, Taylor & Lauren Tilton. 2019. New data? The role of statistics in DH. In Matthew Gold & Lauren Klein (eds.), *Debates in the digital humanities 2019*. Minneapolis: University of Minnesota Press. DOI: 10.5749/9781452963785.

Artstein, Ron & Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics* 34(4). 555–596. DOI: 10.1162/coli.07-034-R2.

Badri, Sushruth. 2024. *Digitising deeds*. https://github.com/CDHUppsala/digitising-deeds.

Bamman, David. 2017. Natural language processing for the long tail. In *Digital Humanities 2017 conference abstracts*, 382–384. Montreal, Canada. https://dh2017.adho.org/abstracts/408/408.pdf.

Baumer, Eric P. S., David Mimno, Shion Guha, Emily Quan & Geri K. Gay. 2017. Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? *Journal of the Association for Information Science and Technology* 68(6). 1397–1410. DOI: 10.1002/asi.23786.

Berglund, Karl & Mats Dahllöf. 2021. Audiobook stylistics: Comparing print and audio in the bestselling segment. *Journal of Cultural Analytics* 6(3). 1–30. DOI: 10.22148/001c.29802.

Berglund, Karl, Mats Dahllöf & Jerry Määttä. 2019. Apples and oranges? Large-scale thematic comparisons of contemporary Swedish popular and literary fiction. *Samlaren: Tidskrift för forskning om svensk och annan nordisk litteratur* 140. 228–260.

Berglund, Karl & Johan Svedjedal. 2025. *Det stora lästa. Fjärrläsningar av klassiska svenska romaners stilvärldar, 1910–1999*. Uppsala: Uppsala University. https://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-557694.

Blei, David M. 2012. Topic modeling and digital humanities. *Journal of Digital Humanities* 2(1). 8–11.

Bode, Katherine. 2023. What's the matter with computational literary studies? *Critical Inquiry* 49(4). 507–529. DOI: 10.1086/724943.

Boyd-Graber, Jordan, Yuening Hu & David Mimno. 2017. Applications of topic models. *Foundations and Trends® in Information Retrieval* 11(2-3). 143–296. DOI: 10.1561/1500000030.

Boyden, Michael, Ali Basirat & Karl Berglund. 2022. Digital conceptual history and the emergence of a globalized climate imaginary. *Contributions to the History of Concepts* 17(2). 95–122. DOI: 10.3167/choc.2022.170205.

Brandsen, Alex, Suzan Verberne, Karsten Lambers & Milco Wansleeben. 2022. Can BERT dig it? Named entity recognition for information retrieval in the archaeology domain. *Journal on Computing and Cultural Heritage (JOCCH)* 15(3). 1–18. DOI: 10.1145/3497842.

Danilova, Vera & Ylva Söderfeldt. 2025. Classifying textual genre in historical magazines (1875-1990). In Anna Kazantseva, Stan Szpakowicz, Stefania Degaetano-Ortlieb, Yuri Bizzoni & Janis Pagel (eds.), *Proceedings of the 9th joint SIGHUM workshop on computational linguistics for cultural heritage, social sciences, humanities and literature (latech-clfl 2025)*, 160–171. Albuquerque, New Mexico. DOI: 10.18653/v1/2025.latechclfl-1.15.

de Bolla, Peter, Ewan Jones, Paul Nulty, Gabriel Recchia & John Regan. 2019. Distributional concept analysis: A computational model for history of concepts. *Contributions to the History of Concepts* 14(1). 66–92. DOI: 10.3167/choc.2019.140104.

Dombrowski, Quinn & Patrick J. Burns. 2023. Language is not a default setting: Countering DH's English problem. In Matthew K. Gold & Lauren F. Klein (eds.), *Debates in the digital humanities 2023*. Minneapolis: University of Minnesota Press. DOI: 10.5749/9781452969565.

Drucker, Johanna. 2017. Why distant reading isn't. *PMLA/Publications of the Modern Language Association of America* 132(3). 628–635. DOI: 10.1632/pmla.2017.132.3.628.

Du, Keli. 2019. A survey on LDA topic modeling in digital humanities. In *Book of abstracts DH2019*. DOI: 10.34894/H9UYPI.

Dubremetz, Marie. 2023. *Philosthetic project*. https://gitlab.com/mardub/philosthetic_pilot.

Ekeman, Karl. 2023. *In want of a sovereign: Metapolitics and the populist formation of the alt-right*. Uppsala University. (Doctoral dissertation).

Fridlund, Mats, Mila Oiva & Petri Paju. 2020. *Digital histories: Emergent approaches within the new digital history*. Helsinki: Helsinki University Press. DOI: 10.2307/j.ctv1c9hpt8.

Grootendorst, Maarten. 2022. *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. DOI: 10.48550/arXiv.2203.05794.

Guldi, Jo. 2023. *The dangerous art of text mining: A methodology for digital history*. Cambridge: Cambridge University Press. DOI: 10.1017/9781009263016.

Hengchen, Simon & Nina Tahmasebi. 2021. A collection of Swedish diachronic word embedding models trained on historical newspaper data. *Journal of Open Humanities Data* 7(2). DOI: 10.5334/johd.22.

Herrmann, J. Berenike, Karina van Dalen-Oskam & Christof Schöch. 2015. Revisiting style, a key concept in literary studies. *Journal of Literary Theory* 9(1). 25–52. DOI: 10.1515/jlt-2015-0003.

Hu, Yan, Qingyu Chen, Jingcheng Du, Xueqing Peng, Vipina Kuttichi Keloth, Xu Zuo, Yujia Zhou, Zehan Li, Xiaoqian Jiang, Zhiyong Lu, Kirk Roberts & Hua Xu. 2024. Improving large language models for clinical named entity recognition via prompt engineering. eng. *Journal of the American Medical Informatics Association* 31(9). 1812–1820. DOI: 10.1093/jamia/ocad259.

Jänicke, Stefan, Greta Franzini, Gerik Scheuermann & Muhammad Cheema. 2015. On close and distant reading in digital humanities: A survey and future challenges. A state-of-the-art (STAR) report. In *Eurographics conference on visualization (EuroVis)*, 83–103.

Jarlbrink, Johan & Fredrik Norén. 2023. The rise and fall of 'propaganda' as a positive concept: A digital reading of Swedish parliamentary records, 1867–2019. *Scandinavian Journal of History* 48(3). 379–399. DOI: 10.1080/03468755.2022.2134202.

Jockers, Matthew L. & Ted Underwood. 2015. Text-mining the humanities. In Susuan Schreibman, Ray Siemens & John Unsworth (eds.), *A new companion to digital humanities*, 291–306. Wiley Online Library. DOI: 10.1002/9781118680605.ch20.

Jurafsky, Daniel & James H. Martin. 2008. *Speech and language processing: An introduction to natural language processing, computational linguistics and speech recognition*. 2nd edn. Saddle River, NJ, USA: Prentice Hall.

La Mela, Matti & Ekta Vats. 2023. Automatic classification of historical texts using a BERT model: News about wild berries, 1860-1910. In *Book of abstracts, DH Benelux*. DOI: 10.5281/zenodo.7990442.

Liu, Bing. 2011. *Web data mining: Exploring hyperlinks, contents, and usage data*. 2nd edn. Heidelberg: Springer. DOI: 10.1007/978-3-642-19460-3.

Maen, Adam. 2024. *Excavating text*. https://github.com/CDHUppsala/excavating-text.

Marsland, Stephen. 2009. *Machine learning: An algorithmic perspective*. Boca Raton, FL: Chapman & Hall/CRC.

McDonough, Katherine, Ludovic Moncla & Matje Van de Camp. 2019. Named entity recognition goes to old regime France: Geographic text analysis for early modern French corpora. *International Journal of Geographical Information Science* 33(12). 2498–2522. DOI: 10.1080/13658816.2019.1620235.

McGillivray, Barbara. 2022. How to use word embeddings for natural language processing. In *Sage research methods: Doing research online*. London: SAGE Publications Ltd. DOI: 10.4135/9781529609578.

Mcgillivray, Barbara, Thierry Poibeau & Pablo Ruiz. 2020. Digital humanities and natural language processing: "Je t'aime... Moi non plus". *Digital Humanities Quarterly* 14(2).

Moretti, Franco. 2000. Conjectures on world literature. *New left review* 2(1). 54–68. DOI: 10.64590/hxj.

Moretti, Franco. 2013. *Distant reading*, vol. 93. New York: Verso.

Ohlsson, Claes, Victor Wåhlstrand Skärström & Henrik Björck. 2022. The market as a concept in Swedish parliamentary records from 1867 to 1970: A mixed methods study. *Digital Humanities in the Nordic and Baltic Countries Publications* 4(2). 22–34. DOI: 10.5617/dhnbpub.11258.

Östling, Robert. 2018. Part of speech tagging: Shallow or deep learning? *Northern European Journal of Language Technology* 5. 1–15. DOI: 10.3384/nejlt.2000-1533.1851.

Piqueras, Matias. 2023. *Alt-right formations*. https://github.com/CDHUppsala/alt-right-formations.

Plank, Barbara. 2016. What to do about non-standard (or non-canonical) language in NLP. In *Proceedings of konvens 2016*, 13–20. DOI: https://doi.org/10.48550/arXiv.1608.07836.

Rapp, Reinhard. 2002. The computation of word associations: Comparing syntagmatic and paradigmatic approaches. In *COLING 2002: The 19th international conference on computational linguistics*. https://aclanthology.org/C02-1007/.

Rettberg, Jill Walker. 2022. Algorithmic failure as a humanities methodology: Machine learning's mispredictions identify rich cases for qualitative analysis. *Big Data & Society* 9(2). DOI: 10.1177/20539517221131290.

Russell, Stuart J., Peter Norvig & John F. Canny. 1995. *Artificial intelligence: A modern approach*. Englewood Cliffs: Prentice Hall.

Schofield, Alexandra. 2022. The possibilities and limitations of natural language processing for the humanities. In James O'Sullivan (ed.), *The Bloomsbury handbook to the Digital Humanities*, 169–178. London: Bloomsbury.

Skeppstedt, Maria, Gijs Aangenendt, Vera Danilova & Ylva Söderfeldt. 2024. Topics in periodicals from the Swedish diabetes association 1949 – 1990: Extending the topic modelling tool Topics2Themes with a timeline visualisation. In *Selected papers from the CLARIN annual conference 2023*. DOI: 10.3384/ecp210015.

Skeppstedt, Maria, Magnus Ahltorp, Kostiantyn Kucher, Gijs Aangenendt, Matts Lindström & Ylva Söderfeldt. 2025. The Word Rain visualisation technique applied to digital history: How to visualise, explore and compare texts using semantically structured word clouds. In Elena Volodina, Gerlof Bouma, Dana Dannélls & Dimitrios Kokkinakis (eds.), *Huminfra handbook: Empowering digital and experimental humanities* (NEALT Proceedings Series 59), 147–182. University of Tartu Library. DOI: 10.58009/aere-perennius0175.

So, Richard Jean. 2017. All models are wrong. *PMLA/Publications of the Modern Language Association of America* 132(3). 668–673. DOI: 10.1632/pmla.2017.132.3.668.

Söderfeldt, Ylva, Karl Berglund & Matts Lindström. 2019. Towards mining the history of the active patient: A mixed-methods discourse analysis of the journal Allergia, 1957-1990. In *Uppsala papers in history of ideas* (19). Uppsala: Uppsala University. http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-394915.

Söderfeldt, Ylva, Andrew Burchell, Julia Reed & Maria Skeppstedt. 2025. Topic timelines for enabling close and distant reading of discursive shifts: A pilot case using periodicals of European diabetes organizations. *Journal of Open Humanities Data* 11(1). DOI: 10.5334/johd.286.

Tahmasebi, Nina & Simon Hengchen. 2019. The strengths and pitfalls of large-scale text mining for literary studies. *Samlaren: Tidskrift för forskning om svensk och annan nordisk litteratur* 140. 198–227.

Tamper, Minna, Petri Leskinen & Eero Hyvönen. 2019. Visualizing and analyzing networks of named entities in biographical dictionaries for digital humanities research. In *International conference on computational linguistics*

and intelligent text processing, 199–214. DOI: 10.1007/978-3-031-24337-0_15.

Tasovac, Toma, Nick Budak, Natalia Ermolaev, Andrew Janco & David Lassner. 2023. Bridging the gap between digital humanities and natural language processing: A pedagogical imperative for humanistic NLP. In Lorella Viola & Paul Spence (eds.), *Multilingual Digital Humanities*, 114–126. London: Routledge. DOI: 10.4324/9781003393696-10.

Thompson, Paul, Riza Theresa Batista-Navarro, Georgios Kontonatsios, Jacob Carter, Elizabeth Toon, John McNaught, Carsten Timmermann, Michael Worboys & Sophia Ananiadou. 2016. Text mining the history of medicine. *PLOS One* 11(1). DOI: 10.1371/journal.pone.0144717.

Underwood, Ted. 2019. *Distant horizons: Digital evidence and literary change*. Chicago: University of Chicago Press.

van Strien, Daniel, Kaspar Beelen, Mariona Ardanuy, Kasra Hosseini, Barbara McGillivray & Giovanni Colavizza. 2020. Assessing the impact of OCR quality on downstream NLP tasks. In *Proceedings of the 12th international conference on agents and artificial intelligence*, 484–496. Valletta, Malta: SCITEPRESS. DOI: 10.5220/0009169004840496.

Vats, Ekta. 2022. *BerryBERT*. https://github.com/CDHUppsala/BerryBERT.

List of abbreviations

HTR	Handwritten Text Recognition
KWIC	Key Word in Context
NER	Named Entity Recognition
NLP	Natural Language Processing
OCR	Optical Character Recognition
POS	Part-of-Speech
TF-IDF	Term Frequency - Inverse Document Frequency

Corresponding author

Gijs Aangenendt
Department of History of Science and Ideas
Uppsala University
gijs.aangenendt@idehist.uu.se

5 Empirisk ordforskning som grund för vidareutveckling av *Svensk ordbok utgiven av Svenska Akademien*

Emma Sköldberg
Göteborgs universitet

Kristian Blensenius
Göteborgs universitet

Empirisk ordforskning är central för det lexikografiska arbetet med att uppdatera *Svensk ordbok utgiven av Svenska Akademien* (SO). Detta kapitel beskriver hur olika korpusar och verktyg, främst från Språkbanken Text och Kungliga biblioteket, används för att välja ut, analysera och revidera uppslagsord. Fokus ligger på böjningsuppgifter och betydelsebeskrivningar. Vi diskuterar metodologiska och praktiska utmaningar, inklusive digitaliseringens påverkan på lexikografiskt arbete. Genom moderna språkteknologiska verktyg har processen blivit mer vetenskaplig och effektiv, samtidigt som behovet av ytterligare textmaterial och metodutveckling kvarstår.

1 *Inledning*

Empirisk ordforskning är avgörande för det lexikografiska arbete som bedrivs inom projektet Svenska Akademiens samtidsordböcker vid Göteborgs universitet. I det här kapitlet redogör vi för hur sådan forskning gagnar vidareutvecklingen av definitionsordboken *Svensk ordbok utgiven av Svenska Akademien* (SO) som beskriver svenskt samtidsspråk. Den senaste upplagan av SO är från 2021 och arbetet inför en uppdatering av ordboken pågår (se avsnitt 2.2). Vi diskuterar utifrån ett antal exempel hur olika slags korpusar och verktyg, vilka tillhandahålls av bl.a. Språkbanken Text, används inom ramen för urval och införlivande av nya uppslagsord, revidering av befintliga ordboksartiklar och utmönstring av uppslagsord ur ordboken. För att avgränsa oss fokuserar vi i hög grad på arbetet med två relativt olika typer av information i ordboksartiklarna, nämligen de informationskategorier som rör uppslagsordens böjning respektive betydelsebeskrivning.

HUM INFRA Emma Sköldberg & Kristian Blensenius. 2025. Empirisk ordforskning som grund för vidareutveckling av *Svensk ordbok utgiven av Svenska Akademien*. In Gerlof Bouma, Dana Dannélls, Dimitrios Kokkinakis & Elena Volodina (eds.), *Huminfra handbook: Empowering digital and experimental humanities* (NEALT Proceedings Series 59), 113–144. University of Tartu Library. DOI: 10.58009/aere-perennius0174 © The authors, CC BY 4.0

Kapitlet är upplagt på följande sätt: i avsnitt 2 redogör vi för våra metalexikografiska utgångspunkter inklusive några centrala begrepp inom lexikografin. Vi presenterar även SO, dess bakgrund och ger exempel på hur en ordboksartikel i verket kan se ut. Avsnitt 3 innehåller redogörelser över hur Språkbanken Text ordforskningsplattform Korp respektive Kungliga bibliotekets textresurser används i det lexikografiska arbetet. Därpå, i avsnitt 4, redogör vi för de tre övergripande uppgifterna i samband med arbetet med SO, dvs. att hitta, välja ut och beskriva nya uppslagsord, att uppdatera och vidareutveckla beskrivningen av redan införlivade ord samt att mönstra ut ord. Avsnitt 5 består av två underavsnitt, ett om hur arbetet med att ge böjningsuppgifter bedrivs och ett annat om hur beskrivning av uppslagsordens semantiska egenskaper går till. Kapitlet avslutas med ett kortfattat slutord.

2 *Bakgrund*

2.1 *Metalexikografiska utgångspunkter*

Svenséns lexikografiska handbok (Svensén 2009: 12–38) innehåller en utförlig presentation av olika slags ordböcker. Med utgångspunkt i ordbokens yttre form brukar man exempelvis skilja på pappersordböcker och elektroniska ordböcker som ordboksappar och webbordböcker. I jämförelse med en tryckt ordbok står mer utrymme till lexikografens förfogande i en elektronisk ordbok och detta kan bl.a. påverka antalet uppslagsord och språkprov i verket. E-ordboken är dessutom mer dynamisk och går lättare att uppdatera.

Ordböcker klassificeras också efter antalet språk de behandlar. De kan vara enspråkiga (monolingvala), tvåspråkiga (bilingvala) eller flerspråkiga. En ordbok kan gälla det allmänna ordförrådet, ett visst fackspråk (såsom det juridiska språket), en geografiskt avgränsad varietet (såsom skånska) eller ord tillhörande en viss stilnivå (såsom slang). Därutöver kan en ordbok behandla många olika aspekter på ett visst ordförråd (t.ex. uppslagsordens uttal, böjning, semantik och etymologi) eller endast en viss aspekt, t.ex. de konstruktioner som uppslagsorden kan ingå i. Olika ordböcker behandlar skilda språkskeden, t.ex. fornsvenska eller modern svenska. Därtill kan de vara synkrona eller diakrona och ha mer eller mindre normativa ambitioner.

Vid praktiskt lexikografiskt arbete tar lexikograferna ofta sin utgångspunkt i de informationskategorier som vanligtvis återfinns i en ordboksartikel (jfr Svensén (2009: 6–8)). Informationskategorierna kan delas in på följande sätt, utifrån vad de bidrar med för uppgifter:

- stavning, uttal, böjning (formell information)
- ordklasstillhörighet, konstruktionssätt (syntagmatisk information)

- betydelsebeskrivning (semantisk information)
- språkexempel
- bruklighetskommentar (pragmatisk information)
- historik (etablering, ursprung och släktskap).

I detta kapitel fokuserar vi som sagt på formell information som rör böjning liksom semantisk information i betydelsebeskrivningarna.

Ordböcker är i hög grad är att betrakta som bruksföremål. Lexikografen måste ha klart för sig vilka den aktuella ordboken i första hand vänder sig till eftersom dess form och innehåll ska anpassas efter målgruppens eller målgruppernas behov. En ordbok som vänder sig till vuxna modersmålstalare behöver ha ett annat innehåll än en ordbok för vuxna inlärare på nybörjarnivå, skolbarn etc. (se t.ex. Atkins & Rundell 2008: 28–33, Tarp 2008, Bergenholtz & Bergenholtz 2011). Lexikografen behöver också ta ställning till vilken eller vilka funktioner ordboken ska fylla eftersom också detta påverkar innehåll och utformning i hög grad. Ska den t.ex. användas vid reception (dvs. vid förståelse av tal och skrift) eller vid produktion (dvs. då man vill formulera sig), eller ska den framför allt dokumentera ett visst språk? Exempelvis är betydelseangivelser viktiga vid reception medan uttals- och konstruktionsuppgifter är relevanta vid produktion. Vid dokumentation fyller historiska uppgifter om uppslagsorden en central roll (se t.ex. Malmgren 2009b).

2.2 *SO och SO-redaktionens uppdrag*

SO kan karakteriseras som en enspråkig, synkron, allmänspråklig och i huvudsak deskriptiv ordbok. Med sina ca 65 000 uppslagsord vänder den sig i första hand till användare som har svenska språket som modersmål men också till vuxna inlärare med goda kunskaper i svenska. Målsättningen är att ordboken ska tillgodose flera funktioner. Reception är den primära funktionen, men SO ska också stödja användarnas språkliga produktion (se vidare bl.a. Malmgren 2009a,b).

SO vilar på vetenskaplig grund. I samband med det redaktionella arbetet förhåller sig lexikograferna till såväl etablerad kunskap, återgiven i bl.a. Svensén (2009), Durkin (2015) och Jackson (2022), som till de senaste rönen inom lexikalisk och metalexikografisk forskning, t.ex. resultat av användarstudier (se bl.a. Hult 2016).

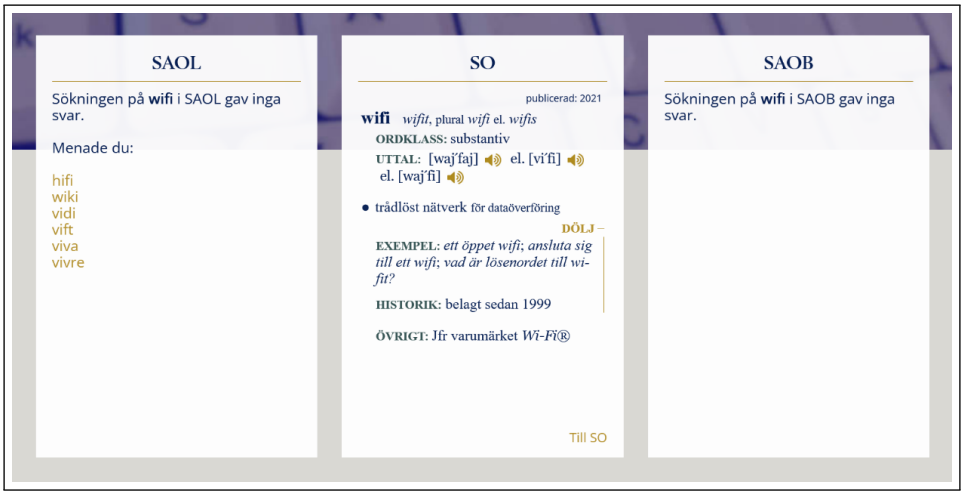
Ordboken bygger på en delmängd av en lexikalisk databas (på ca 200 000 ingångar) som förvaltas och vidareutvecklas vid Göteborgs universitet (GU). Arbetet med databasen påbörjades vid dåvarande Språkdata vid Göteborgs

universitet på 1970-talet i anslutning till projektet ”Lexikalisk databas”. Projektet hade två huvudsyften: dels att ”etablera ett välstrukturerat lexikon lagrat i form av ett länkat nätverk (databas)”, dels att ”databasen skall läggas till grund för en ny svensk enspråkig definitionsordbok, Svensk ordbok” (Ralph m. fl. 1977: 3). Det mer teoretiska syftet var alltså lika centralt som det praktiska, att skapa en ordbok, vilket visar sig i den omfattande och teoretiskt genomarbetade dokumentationen av struktureringen av det lexikaliska innehållet. Ett exempel avser definitionsformatet i databasen: definitionen skulle vara fullständigt utbytbar mot definiendum, dvs. det ord som ska definieras, ”i alla syntaktiska (och även morfologiska) kontexter, vid behov efter applikation av någon naturlig syntaktisk transformation” (Järborg 1989: 22). Detta krav medförde att olika ordklasser fick olika definitionsformat: ett substantiv skulle beskrivas med en obestämd nominalfras, ett adjektiv med en relativsats eller adjektivfras, ett verb med en infinitivfras etc. (se även Josephson 2022: 51, 254–255 om Lexikalisk databas och dess roll inom svensk språkvetenskap).

Den andra upplagan av SO från 2021 är en vidareutveckling av en rad lexikografiska verk vilka har publicerats från 1986 och framåt (se t.ex. Malmgren 1992, 2009a; Malmgren & Sköldberg 2013: 125–126och Borin & Holmer 2024 för en tillbakablick över dessa; se även Holmer m. fl. 2015). Den andra och senaste upplagan publicerades, till skillnad från tidigare upplagor, endast digitalt, i form av gratis nedladdningsbara appar och på Svenska Akademiens ordboksportal, Svenska.se. Se t.ex. Sköldberg (2022, 2024) för redogörelser för viktigare redaktionella ställningstaganden inför denna upplaga samt Trap-Jensen (2022) för en recension av verket.

Sedan år 2010 ägs orddatabasen som SO baseras på av Svenska Akademien och arbetet med denna regleras av ett avtal mellan Svenska Akademien och GU. Den vidareutveckling av databasen, som GU ska ombesörja, består bl.a. i att nya ord, ordformer och ordförbindelser ska införas i databasen så att SO kan utges i bokform och/eller i elektronisk form minst vart tionde år. Vidare ska så fullständiga upplysningar som möjligt beträffande stavning, uttal, böjning, stilvärde och betydelse, illustrerade med språkexempel, ges om alla viktiga ord och uttryck i det svenska ordförrådet, och information ska lämnas om ordens fraseologi och konstruktionssätt. Detta nödvändiggör även en översyn av befintliga ordboksartiklar. Det lexikografiska revideringsarbetet sker i Karp, Språkbanken Texts dataredigeringsplattform, i vilken det är möjligt att söka i och redigera olika slags lexikografiska resurser (se vidare Borin, Sköldberg m. fl. 2025 om plattformen).

I ordboksportalen Svenska.se kan man söka i Svenska Akademiens tre ordböcker, det mer normativa verket *Svenska Akademiens ordlista* (SAOL 14 2015), SO (2021) och den historiska ordboken *Svenska Akademiens ordbok*



Figur 1: SO-artikeln *wifi* i Svenska.se vid sökning i Svenska Akademiens tre ordböcker på Svenska.se

(SAOB 1898–2023) på samma gång (se vidare om Svenska.se i bl.a. Lönnroth 2018 och Bäckerud m. fl. 2020). I portalen kan man även söka i ett verk i taget och användarna har därmed också tillgång till en renodlad SO-webbsida. En fördel med just webbversionen av SO är att den är lättillgänglig för skribenter som arbetar vid datorn, men appversionerna har flera fördelar, t.ex. bättre och fler sökfunktioner liksom funktionen ”Dagens ord”. I figur 1 återges (något beskuret) webbsidans gränssnitt och sökresultatet för substantivet *wifi*. SO-artikeln återfinns i mittenspalten. Som synes är det aktuella uppslagsordet (ännu) inte upptaget i SAOL eller i SAOB.

Av figur 1 framgår att beskrivningen av uppslagsordet *wifi* är förhållandevis kortfattad. I fråga om formell information anges enbart *en* stavning. Däremot varierar ordet böjningsmässigt i de material som redaktionen konsulterat, vilket lett till att två pluralformer ges (*wifi* eller *wifis*). Som synes återfinns ett ”el.” (dvs. ”eller”) mellan formerna, vilket ska tolkas som att formerna är i princip likvärdiga (se s. X i inledningen till SO (2009); se även avsnitt 5.1).

Uttalet ges både i skrift och med länkade ljudfiler med inläst uttal. Som påpekas av bl.a. Sköldberg (2017: 126) är det, om man jämför med uppgifter om exempelvis ords stavning och böjningsmönster, svårare för lexikografen att bilda sig en uppfattning om hur ord uttalas. Detta beror inte minst på bristen på större svenska talspråkskorpusar. SO-lexikograferna har här gjort bedömningen att ordet *wifi* uttalas på tre sätt i svenskan: [waj´faj] el. [vi´fi] el. [waj´fi].

När det kommer till syntagmatisk information meddelas att ordet tillhör ordklassen substantiv. *Wifi* har också endast en betydelse, vilken förmed-

las med definitionen ”trådlöst nätverk” och definitionstillägget ”för dataöverföring” (i mindre stil) (se vidare om betydelsebeskrivningarna i SO i avsnitt 5.2). Artikeln innehåller därutöver tre kortare språkprov, varav det sista utgörs av en fråga: *ett öppet wifi*; *ansluta sig till ett wifi*; *vad är lösenordet till wifit?*. Målsättningen är att exemplen ska illustrera uppslagsordets fraseologi och fungera som ett komplement till definitionen. Någon pragmatisk information ges inte i artikeln, vilket tyder på att ordet bedöms vara omarkerat med avseende på bl.a. stil och bruklighet. (Se Svensén 2009: 315 för genomgång av olika sätt att markera att ett uppslagsord avviker från huvuddelen av de lexikala enheter som beskrivs.) Till skillnad från många andra ordboksartiklar i SO innehåller artikeln *wifi* inga hänvisningar till betydelsemässigt relaterade, uppslagsord. Slutligen, angående etablering, ursprung och släktskap, anges att ordet är belagt i skrivna svenska texter sedan 1999, men det finns ingen explicit uppgift om var det första belägget har påträffats. Lite speciellt i just denna ordboksartikel är också upplysningen att uppslagsordet i hög grad sammanfaller med det registrerade varumärket *Wi-Fi* (se vidare bl.a. Svensén 2009: 428–429 och Agerbo 2018: 198–202 om varumärken ur ett lexikografiskt perspektiv).

I SO (2021) har redaktionen utnyttjat många av den digitala ordbokens fördelar och bl.a. utvecklat den interna sökstrukturen i form av ett stort antal nytillagda klickbara länkar. Dessa underlättar för användarna att röra sig inom ordboken. De böjningsformer som anges är också fullt utskrivna. (I tidigare upplagor hade man, för att förkorta, ersatt ordets uppslagsform med ett tilde och skrivit ”*~t*, plural *wifi* el. *~s*”.) Det kan dock konstateras att SO ännu är starkt präglad av det tryckta formatet. I en elektronisk ordbok, och då i synnerhet i en webbordbok, behöver lexikografen inte använda sig av samma komprimeringsstrategier som i en tryckt ordbok, men många konventioner som utvecklades i samband med utarbetandet av tryckta ordböcker håller i sig ännu och kan vara svåra att frigöra sig ifrån (se vidare bl.a. Rundell 2015; se även Trap-Jensen 2002 och Holmer 2022: 76, 85–88). I artikeln *wifi* är det främst de länkade ljudfilerna med inläst uttal som tydliggör att det rör sig om en digital resurs. Men också huvudordet i definitionen, ”nätverk” hade kunnat utgöra en klickbar länk till uppslagsordet *nätverk* så att ordboksanvändare vid behov lätt hade kunnat se vad det ordet betyder. Vidare hade vissa förkortade ord kunnat skrivas ut. Fler böjningsformer hade även kunnat presenteras, t.ex. genitivformerna *wifis* och *wifits*.

Den senaste upplagan når via Svenska.se ut till en bredare publik än de tryckta föregångarna. Dagens SO-användare har troligen skiftande erfarenheter av att utnyttja ordböcker och fler användare än tidigare torde utgöras av svenskinlärare.

I och med att SAOL och SO, som båda är samtidsordböcker, visas bredvid varandra i portalen har även deras respektive funktioner diskuterats inom den projektgrupp som står bakom dem båda. Målet har blivit att renodla SO och dess perspektiv. Målsättningen har också lett till att SO (2021) är mer deskriptiv än exempelvis 2009 års upplaga (se vidare Sköldberg 2017, 2024).

Värt att notera är emellertid att det är svårt att tala om renodlat deskriptiva eller renodlat normativa ordböcker. Det är snarare fråga om en gradskillnad dem emellan. Sköldberg (2017: 124) konstaterar att en normativ ordbok alltid bygger på en deskriptiv grund och att en deskriptiv ordbok alltid representerar lexikografens språkliga värderingar. Också deskriptiva ordböcker har således en normativ dimension (jfr även Svensén 2009: 24). Trap-Jensen (2002: 64–65) påpekar också att oavsett den enskilda ordbokens deskriptiva eller normativa inriktning, så betraktar användarna ofta det som behandlas i en ordbok som det korrekta.

3 *Resurser och verktyg vid lexikografiskt arbete*

Beskrivningar i ordböcker baseras på evidens som samlas ur stora mängder text, korpusar. I skrivande stund (september 2024) har ordboksprojektet vid Göteborgs universitet via Språkbanken Texts samlingar tillgång till bl.a. mer än 16 miljarder token med moderna texter. Tillgången till material är högst relevant med tanke på att SO, sedan starten, i stor utsträckning är korpusbaserad (se vidare Malmgren 2009a: 16). Lexikografen behöver ha tillgång till omfattande korpusar med olika slags texter från olika tidsepoker för att kunna hitta nya ord, undersöka ovanliga ord, studera ords spridning, analysera ords formella och innehållsmässiga utveckling, för att nämna några aspekter. Innehållet i textmaterialen ligger också till grund för morfologiska och syntaktiska språkexempel i ordboken. För att lexikografen ska kunna hantera all denna data och se mönster, krävs det olika slags språkteknologiska metoder och verktyg. Sådana metoder och verktyg leder till att det lexikografiska arbetet kan effektiviseras och bli mer objektivt.

I det följande beskrivs Korp och Kungliga bibliotekets Svenska tidningar med betoning på de delar som är viktigast för SO-redaktionens arbete.

3.1 *Korp – Språkbanken Texts ordforskningsplattform*

Som redan antytts spelar innehållet i olika slags referensmaterial sedan länge en avgörande roll inom lexikografisk praktik. För svenskans del är innehållet i ordforskningsplattformen Korp ytterst viktigt (se Borin, Forsberg m. fl. (2025) för en presentation av plattformen). Korp innehåller textsamlingar,

korpusar, med text från exempelvis dagstidningar, skönlitteratur och sociala medier, vilka är uppmärkta med bl.a. språkliga attribut. Det moderna materialet i Korp spänner över 1900- och 2000-talet (https://spraakbanken.gu.se/verktyg/korp/anvandarhandledning). Där finns även delkorpusarna Kubord (se avsnitt 3.2). Det empiriska material som legat till grund för flertalet äldre ordböcker har i huvudsak bestått av tryckt och redigerat skriftspråk i form av tidningstext, men även en del romantext. Det är emellertid mycket viktigt att lexikograferna har tillgång till många olika slags texter – såsom de har via Korp – för att de ska kunna ge en så rättvisande bild av hur språket faktiskt används (se vidare Holmer 2022: 85–91 om val av textunderlag vid ordboksarbete; se även Forsberg & Holmer 2024).

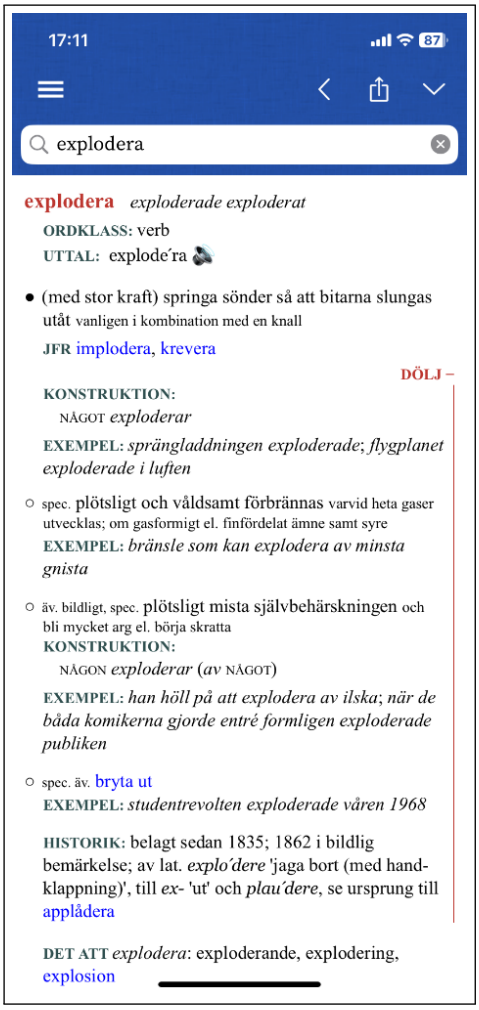
I Korps sökgränssnitt finns tre visningslägen: KWIC-konkordansvyn, Statistikvyn och Ordbildsvyn (Borin, Forsberg m. fl. 2025). Som utgångspunkt för en presentation av dessa tre visningslägen (och deras respektive användningsområden) väljer vi verbet *explodera*. Hur den aktuella ordboksartikeln ser ut i SO (2021) framgår av figur 2.

Av ordboksartikeln i figur 2 framgår bl.a. att ordet *explodera* böjs *exploderade* och *exploderat* i preteritum respektive i supinum. Vidare har ordet följande huvudbetydelse: ’(med stor kraft) springa sönder så att bitarna slungas utåt’. Definitionen följs av definitionstillägget ’vanligen i kombination med en knall’. Parentesen runt ”med stor kraft” betyder att just denna betydelsekomponent kan vara aktuell när ordet används men att det inte alltid är så. Därtill visas tre underbetydelser: ’plötsligt och våldsamt förbrännas (…)’, ’plötsligt mista självbehärskningen (…)’ och ’bryta ut’. Samtliga underbetydelser anges utgöra betydelsespecialiseringar i förhållande till huvudbetydelsen. Såväl den andra som den tredje underbetydelsen är därtill bildlig.

Vad kan man då få fram för information om verbet *explodera* i Korp? Enligt Atkins & Rundell (2008: 104) är KWIC-konkordanser centrala inom korpusbaserad lexikografi. En fördel med konkordanser är att man kan sortera dem på vänster- och högerkontext. De gör det också möjligt för lexikografen att överblicka ords kontexter och därigenom kunna urskilja återkommande mönster för hur ord används.

I figur 3 återges ett konkordansutdrag ur delkorpusen SVT Nyheter (2022) (på drygt 13 miljoner token; Språkbanken Text u. å.) i Korp. Utdraget gäller alla ordformer med grundformen *explodera* (med sammanlagt 288 belägg i materialet).

KWIC-vyn visar bl.a. olika böjningsformer av ordet *explodera*, främst infinitiv, preteritum och supinum. Vidare återfinns flera av de betydelser som verbet anges ha i SO. Exempelvis är huvudbetydelsen aktuell i ett flertal av konkordansraderna, t.ex. i ”Husbil exploderade i Älvsjö” (rad 11), ”[…] en kraftig bomb exploderade […]” (rad 24) samt ”Roboten exploderade strax



Figur 2: Artikeln *explodera* i SO (2021) (app för iOS)

efter att den avfyrats […]" (rad 25). Att just huvudbetydelsen är så vanlig i denna korpus kan eventuell kopplas till att materialet består av nyhetstexter. En av underbetydelserna är aktuell i fall som "Vilket fick San Siro-publiken att explodera av glädje." (rad 1) och "[…Anthony Modeste] stormade fram och nickade in 2-2, vilket fick hemmaarenan att explodera." (rad 3).

Bland konkordansraderna finns det också en bildlig användning av *explodera* som inte behandlas i SO. Se t.ex. rad 16 med meningen "Utbudet har exploderat på Hemnet, från 430 objekt för en månad sen till 610 i Uppsala […]". Verbet *explodera* kan då sägas betyda 'öka kraftigt'. En granskning



SVT NYHETER 2022

Vilket fick San Siro-publiken att	**explodera**	av glädje.
Mycket var bra han gjorde fina galoppombyten men	**exploderade**	sen lite.
·este stormade fram och nickade in 2-2, vilket fick hemmaarenan att	**explodera**	.
; kunde inte resa mig på grund av smärtan och mina ben höll på att	**explodera**	i klättringen, säger van Vleuten.
Efter två mållösa perioder	**exploderade**	Buffalo Sabres med sex mål i den sista perioden.
– Vi har fått uppleva hur damfotbollen har	**exploderat**	på riktigt och vi har blivit förebilder och idoler – och kunnat driva vikt
· även Roope Hintz med sitt tredje mål för matchen, varpå publiken	**exploderade**	och kastade in sina huvudbonader på isen som hyllning till finländare
– Det	**exploderade**	utan att jag hade räknat med det, säger han.
Den verkar ha haft gasolflaskor som kan ha	**exploderat**	.
Vittnesuppgift: Svart väska	**exploderade**	
• Husbil	**exploderade**	i Älvsjö
Flaskan ska sedan ha	**exploderat**	så att de fick ämnet över sig, vilket ledde till hudrodnad och klåda.
Flera medier rapporterar att vattenskotern	**exploderade**	, men enligt polisen finns det inga uppgifter om detta.
– Det har	**exploderat**	sedan årsskiftet.
Sen	**exploderar**	rutan och en bit av en vägskylt tränger in i framrutan.
Utbudet har	**exploderat**	på Hemnet, från 430 objekt för en månad sen till 610 i Uppsala, säger
– Jag blev så glad inombords och det kändes som om något	**exploderade**	eftersom jag blev så glad när vi sågs.
Rasmus bil	**exploderade**	på parkeringen: ” Helvetet släpptes lös ”
– Än så länge vet vi inte vad som	**exploderat**	, men
·ppsdelar fattas, säger Rasmus Sundström, som satt i bilen när den	**exploderade**	i lördags i Enköping.
– För fem år sedan fick det en egen vägg i butiken, efter det har det	**exploderat**	, säger Henry Brandt, butikschef på Partyland i Gränbystaden.
– Det var ett 20-tal elever och lärare vid stugan när det	**exploderade**	och en gasbehållare började brinna.
Det rör sig troligtvis om en ballistisk robot som	**exploderade**	relativt kort efter att den avfyrats, på en höjd under 20 kilometer, upp
56 människor har omkommit efter att en kraftig bomb	**exploderade**	i en shiamuslimsk moské i provinshuvudstaden Peshawar i nordvästr
Roboten	**exploderade**	strax efter att den avfyrats, enligt Sydkoreas militär.

Figur 3: Del av konkordansutdrag för *explodera* i alla ordformer i SVT Nyheter 2022 i Korp

av fler användningar av verbet i samma korpus visar att det finns fler fall där ordet tycks betyda detsamma, bl.a. ”Skulderna har exploderat i storlek [...]” och ”svenskarnas poddlyssnande exploderar”. Detta faktum talar för att ännu en bildlig användning av ordet ska inkluderas i nästa upplaga av ordboken.

Enligt Statistikvyn är det ordformen *exploderade* som är vanligast förekommande i delkorpusen. Den här typen av resultat kan ligga till grund för beslut om vilken eller vilka böjningsformer som kan vara lämpliga att använda i ordbokens språkexempel för att dessa ska spegla bruket av ordet så bra som möjligt. Via denna vy kan man relativt lätt också jämföra bruket av olika stavningsvarianter av ett och samma ord och, om man söker på ordet som för- eller efterled, studera vilka sammansättningar och avledningar som ett ord ingår i.

Även Ordbildsvyn är ett viktigt verktyg för lexikografen. I en ordbild visas det sökta ordet tillsammans med andra ord, vilka ordet har olika syntaktiska relationer till, i den aktuella korpusen. För ett verb visas bl.a. listor över återkommande subjekt och objekt och för ett substantiv visas vanliga adjektiv och verb. Ordbildsvyn bidrar till att lexikografen bättre kan överblicka vad



Figur 4: Ordbild av verbet *explodera* i delkorpusen SVT Nyheter 2022 i Korp.

som förekommer i det sökta ordets kontext, vilket är nödvändigt för att hen ska kunna åstadkomma en så rättvisande bild av ordets syntagmatiska egenskaper som möjligt. Men ordbilder kan även tydliggöra olika betydelser hos ett ord (se avsnitt 5.2). I figur 4 återges en ordbild av verbet *explodera* (baserad på korpusen SVT Nyheter 2022).

Av figur 4 framgår bl.a. att *bomb*, *petflaska* eller *gasflaskor*, men också *intresse* och *efterfrågan* är vanliga subjekt till *explodera* i den aktuella korpusen. De tre första orden kan kopplas till huvudbetydelsen i SO och de två senare till den ännu inte beskrivna underbetydelsen. Vidare listas vanligare adverbial som t.ex. *i luft* och *vid trappuppgång*. Mittenkolumnen visar att verbet oftast inte tar objekt (de övriga fallen synes vid en närmare kontroll vara felparsade subjekt).

Ur lexikografisk synpunkt kan flera av de ord som uppträder i ordbilden för SVT Nyheter 2022 vara med i SO-artikeln *explodera*, men för att beskrivningen av det aktuella uppslagsordet ska bli så heltäckande som möjligt krävs det att flera olika slags texter, bl.a. nyhetstexter, skönlitteratur och texter publicerade på sociala medier, ligger till grund för ordbildsvisningarna (se vidare om Korps ordbild ur ett lexikografiskt perspektiv i Sköldberg 2022).

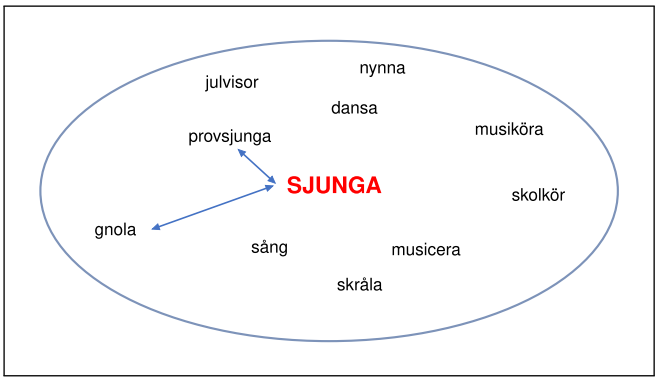
3.2 *KB:s Svenska tidningar och Kubord*

Ytterligare en viktig resurs i det lexikografiska arbetet är de tidningsmaterial som tillhandahålls av Kungliga biblioteket (KB). Tjänsten Svenska tidningar innehåller digitaliserade tidningar och tidskrifter från 1645 till idag. I dagsläget (september 2024) rör det sig om material från närmare 2 000 olika tidnings- och tidskriftstitlar.

SO-lexikograferna söker i KB:s tidningsmaterial främst vid datering av nya uppslagsord, dvs. för att finna första belägg i skrift. Med tiden har det också blivit uppenbart att många av de befintliga angivelserna i SO som rör uppslagsordens förstabelägg kan revideras efter slagningar i KB:s material, och det beror på att den allmänna tillgången till skrivna material är en helt annan idag än under tidigt 1990-tal då många av årtalen skrevs in i SO-databasen. Inte sällan uppmärksammas redaktionen av intresserade ordboksanvändare på att en angivelse om första belägg kan revideras. Att uppdatera alla uppslagsord på denna punkt är en mycket omfattande uppgift och den prioriteras inte just nu inom ordboksprojektet. Men några av de uppslagsord som kommer att få uppdaterade årtalsuppgifter i nästa upplaga av SO är *barbecue* (nuvarande förstabelägg i SO: 1989; uppdaterat förstabelägg: 1868), *espresso* (nuvarande angivelse: ca 1960; uppdaterat förstabelägg: 1924), *illustrativ* (nuvarande förstabelägg: 1906; uppdaterat förstabelägg: 1863) och *tsunami* (nuvarande förstabelägg: 1973; uppdaterat förstabelägg: 1942).

Av upphovsrättsliga skäl är tillgången till de tidningstexter som KB förvaltar begränsad, och detta vållar bekymmer för lexikografer och andra språkforskare som undersöker det svenska ordförrådet, liksom för exempelvis historiker och samhällsvetare. KB-labb vid Kungliga biblioteket och Språkbanken Text har av detta skäl tagit sig an uppgiften att skapa datasamlingar baserade på samtida tidningstexter som kan stödja ordforskning utan att för den sakens skull kränka upphovsrätten. Samarbetet mellan KB-labb och Språkbanken Text har bland annat resulterat i datasamlingarna Kubord 1 och Kubord 2, som är fritt tillgängliga via Korp. Därtill är ännu en datasamling, Kubord-fasttext, på väg att bli publik (se bl.a. Forsberg & Holmer 2024 om dessa samlingar).

Kubord 1 bygger på drygt 80 delsamlingar med olika pressmaterial, sammanlagt lite mer än 3 miljarder token. Tidningsmaterialen är publicerade mellan åren 2000–2021 i bland annat nationella morgon- och kvällstidningar liksom i landsortspress. På grund av upphovsrätten kan Kubord 1 inte göras sökbar på samma sätt som andra korpusar i Korp. Det går att slå upp ord, men förekomsterna visas inte i konkordansformat. Som redan framgått är tillgången till ords kontext viktigt vid lexikografiskt arbete (se avsnitt 3.1).



Figur 5: Vektorrymd med några grannar för verbet *sjunga* (exemplet är hämtat ur Forsberg & Sköldberg, under utgivning)

Det sammanhang som orden används i behövs för att lexikografen ska kunna avgöra ett ords betydelser och fastslå i vilka konstruktioner ordet typiskt uppträder i. I Kubord 1 kan man dock få bl.a. detaljerad frekvensinformation om de ord man vill undersöka och detta räcker långt i samband med vissa lexikografiska uppgifter. Exempelvis kan man, på basis av Kubord 1, undersöka likheter och skillnader i ordförrådet mellan två årgångar av en och samma tidning. Sådana jämförelser är användbara vid det nyordsarbete som bedrivs inom ordboksprojektet (se avsnitt 4.1). Innehållet i Kubord 1 kan även fungera som stöd för utmönstring av uppslagsord (se avsnitt 4.3).

Kubord 2, som i stora drag innehåller samma material som Kubord 1, ger användaren tillgång till statistik över enstaka ord men också över ordpar som står i syntaktisk relation till varandra, såsom verb–subjekt eller substantiv–attribut. Denna kontextuella information möjliggör ordbilder för de ord som förekommer i korpusarna.

Den tredje Kubord-resursen, dvs. Kubord-fasttext, spelar en viktig roll i samband med SO-redaktionens studier av ordvektorer i det lexikografiska arbetet (se Forsberg & Sköldberg 2022, Bouma m. fl. 2024 och Forsberg & Sköldberg under utgivning).Arbetet med ordvektorer ger information om ord vars språkliga kontexter liknar varandra. På så sätt kompletterar ordvektorerna de uppgifter lexikografen får fram via konkordanser och ordbilder. Figur 5 illustrerar en vektorrymd för verbet *sjunga*. På bilden återges ett urval ord som hamnar nära, eller som man säger, utgör grannar.

Bland grannarna i en vektorrymd återfinns till exempel ofta semantiskt besläktade ord till det ord som studeras. Dessa grannord kan ligga till grund för fler hänvisningar till synonyma, antonyma och kohyponyma ord inom

ordboksartiklarna. Studier av vektorrymder kan därmed leda till tillägg i ordboken som förtydligar kopplingarna mellan dess uppslagsord. Bland grannarna finns det också många sammansättningar som kan tjäna som morfologiska språkprov i ordboksartiklar. Därtill kan ordvektorer ringa in olika slags semantiska fält och ge information om ords värdeladdning. Bland grannarna till verbet *sjunga* återfinns sammansättningar som *provsjunga*, mer eller mindre synonyma ord som *skråla*, och ord som kan föras till samma semantiska fält, t.ex. *musicera*. Där finns även för detta verb typiska objekt som *sång*. De aktuella grannarna står således i olika relation till det undersökta verbet. Grannarnas avstånd från *sjunga* ger också en indikation om graden av likhet i fråga om ordens kontext. Grannen *provsjunga* ligger närmare *sjunga* än ett annat verb, *gnola*. Det innebär att de sammanhang där *provsjunga* används liknar mer de sammanhang där *sjunga* förekommer, jämfört med de kontexter där *gnola* används (se bl.a. Bouma m. fl. 2024 och Forsberg & Sköldberg under utgivning för fler exempel).

I nuläget sträcker sig Kubord-materialen som sagt fram till och med år 2021, men materialen kommer successivt att kompletteras med nyare tidningstexter, vilket är avgörande för ordforskare som inriktar sig på bl.a. nyord.

4 *Tre övergripande uppgifter i arbetet med SO*

Det redaktionella arbetet med SO kan delas in i tre huvuduppgifter: 1) nyordsjakt, urval av uppslagsord och författande av nya ordboksartiklar, 2) revidering av befintliga ordboksartiklar och 3) utmönstring av ordboksartiklar.

4.1 *Nyordsjakt, urval av uppslagsord och författande av nya ordboksartiklar*

SO-redaktörerna hyser av förståeliga skäl stort intresse för nya svenska ord. Med nya ord menas här ord som skapats och spridits i svenska texter på senare år, och därmed kan sägas ingå i svenskan. Ett exempel är substantivet *lårskav* (’skavsår på lårets insida orsakat av långvarig friktion’) som används allt mer sedan 2018.[1] Med nya ord avses också (fack)ord som på senare tid ökat i användning i allmänspråket, t.ex. växtbeteckningen *jätteloka* som återfinns i svenska texter sedan åtminstone 1961. Nya svenska ord kan antingen bli uppslagsord eller fungera som språkexempel i form av sam-

1 Ordet är med på Språkrådets nyordslista för året 2018 (se https://www.isof.se/download/18.1e4309991774c3fe50b6ea40/1612877399497/Nyordslistan%202018.pdf; se även Karlsson 2021). Ordet behandlas sedan år 2021 i SO.

mansättningar eller avledningar i ordboken. Utöver den här typen av ord är redaktörerna också intresserade av äldre ord som av någon anledning inte fångats upp och införlivats i tidigare upplagor av SO, t.ex. *torrvara*, som är belagt i svenska tidningar åtminstone sedan 1730-talet och som nu är inlagt i SO-databasen inför nästa uppdatering. Nyare ord intresserar för övrigt inte bara lexikografer utan även andra språkvetare, språkteknologer, historiker och samhällsvetare, eftersom ett språks ordförråd och dess utveckling säger något om det omkringliggande samhället och hur det förändras.

För att bli varse nya ord som kan läggas in i databasen använder sig medarbetarna inom ordboksprojektet av en metod som i korthet går ut på att ordförrådet i en viss korpus jämförs med ordförrådet i en annan korpus och så tar de fasta på de ord som finns i den ena textsamlingen men inte i den andra. De ord som är nyskapade eller har ökat i bruk jämförs sedan med de ord som redan behandlas i ordboken och så får redaktionen en lista över s.k. nyordskandidater att välja ur (se Holmer m. fl. 2024). En jämförelse t.ex. mellan de ord som används i Göteborgs-Posten år 2020 och 2021 visar bland annat att följande ord har ökat i användning mellan de två åren: *barnfridsbrott, charkbricka, fucking, gastropub, hjärndimma, höghöjdsbana, incel, intimitetskoordinator, kulturtolk, miso, mockumentär, mobildata, nagelsalong, naturhus, nettonollutsläpp, näringsterapeut, parallellsamhälle, powerwalk, preklinisk, prosecco, reduktionsplikt, sexsomni, skidalpinism, skyddsperson, streetfood, trygghetspension, utsläppssiffra, verkställighetshinder* och *yes*. Många av dessa ord illustrerar hur flertalet svenska ord bildas. De säger även något om vår samtid, om förändringar i samhället, om nya vetenskaper och ny teknik, om mer eller mindre lyckliga händelser etc. (se Karlsson 2021).

Kubord-materialen kan även jämföras med innehållet i andra slags texter i Korp. Ett flertal väletablerade, men ofta mer informella, svenska ord som ännu inte finns med i SO, har tydliggjorts genom jämförelser mellan Kubord och Poeter.se, en webbsida där skribenterna kan ladda upp sina egna alster. Exempel på sådana ord är *bokmärka, healer, lost, photoshoppa, repeat och vänförfrågan*. I arbetet med SO, som ska beskriva modern svenska, är det viktigt att bruket av också mer vardagliga ord undersöks så att även de orden kan komma i fråga som uppslagsord.

Vid själva urvalet av uppslagsord beaktas i hög grad aspekter som frekvens i allmänspråket och spridning i olika texter. När det kommer till sammansättningar spelar ordens grad av semantisk genomskinlighet en viktig roll. Lexikaliserade sammansättningar tas med i SO, men mer genomskinliga sammansättningar behandlas bara i begränsad omfattning (se t.ex. s. IX i inledningen till SO 2009). Användningen av de ord som väljs ut analyseras grundligt och de aktuella orden förses sedan, i enlighet med avtalet mellan GU och Svenska Akademien, med en rad upplysningar (se avsnitt 2.2).

4.2 *Revidering av befintliga ordboksartiklar*

Som redan framgått har innehållet i SO utvecklats under decennier. En mycket stor andel av SO:s artiklar tillkom på 1980-talet inför den första upplagan av verket och dessa artiklar behöver av olika skäl bearbetas. En anledning är att språket har förändrats sedan ordboksartiklarna skrevs. Exempelvis kan, som vid fallet *explodera*, en betydelse ha tillkommit. Ett ytterligare sådant exempel är *tänka* som, troligen under inflytande från engelskan, numera också betyder 'tycka' (se t.ex. Blensenius 2015). Ett annat skäl är att dagens lexikografer har tillgång till större korpusar och till korpusar med olika slags språk. Exempelvis kan större korpusar med texter från sociala medier innehålla stavningsvarianter som man inte finner i mindre korpusar baserade på redaktionella texter. Här kan förhållandet mellan stavningsvarianterna *spray* och *sprej* nämnas. I tidigare upplagor, inklusive SO (2009), finner man angivelsen "**sprej** el. **spray**", men i SO (2021) har varianternas inbördes ordning ändrats. Angivelsen "**spray** eller **sprej**" är helt enkelt mer rättvisande när det gäller hur ordet stavas i olika texter av idag. En tredje orsak till att äldre ordboksartiklar behöver revideras är att SO nu enbart är digital och att redaktionen så långt det är möjligt vill utnyttja det digitala formatet och bl.a. skapa tydligare samband inom ordboken med hjälp av hänvisningar i form av klickbara länkar.

För att säkra SO:s kvalitet är det nödvändigt att närma sig ordboksinnehållet ur flera perspektiv. Granskningen av materialet kan med fördel ske genom 1) läsning av ordboksartiklar från bokstäverna A–Ö, 2) närstudier av en informationskategori i taget (t.ex. uttal, böjning, bruklighetskommentar) och 3) tematiska genomgångar av olika domäner (juridik, medicin etc.). Som exempel kan nämnas att redaktionen inför SO (2021) gjorde en omfattande genomgång av många tusen konstruktionsuppgifter, som finns typiskt i verbartiklar (men även i ett stort antal substantiv- och adjektivartiklar) och som har till uppgift att ange hur uppslagsord konstrueras med andra ord (se Blensenius 2019). En annan omfattande genomgång gällde ord som kan föras till ämnet pedagogik. Mellan olika upplagor har det omgivande samhället, inklusive det svenska utbildningssystemet, förändrats. Dessa förändringar kan också avläsas i språkbruket såsom det återspeglas i bl.a. Språkbanken Texts korpusar. Ett konkret fall är uppslagsordet **rektor.** Jämför hur substantivet beskrivs i SO (2009) respektive i SO (2021) (de båda artiklarnas innehåll återges i exempel 1 och 2):

(1) **rektor** […]
• […] (titel för) högste ledare för skola eller högskola {JFR **studierektor**}: *rektorsexpedition; bråkstakarna skickades till ~; ett enskilt samtal med ~*
KONSTR.: ~ (*för* el. *på* el. *vid ngt*)
HIST.: sedan 1612; av lat. *re'ctor* 'styresman'; till **regera** (SO 2009)

(2) **rektor** […]
• […] (titel för) högsta administrativa och pedagogiska ledaren för skola, universitet, högskola eller annan utbildningsinstitution
JFR **studierektor**
SAMMANSÄTTN./AVLEDNING: *rektorsexpedition*; *rektorsutbildning*;
KONSTRUKTION:
▸ *rektor* (*för*/*på*/*vid* NÅGOT)
EXEMPEL: *eleven bad om ett enskilt samtal med rektor; som rektor för skolan har hon både personal- och budgetansvar; en biträdande rektor med ansvar för den övergripande pedagogiska planeringen*
HISTORIK: belagt sedan 1612; av lat. *re'ctor* 'styresman'; till **regera** (SO 2021)

Av ordboksartikeln från 2021 framgår det bl.a. att rektorer med tiden har fått förändrade arbetsuppgifter och att avståndet mellan rektor och elever har minskat.

Ytterligare en specialsatsning gjordes på s.k. kontroversiella ord med koppling till olika diskrimineringsgrunder. Några sådana ord är *hora*, *blatte*, *rödskinn*, *funktionsvariation*, *miffo*, *bög* och *hedersrelaterad* (se t.ex. Petersson & Sköldberg 2020a,b). Arbetet gick bl.a. ut på att revidera uppsättningen med uppslagsord, att omformulera betydelsebeskrivningar och att modernisera språkprov så att de blev mer könsneutrala. Till arbetet kom även att uppdatera bruklighetskommentarer i artiklarna, t.ex. ”kan uppfattas som nedsättande” och ”stark nedsättande”. Petersson & Sköldberg (2020a: 112) lyfter dock de begränsningar som innehållet i korpusar och andra textdatabaser har i just detta sammanhang:

> En relevant fråga […] är hur lexikografer hittar de ord som är – eller kan uppfattas som – diskriminerande i den databas som de arbetar med och ska revidera. […]. [S]pråket utvecklas snabbt och ett ord som tidigare varit förhållandevis neutralt kan fort bli kontroversiellt […]. En central fråga är också hur lexikografen avgör vad ett ord har för laddning. Det kan finnas källor att utgå från, såsom t.ex. RFSL:s ordlista, men de passar inte alltid lexikografens syfte.

Det är alltså uppenbart att innehållet i korpusar utgör ett bättre stöd i samband med vissa lexikografiska uppgifter, såsom vid angivelse av ett ords böjning, än vid andra arbetsuppgifter.

4.3 *Utmönstring av uppslagsord*

Ordförrådet utvecklas hela tiden och det är naturligt att vissa av orden i olika upplagor av SO, inklusive den andra från 2021, inte längre hör till det samtida svenska ordförrådet. Dessa ord riskerar att mönstras ut. Förteckningar med så kallade utmönstringskandidater utarbetas genom att listan med uppslagsord i SO jämförs med ordförrådet i korpusar över samtida språk i Korp. Den uppsättning korpusar som SO-redaktionen på senare tid valt att jämföra uppslagsorden med består av bl.a. tidningstexter och nyhetstexter från SVT (båda materialen är från åren 2010–2021), innehållet i den svenska versionen av Wikipedia samt texter publicerade på sajten Poeter.se med användargenererat innehåll. Allt som allt innehåller korpusen ca 3 miljarder token. En jämförelse mellan uppslagsorden och ordförrådet i korpusarna visar att SO innehåller ett stort antal ord som inte förekommer – i någon av deras respektive böjningsformer – i de aktuella korpusarna (jfr Holmer m. fl. 2024 för liknande jämförelser mellan förteckningen över uppslagsord i SAOL och innehållet i samma textmaterial). Exempel på sådana uppslagsord är *kabinettspolitik, kakelugnsnisch, kakelugnssättare, kallhus, kalstammig, kardioskleros, kartusianerorden, kassaköp, katalysatorrenad, katetrisera, kautschukparagraf* respektive *raderkniv, radiokol, [1]rallentando, [2]rallentando, realkatalog, reprografi, rimsalta, rousseauan, rönnbärsfilosofi*.

En mer kvalitativ genomgång av listan med utmönstringskandidater visar att många av dessa utgörs av beteckningar på föremål, företeelser, aktiviteter etc. som är ovanligare, knappt används eller är inaktuella i det svenska samhället av idag (t.ex. *kakelugnssättare* och *raderkniv*). Bland utmönstringskandidaterna finns också många fackord, exempelvis det medicinska ordet *kardioskleros* liksom de homografa musiktermerna *rallentando* (adverb och substantiv). Eventuellt ska ord som dessa inte vara med i allmänspråkliga SO utan listas i fackordböcker av olika slag (jfr avsnitt 2.1).

En annan lexikografisk arbetsuppgift är att granska det inte obetydliga antal ord som är försedda med bruklighetsbeteckningarna ”något ålderdomligt” och ”ålderdomligt”. I SO (2021) finns det ca 1 000 ordboksartiklar som innehåller etiketten ”något ålderdomligt”, t.ex.: *barnförlamning, bröstsocker, grosshandlare, hemlighus, judekörs, malträtera, mentalhygien, sedlighetsbrott, självspilling, stilettkäpp, tankomat, tattare, tvivelsjuk, törhända, vanartig, vidlåda, översiggiven*. Därtill finns det ca 600 ordboksartiklar med etiketten ”ålderdomligt”, exempelvis *dårhus, fjärdingsväg, huskors* och *rucklare*. Även

bland dessa ord finns beteckningar på äldre föremål och produkter, t.ex. *bröstsocker* och *stilettkäpp*. Det finns också beteckningar (t.ex. *barnförlamning* och *judekörs*) som av olika anledningar kommit att ersättas av andra svenska beteckningar (i detta fall *polio* och *japansk lykta*).

Många uppslagsord i SO skulle därmed kunna utmönstras men frågan är om det är nödvändigt. Eftersom också nästa upplaga enbart blir digital finns det, åtminstone i teorin, oändligt med plats i också den (se t.ex. Rundell 2015). Detta innebär att flertalet ord kan vara kvar i SO, även om de är mindre typiska för modernt språkbruk. Å andra sidan, och med tanke på visningen av de tre ordböckerna i Svenska.se, är det inte självklart att synkrona SO, som ska spegla samtida svenska, behöver redogöra för ord som är utförligt beskrivna i SAOB, t.ex. *antvarda, banesår, jungfrudom, kormoran, kruserlig, ostärkt* och *paltig*. Användare som är intresserade av ord som inte längre används kan förutom att konsultera SAOB söka i online-resursen SAOLhist.se vilken bl.a. innehåller SO (2009).

5 *Två informationskategorier i fokus*

I det följande resonerar vi kring två relativt olika informationskategorier i SO och hur lexikografen använder olika empiriska material och verktyg i samband med arbetet med dessa.

5.1 *Information om uppslagsordens böjning*

I SO redovisas böjningsformer på följande sätt för böjliga ord (se SO 2009: XII): för substantiv ges i regel bestämd form singular och obestämd form plural (t.ex. *hunden hundar*), för verb i regel preteritum- och supinumformerna (t.ex. *hoppade hoppat*) och för adjektiv vanligtvis neutrum- och pluralformen (t.ex. *gult gula*). Sedan finns det en rad avvikelser från huvudmönstret. Många substantiv är mer eller mindre oböjliga, och då anges ”ingen böjning”, t.ex. för *fanders*, enstaka verb böjs ogärna i vissa former, t.ex. *stinga*, där upplysningen ”preteritum undviks” ges, och vissa adjektiv böjs ogärna i neutrum, t.ex. *rigid*, vilket också anges i böjningsinformationen.

Vissa ord böjs oregelbundet (t.ex. verb som *gå gick gått* och adjektiv som *bra bättre bäst*), men dessa utgör sällan några utmaningar i ordboksarbetet eftersom det främst är gamla ord som kan föras till denna kategori. Böjningsmönstren kan emellertid förändras över tid även för befintliga ord, som när svaga böjningsmönster tar över från starka böjningsmönster. Ett exempel på detta är mönstret *simma sam summit* som konkurrerar med det i dag betydligt frekventare *simma simmade simmat* (att *sam summit* är mindre

frekventa former markeras i SO av att de föregås av ”äv.”). Regionala varianter som preteritumformen *lös* till *lysa* är också ett exempel på ord där böjningsförändringar över tid är möjliga och därmed särskilt viktiga att kunna följa.

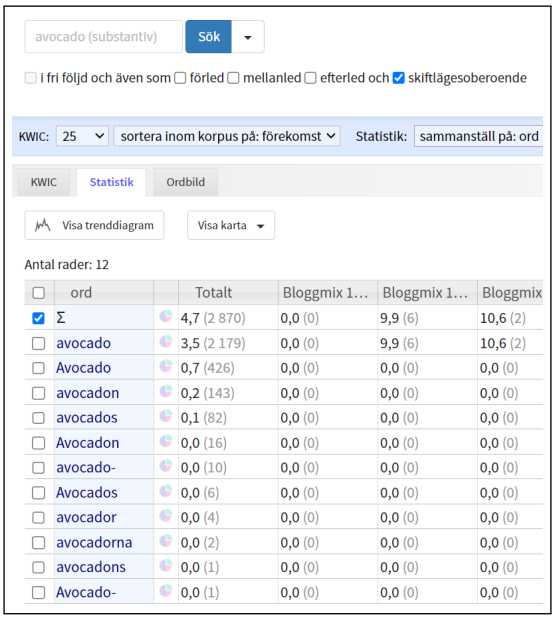
För att ta reda på hur ordens böjningsmönster ser ut i faktiskt språkbruk används Korp och då främst Statistikvyn på ordformsnivå. De korpusar som konsulteras är framför allt nyhetskorpusar, men även skönlitteratur, akademisk text och vissa sociala medier. Eftersom orden ofta är lemmatiserade i lexikonresursen Saldo (se t.ex. Borin & Holmer 2024: 44–45) och därmed innehåller samtliga brukliga böjningsmönster, räcker det ofta att söka på en viss grundform med alla ordformer i en viss korpus. I statistikvisningen ges både frekventa former och ibland eventuella variationer i bruket. SO har som nämnts en deskriptiv inriktning (se avsnitt 2.2), vilket medför att frekvens i hög grad styr hur eventuella variantformer presenteras: om en viss pluralform är frekventast, kommer den att redovisas först.

Utöver Korp kan andra textdatabaser konsulteras för kontroll av specifika ordformer, exempelvis Retriever/Mediearkivet, där de senaste tidnings- och tidskriftstexterna finns representerade och sökbara. Här finns t.ex. även facktidskrifter och (transkriberade) radioprogram representerade, där man kan göra stickprov.

Något lexikografen ibland funderar på i arbetet med böjningsangivelser är ord som böjs på flera olika sätt i bruket. Hur många varianter ska då ges i en mer deskriptiv ordbok som SO, och hur ska dessa förhålla sig till dem i den mer normativa ordlistan SAOL? Ett exempel är uppslagsordet *papper*, för vilket SAOL i artikelhuvudet ger böjningsformerna *papperet* el. *pappret*; plural *papper*, bestämd plural *papperen* el. *pappren*. I SO redovisas fler böjningsformer än så, i alla fall för bestämd plural: *pappret* el. *papperet*, plural *papper*, bestämd plural *pappren* el. *papperna* el. *papperen* el. *papprena*. Hur många alternativa böjningsformer som ska ges och hur de ska redovisas i förhållande till SAOL är fortfarande en öppen fråga.

En annan kategori av ord som böjs på flera olika sätt utgörs av ord som ansluter till flera böjningsmönster, framför allt sådana som har inspirerats av andra språk. Ett sådant fall utgör s-pluralen, som i *enchilada-s* och *wrap-s*. Pluraländelsen berör framför allt nyare ord. Om man skulle lägga in ord som *boomer*, *pinata/piñata* och *youtuber* i SO, skulle man få anledning att ta reda på hur dessa ord böjs i bruket. Här kan man ta hjälp av t.ex. ordformsfrekvenser i Kubord-korpusarna. Det verkar som om s-pluralen dominerar klart om man söker efter ord som börjar på t.ex. ”boom” och ”pinat”. I fallet *pinata/piñata* används förvisso *pinator/piñator* också i viss mån.

Ett annat intressant exempel är de ord där en viss böjningsform sticker ut frekvensmässigt. Att pluraler som *britter* och *pengar* tenderar att vara



ord	Totalt	Bloggmix 1…	Bloggmix 1…	Bloggmix
Σ	4,7 (2 870)	0,0 (0)	9,9 (6)	10,6 (2)
avocado	3,5 (2 179)	0,0 (0)	9,9 (6)	10,6 (2)
Avocado	0,7 (426)	0,0 (0)	0,0 (0)	0,0 (0)
avocadon	0,2 (143)	0,0 (0)	0,0 (0)	0,0 (0)
avocados	0,1 (82)	0,0 (0)	0,0 (0)	0,0 (0)
Avocadon	0,0 (16)	0,0 (0)	0,0 (0)	0,0 (0)
avocado-	0,0 (10)	0,0 (0)	0,0 (0)	0,0 (0)
Avocados	0,0 (6)	0,0 (0)	0,0 (0)	0,0 (0)
avocador	0,0 (4)	0,0 (0)	0,0 (0)	0,0 (0)
avocadorna	0,0 (2)	0,0 (0)	0,0 (0)	0,0 (0)
avocadons	0,0 (1)	0,0 (0)	0,0 (0)	0,0 (0)
Avocado-	0,0 (1)	0,0 (0)	0,0 (0)	0,0 (0)

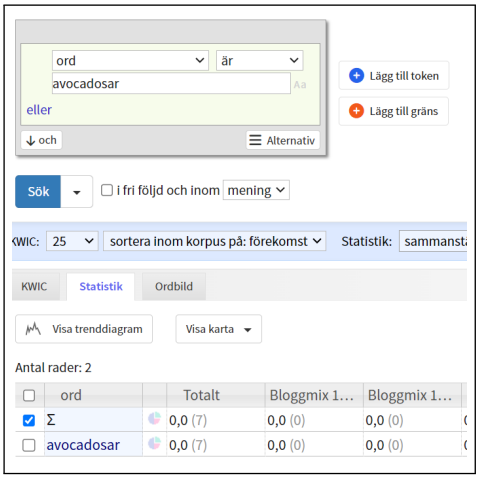
Figur 6: Statistik över sökträffar för *avocado* i alla ordformer som ges i Korp (korpusen Bloggmix)

frekventare än uppslagsformerna *britt* och *peng* är välkänt och anges med anmärkningen ”vanligen plural” i SO. Blensenius & Martens (2020) finner emellertid fler användningar av pluralformer för substantiv som avviker i förhållande till användningen av grundformen, i deras fall i ett material bestående av diskussionsforum. Pluralen *räkor* är där betydligt frekventare än *räka*, *trakter* klart frekventare än *trakt* och *vittnesutsagor* klart frekventare än *vittnesutsaga*. Denna typ av uppgifter går såklart att få fram genom statistik i Korp, men det hade varit bra att integrera någon form av enkel frekvensvisning av ordformer vid varje uppslagsord i redigeringsgränssnittet.

Lexikografen behöver vidare vara medveten om att ordformer kan saknas i ett visst böjningsparadigm för en grundform i Korp. Då kan hen behöva söka efter specifika ordformer. Ett exempel är den vardagliga pluralformen *avocadosar*/*avokadosar* (till *avocado*/*avokado*), som inte återfinns genom sökning efter *avocado*/*avokado* i någon böjningsform. I figur 6 ges exempel med sökning efter *avocado* i alla ordformer som inkluderats i böjningsparadigmet.

Det finns också andra sätt att söka efter potentiella ordformer, t.ex. att sökuttrycket ”Börjar med” *avocado* i Utökad sökning. En sökning efter specifikt *avocadosar* i samma korpus visar t.ex. att *avocadosar* förekommer 7 gånger i korpusen (se figur 7).

Fall som *avocado*/*avokado*, som inte uppvisar någon större stamvariation,



Figur 7: Statistik över sökträffar för ordformen *avocadosar* i Korp (korpusen Bloggmix)

är alltså lätta att söka fram på olika sätt i Korp, men vissa former kan vara svårare att få fram. Ett exempel på en sådan form är den regionala preteritumvarianten *lös* till *lysa*, som togs upp ovan, som inte verkar finnas i *lysa*:s böjningsparadigm i Korp (lysa och lysa[2] i Saldo).

5.2 *Information om uppslagsordens betydelser*

Svenséns lexikografiska handbok (Svensén 2009) innehåller ett fylligt kapitel om hur betydelsebeskrivningar i olika ordböcker kan se ut. I boken finns också utförlig information hur själva arbetet med att fastlägga betydelser går till och vad som kan påverka den betydelsebeskrivning som lexikografen väljer att använda (se även t.ex. Atkins & Rundell 2008: 263–315). Svensén (2009: 205–211) konstaterar bl.a. att innan lexikografen ger sig i kast med att beskriva ett uppslagsords betydelse måste hen ta ställning till om ordet ska anses ha en enda betydelse eller flera delbetydelser. Enligt författaren grundar sig detta ställningstagande i regel på korpusdata, vilket bidrar till att betydelsebeskrivningarna blir mer precisa och rättvisande. Olika slags kriterier, såsom morfologiska, syntagmatiska, semantiska, paradigmatiska och pragmatiska, styr också om analysen resulterar i att det aktuella uppslagsordet tilldelas en eller flera betydelser.

Lew (2013: 287) lyfter å sin sida fram två motsatta strategier, kända som *lumping* och *splitting*, vad gäller betydelseindelning i enspråkiga ordböcker. Den första strategin går ut på att minimera antalet betydelser så att var och en

av dem täcker så stort semantiskt innehåll som möjligt. Den andra strategin tenderar att resultera i ett större antal finindelade betydelser. Lexikografens val av strategi beror ofta på den typ av ordbok som hen arbetar med och ordbokens målgrupp (Atkins & Rundell 2008: 267–268). Den semantiska beskrivningen i en ordbok som främst vänder sig till inlärare på nybörjarnivå kännetecknas ofta av lumping medan den semantiska beskrivningen i en ordbok som vänder sig till lingvister etc. kännetecknas av splitting.

Svensén tar också upp de två olika sätt som betydelserna hos polysema ord kan struktureras. Enligt författaren (2009: 211–212) kan de olika betydelserna ordnas *linjärt*, dvs. som ett antal diskreta enheter i en sekvens, eller *hierarkiskt*, som ett antal huvudbetydelser, till vilka grupper av underbetydelser är associerade. För att förtydliga kan ett ord som *aptitretare* nämnas. Detta ord har dels en mer bokstavlig, dels en mer bildlig betydelse. När polysemistrukturen beskrivs linjärt i ordboken är dessa två betydelser listade som betydelse 1 och betydelse 2 och när polysemistrukturen är hierarkisk listas de två betydelserna som betydelse 1 och betydelse 1a eftersom den andra betydelsen anses vara en bildlig underbetydelse till den första.

I SO ordnas betydelserna, i enlighet med riktlinjerna i Ralph m. fl. (1977) och Järborg (1989), hierarkiskt. I ordboken specificeras dessutom förhållandet mellan huvudbetydelserna och deras respektive underbetydelser (i termer av betydelseutvidgning, betydelsespecialisering, metaforisk användning etc.). I SO är analysen och distinktionen mellan olika betydelser relativt ”finkornig”. SO-lexikograferna är med andra ord snarare splitters än lumpers.

Svensén (2009: 214ff.) tar därtill upp olika typer av ordförklaringar, t.ex. lexikografiska definitioner och synonymförklaringar. Som redan framgått innehåller bl.a. Ralph m. fl. (1977) och Järborg (1989) information om definitionsformat i SO (se avsnitt 2.2) och de riktlinjer som uttrycks där är i många avseenden giltiga än idag. Men innehållet i skrifterna är inte heltäckande. Exempelvis finns det inget skrivet om hur interjektioner ska beskrivas (se Landqvist & Sköldberg 2024). Den nuvarande SO-redaktionen försöker följa traditionen och påtalade principer, men väljer ibland att frångå vissa direktiv bl.a. för att förenkla, inte minst med tanke på nya användargrupper av SO och deras behov.

Eftersom de semantiska uppgifterna i SO är så centrala utgör en viktig del av ordboksarbetet att undersöka *om* betydelserna hos uppslagsorden har förändrats – och i så fall *hur*. Lexikografen upptäcker att den befintliga beskrivningen behöver modifieras, att nya betydelser behöver läggas till m.m. i samband med olika arbetsmoment, t.ex. vid undersökning av nyordskandidater. Som exempel kan nämnas att det bland de senaste nyordkandidaterna till SO finns ett antal sammansättningar som slutar på substantivet *ring*, bl.a.

pedofilring, *narkotikaring* och *reko-ring*. Slagningar i korpusar visar också på fraser i vilka samma betydelse hos *ring* är aktuell, t.ex. *en ring av pedofiler*. Denna betydelse, som kan ha sitt ursprung i engelskans *circle* och som kan jämföras med en betydelse hos ordet *krets*, är inte registrerad i SO sen tidigare. En figurativ underbetydelse till ordet *ring* behöver läggas till i ordboken.

Vidare kan lexikografen bli varse nya betydelser när hen använder sig av Ordbildsvyn i jakt på kollokationer och andra återkommande ordkombinationer. Ett exempel är uppslagsordet *riktmärke* som enligt SO (2021) betyder 'märke som man tar sikte på för att hålla rak kurs t.ex. vid segling'. En ordbildsvisning i Kubord 2 visar att det bland de vanligare framförställda attributen till ordet finns adjektiv som *vetenskaplig* och *ideologisk*. Bland de efterställda attributen finner man ord som *asylmottagande*, *gruppstorlek* och *löneökning*. Dessa ord indikerar att ordet *riktmärke* också används i bildlig bemärkelse, vilket stöds av sökningar i andra korpusar. I exempelvis SVT 2020, som erbjuder KWIC-visningar, uppträder ordet i bl.a. följande meningar (vår fetstil):

(3) - Det är två meters avstånd som är **riktmärket** för att kunna minska smittspridningen.

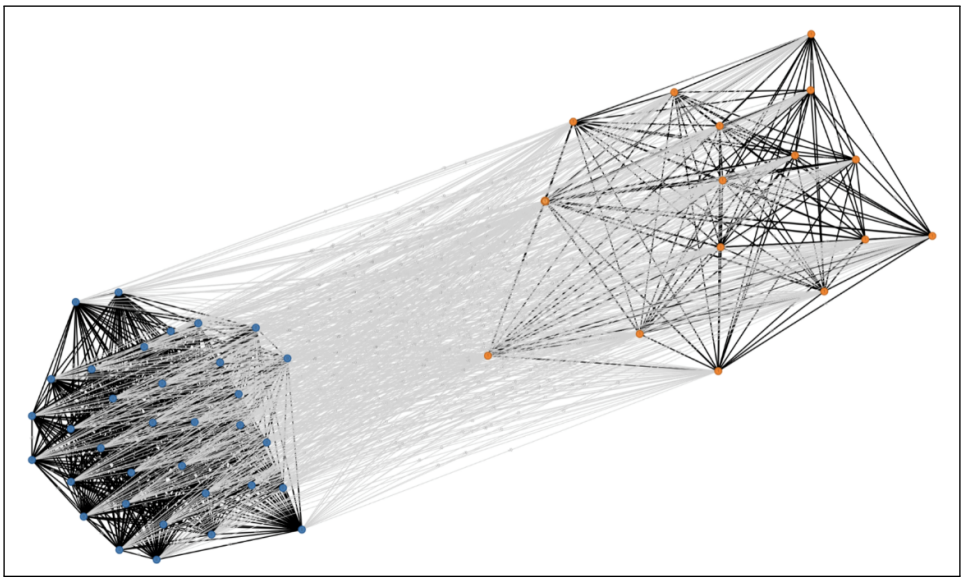
(4) Redan 2016 kom Skolverket med ett **riktmärke** på 6–12 barn för grupper med 1–3-åringar.

(5) Enligt Trafikverket kommer de arbeten som utförs dagtid inte heller att överskrida **riktmärken** för buller- och stomljud i närliggande bostäder.

Allt talar för att substantivet, utöver sin nuvarande betydelse, kommer att ha en bildlig underbetydelse i nästa upplaga av SO.

Andra slags närläsningar bär också frukt. Exempelvis har flertalet interjektionsartiklar reviderats som en följd av en jämförelse mellan innehållet i befintliga interjektionsartiklar och hur interjektioner används i mer talspråknära texter i Korp (se Landqvist & Sköldberg 2024). Undersökningarna har bl.a. lett till att flera interjektioner i nästa upplaga kommer att beskrivas utifrån den eller de funktioner som de fyller istället för med traditionella definitioner med vars hjälp man försöker återge interjektionens betydelse (se även Svensén (2009: 241)). Ett exempel på en sådan interjektion är *heja* som i nuvarande upplaga har betydelsebeskrivningen 'må det gå bra för viss deltagare i tävling eller dylikt'. I nästa upplaga av ordboken kommer uppslagsordet istället att följas av beskrivningen "uttryck för önskan om framgång för viss deltagare i tävling etc.".

Redaktionen uppmärksammas också på svagheter i betydelsebeskrivningarna av ordbokens användare. Ett exempel på detta är den användarkom-



Figur 8: DURels semantiska kluster för adjektivet *enkelspårig*

mentar som gällde uppslagsordet *fruga* med betydelsen 'hustru till viss man'. Liksom många andra ord som försetts med mer inkluderande betydelsebeskrivningar inför SO (2021) behöver också denna betydelse omformuleras så att den blir mer könsneutral (se Petersson & Sköldberg 2020a).

Därtill är SO:s redaktionsmedlemmar delaktiga i studier som rör datorstödda metoder för att upptäcka semantisk förändring på ett systematiskt sätt. Bl.a. har Sköldberg m. fl. (2024) genomfört olika slags experiment i anslutning till annoteringsverktyget DURel som använder sig av språkmodeller för automatisk semantisk analys av ord (se vidare Schlechtweg m. fl. 2018, 2024 om verktyget). I DURel kan man studera ett ords betydelser genom att titta på den semantiska likheten mellan olika meningar som innehåller ordet. Programmet skapar en visualisering med kluster av lika meningar, som är tänkta att motsvara fokusordets eventuella olika betydelser. Inom just dessa experiment är det SVT-korpusen (cirka 200 miljoner tokens; Språkbanken Text u. å. med texter publicerade mellan 2004 och 2021 som använts.

För att illustrera hur gränssnittet i DURel fungerar kan adjektivet *enkelspårig* nämnas. I SO har detta ord dels en huvudbetydelse ('som har bara ett spår'), dels en bildlig underbetydelse ('ensidig, ytlig'). Efter att ha laddat upp 50 slumpmässigt valda meningar ur SVT-korpusen som innehåller ordet i DURel, utfördes alla ytterligare steg i verktygets gränssnitt. Figur 8 visar de två resulterande DURel-klustren för *enkelspårig*, ett med blå prickar och

ett annat med orange prickar, motsvarande de två betydelserna hos detta uppslagsord i SO.

I DURels front-end-gränssnitt kan man, genom att klicka på de blå prickarna, se korpusbelägg som ”Den enkelspåriga järnvägen mellan Motala och Hallsberg är idag en flaskhals […]”. Detta exempel illustrerar sålunda huvudbetydelsen hos adjektivet. Om man klickar på de orange prickarna finner man korpusbelägg som ”De tror att vi är enkelspåriga lantisar, de tror att vi är trångsynta, att vi är rasister och homofober.” Här är således den bildliga betydelsen, som motsvarar underbetydelsen i SO, relevant. Sammanfattningsvis har DURel i detta fall lyckats peka ut de två betydelserna hos ordet.

I ett specifikt experiment utgick Sköldberg m. fl. (2024) från de uppslagsord i SO som endast har en betydelse. Därefter studerade de hur många betydelsekluster av respektive ord som DURel skapade baserat på användningar i SVT-korpusen. Två eller flera semantiska kluster kan tyda på att ordet har mer än en betydelse i texterna och att den semantiska beskrivningen i SO bör revideras. Sådana sökord kan prioriteras för manuell inspektion vid revideringsarbetet inför nästa upplaga. Ett exempel från detta experiment är substantivet *lydnad*, för vilket DURel upptäckte två kluster. Verktyget indikerar med andra ord att substantivet *lydnad* har minst två olika betydelser vilket sedan bekräftades av lexikograferna som inspekterade klustren och korpusutdragen manuellt. De fann att det ena klustret inkluderade korpusexempel som ”Nunnor följer de tre klosterlöftena: att leva i celibat, fattigdom och lydnad.” Detta exempel illustrerar betydelsen som redan finns nedtecknad i SO. Det andra klustret innehåller däremot korpusexempel med ordkombinationer som ”tävla i lydnad” och ”nordiska mästerskap i lydnad”. De representerar en betydelsespecialisering som är aktuell inom hund- och ridsport. Denna specifika betydelse av substantivet finns inte upptagen i den aktuella utgåvan av SO och kommer att läggas till i nästa uppdatering.

Genom studierna i anslutning till DURel har flera insikter kunnat göras. Även om kvaliteten på SO generellt sett är hög, stödjer undersökningarna tidigare slutsatser av t.ex. Blensenius m. fl. (2021) som konstaterar att den semantiska sidan i ordboken kan vidareutvecklas. Experimenten är fortfarande småskaliga, men resultaten är ändå lovande. De datorstödda metoderna förväntas effektivisera och förbättra det lexikografiska arbetet på flera sätt. Hittills har arbetet bland annat uppmärksammat lexikograferna på underbetydelser som bör läggas till ordboken. Även saknade huvudbetydelser har upptäckts. Arbetet har också visat att vissa betydelsebeskrivningar i SO är för generella och att de bör delas upp i enlighet med gällande principer för de semantiska beskrivningarna i ordboken.

6 *Slutord*

I detta kapitel har vi redogjort för den empiriska ordforskning som ligger till grund för det pågående lexikografiska arbetet inom ett ordboksprojekt vid Språkbanken Text, Göteborgs universitet. Arbetet inom projektet syftar bl.a. till att vidareutveckla samtidsordboken *Svensk ordbok utgiven av Svenska Akademien* (SO). Den senaste upplagan av verket publicerades år 2021 och arbete inför en uppdaterad version av ordboken pågår. I kapitlet har vi visat hur olika slags korpusar och verktyg, som tillhandahålls av Språkbanken Text, och andra slags textmaterial, vilka tillgängliggjorts av Kungliga biblioteket, används inom uppdateringsarbetet. Innehållet i olika textmaterial är centralt vid författande av nya ordboksartiklar, uppdatering av befintliga artiklar liksom vid utmönstring av artiklar som inte hör hemma i en mer deskriptiv samtidsordbok. Innehållet i olika slags korpusar är exempelvis avgörande för uppgiften att ge information om uppslagsordens böjning och semantik.

Arbetet med att välja ut lämpliga uppslagsord utgör en viktig uppgift för lexikografen. När orden har valts ut ska de, efter noggrann analys, förses med uppgifter om stavning, uttal, böjning, betydelse, kombinatoriska egenskaper, historik m.m. Dessutom ska språkexempel väljas ut, som visar hur uppslagsorden används. Men ordboksarbetet innefattar sålunda också revidering av befintliga artiklar och även det arbetet inbegriper en rad informationskategorier. Med andra ord aktualiseras flera olika språkvetenskapliga forskningsområden i samband med ordboksarbete. Men lexikografiskt arbete inbegriper också tvärvetenskapliga frågeställningar som vetter mot digital humaniora. Exempelvis aktualiseras frågor kring behandling, analys och presentation av stora och komplexa datamängder.

Praktiskt lexikografiskt arbete är traditionsbundet och alla olika upplagor av SO följer vissa principer som redovisas i t.ex. Ralph m. fl. (1977) och Järborg (1989). Men dessa principer nedtecknades för flera decennier sedan och mycket har förändrats sedan dess, inte minst tillgången till olika slags textmaterial och språkteknologiskt baserade verktyg. Därtill har SO gått från att vara en tryckt ordbok till att vara en heldigital produkt. Det senare har i sin tur medfört att ordboken når ut till andra slags användare (med andra behov). Faktorer som dessa påverkar hur nuvarande redaktion väljer att förhålla sig till de principbeslut som en gång fattades. Exempelvis kan det leda till att lexikografen väljer att avvika från de en gång fastlagda definitionsformaten för att betydelsebeskrivningen ska bli mer begriplig.

Språkbanken Texts korpusar, liksom de tidningsmaterial som är tillgängliga via Kungliga biblioteket, har många styrkor. Materialen utgör ett utmärkt stöd i anslutning till vissa lexikografiska arbetsuppgifter, inte minst när lexikografen vill studera hur ett ord stavas eller böjs. Det är förhållandevis enkelt

att ställa eventuella varianter mot varandra och därigenom avgöra vilka former som ska upptas i ordboken och i vilken ordning de ska presenteras. Textmaterialen är också ovärderliga när lexikografen söker information om hur ord kombineras med varandra eller när hen vill finna konkreta exempel på hur enskilda ord eller uttryck används.

Däremot är det inte lika lätt att, utifrån de textmaterial som står till buds idag, avgöra vad ett ord har för emotiv laddning. Därtill kan bristen på talspråkskorpusar leda till svårigheter att hitta information om hur ord uttalas och hur utpräglat talspråksartade ord, såsom många interjektioner, används.

En sak står dock klar. Tack vara nuvarande korpusar i Språkbanken Text, de texter som är publika via Kungliga biblioteket samt moderna språkteknologibaserade metoder, har det lexikografiska arbetet med bl.a. SO blivit mer vetenskapligt baserat och effektivt. SO-redaktionen arbetar också aktivt för tillgängliggörande av nya textmaterial, metodutveckling och ökat samarbete mellan olika kompetenser för att den vetenskapliga kvaliteten på ordboksarbetet ska bli ännu högre.

Referenser

Agerbo, Heidi. 2018. Funktionsteoretisk diskussion af brugerreaktioner på ordbogsartikler – med fokus på varemærker, omdiskuterede ord og eksempelsætninger. *LexicoNordica* 25. 195–214.

Atkins, B. T. Sue & Michael Rundell. 2008. *The Oxford guide to practical lexicography*. Oxford: Oxford University Press.

Bergenholtz, Henning & Inger Bergenholtz. 2011. A dictionary is a tool, a good dictionary is a monofunctional tool. I Pedro A. Fuertes-Olivera & Henning Bergenholtz (utg.), *e-Lexicography: The internet, digital initiatives and lexicography*, 187–207. London/New York: Continuum.

Blensenius, Kristian. 2015. Jag tänker att. *Språktidningen* 1. 58–61.

Blensenius, Kristian. 2019. Revision av konstruktionsuppgifter i Svensk ordbok utgiven av Svenska Akademien. *LexicoNordica* 26. 203–223.

Blensenius, Kristian & Monica Martens. 2020. Analys av relativa ordformsfrekvenser för en bättre ordbok. *Svenskans beskrivning 37*. 56–69.

Blensenius, Kristian, Emma Sköldberg & Erik Bäckerud. 2021. Finding gaps in semantic descriptions: Visualisation of the cross-reference network in a Swedish monolingual dictionary. *Proceedings of the Electronic lexicography in the 21st century (eLex) 2021 conference*. 247–258.

Borin, Lars, Markus Forsberg, Martin Hammarstedt, Louise Holmer & Arild Matsson. 2025. Korp: Språkbanken's word research platform. I *Sixty years of Swedish computational lexicography*. Dana Dannélls, Kristian Blensenius & Lars Borin (utg.). De Gruyter. 175–194. DOI: 10.1515/9783111577234-010.

Borin, Lars & Louise Holmer. 2024. Tradita innovare, innovata tradere: The Gothenburg approach to computational lexicography. *Proceedings of the Huminfra Conference (HiC 2024)*. 41–50.

Borin, Lars, Emma Sköldberg, Ann Lillieström, Nick Smallbone, Maria Öhrman, Jonatan Uppström & Louise Holmer. 2025. Karp: Språkbanken's data editing platform. I *Sixty years of Swedish computational lexicography*. Dana Dannélls, Kristian Blensenius & Lars Borin (utg.). De Gruyter. 195–210. DOI: 10.1515/9783111577234-011.

Bouma, Gerlof, Markus Forsberg, Justyna Sikora & Emma Sköldberg. 2024. Konsten att bedriva svensk ordforskning utan att kränka upphovsrätten. *Proceedings of the Huminfra Conference (HiC 2024)*. 161–167.

Bäckerud, Erik, Pär Nilsson & Emma Sköldberg. 2020. Så används Svenska Akademiens ordböcker på nätet: Implicit och explicit feedback från användarna. *Nordiska studier i lexikografi 15*. 91–101.

Durkin, Philip (utg.). 2015. *The Oxford handbook of lexicography*. Oxford: Oxford University Press.

Forsberg, Markus & Louise Holmer. 2024. Datatillgång, metodutveckling och lexikografiskt arbete vid Språkbanken Text. *LexicoNordica* 31. 61–79.

Forsberg, Markus & Emma Sköldberg. 2022. Ordvektorer i lexikografiskt arbete. I Elena Volodina, Dana Dannélls, Aleksandrs Berdicevskis, Markus Forsberg & Shafqat Virk (utg.), *Live and learn: Festschrift in honor of Lars Borin*, 37–41. Gothenburg: Department of Swedish, Multilingualism, Language Technology, University of Gothenburg.

Holmer, Louise. 2022. *Neutrala substantiv på* -ande *i text och ordbok*. Gothenburg: Meijerbergs institut för svensk etymologisk forskning.

Holmer, Louise, Ann Lillieström, Emma Sköldberg & Jonatan Uppström. 2024. Time to say goodbye revisited: On the exclusion of headwords from the Swedish Academy Glossary (SAOL). *Proceedings of the European Association for Lexicography (EURALEX) 2024*. 443–452.

Holmer, Louise, Monica von Martens & Emma Sköldberg. 2015. Making a dictionary app from a lexical database: The case of the Contemporary Dictionary of the Swedish Academy. *Proceedings of the Electronic lexicography in the 21st century (eLex 2015) conference*. 32–50.

Hult, Ann-Kristin. 2016. *Ordboksanvändning på nätet: En undersökning av användningen av Lexins svenska lexikon*. Gothenburg: University of Gothenburg. (diss.).

Jackson, Howard (utg.). 2022. *The Bloomsbury handbook of lexicography*. 2. utg. London: Bloomsbury Academic.

Josephson, Olle. 2022. *Språkpolitik*. 2. utg. Stockholm: Morfem.

Järborg, Jerker. 1989. *Betydelseanalys och betydelsebeskrivning i Lexikalisk databas: Preliminär version*. (Research Reports from the Department of Swedish No. GU-ISS-89-01) Gothenburg: Department of Swedish, University of Gothenburg.

Karlsson, Ola. 2021. *Lesserwisser*, *lårskav* och *läppstiftseffekt*: Presentation och problematisering av urvalskriterierna för den svenska nyordslistan. *LexicoNordica* 28. 101–120.

Landqvist, Hans & Emma Sköldberg. 2024. Interjektioner som lexikografisk utmaning: En fallstudie av interjektioner med engelskt ursprung utifrån Svensk ordbok utgiven av Svenska Akademien. *ASLA:s skriftserie/ASLA Studies in Applied Linguistics* 31. 26–55.

Lew, Robert. 2013. Identifying, ordering and defining senses. I Howard Jackson (utg.), *The Bloomsbury companion to lexicography*, 284–302. London: Bloomsbury Publishing.

Lönnroth, Harry. 2018. Portalen svenska.se: En ny digital samlingsplats för språkresurser från Svenska Akademien. *LexicoNordica* 25. 281–292.

Malmgren, Sven-Göran. 1992. From Svensk ordbok ('A dictionary of Swedish') to Nationalencyklopediens ordbok ('The Dictionary of the National Encyclopedia'). *Proceedings of the European Association for Lexicography (EURALEX) 1992*. 485–491.

Malmgren, Sven-Göran. 2009a. Från Nationalencyklopedins ordbok (1995–96) till Svensk ordbok utgiven av Svenska Akademien (2009). *LEDA-Nyt* (47). 14–20.

Malmgren, Sven-Göran. 2009b. On production-oriented information in Swedish monolingual defining dictionaries. I Sandro Nielsen & Sven Tarp (utg.), *Lexicography in the 21st century: In honour of Henning Bergenholtz*, 93–102. Amsterdam: John Benjamins.

Malmgren, Sven-Göran & Emma Sköldberg. 2013. The lexicography of Swedish and other Scandinavian languages. *International Journal of Lexicography* 26(2). 117–134. DOI: 10.1093/ijl/ect008.

Petersson, Stellan & Emma Sköldberg. 2020a. Beskriva utan att diskriminera. Representation av könsidentitet och sexuell läggning i Svensk ordbok. *LexicoNordica* 27. 97–115.

Petersson, Stellan & Emma Sköldberg. 2020b. To discriminate between discrimination and inclusion: A lexicographer's dilemma. *Proceedings of the European Association for Lexicography (EURALEX) 2021*. 382–386.

Ralph, Bo, Jerker Järborg & Sture Allén. 1977. *Svensk ordbok och lexikalisk databas: Förstudierapport*. Gothenburg: Department of Computational Linguistics, University of Gothenburg.

Rundell, Michael. 2015. From print to digital: Implications for dictionary policy and lexicographic conventions. *Lexikos* 25. 301–322.

SAOB. 1898–2023. *Svenska Akademiens ordbok*. Lund: Gleerups. https://www.saob.se/%20and%20https://svenska.se/.

SAOL 14. 2015. *Svenska Akademiens ordlista*. 14. utg. Stockholm: Norstedts. svenska.se.

Schlechtweg, Dominik, Sabine Schulte im Walde & Stefanie Eckmann. 2018. Diachronic usage relatedness (DURel): A framework for the annotation of lexical semantic change. *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics / Human Language Technologies (NAACL/HLT) 2018: Volume 2 (Short Papers)*. 169–174. DOI: 10.18653/v1/N18-2027. https://aclanthology.org/N18-2027.

Schlechtweg, Dominik, Shafqat Virk, Pauline Sander, Emma Sköldberg, Lukas Theuer Linke, Tuo Zhang, Nina Tahmasebi & Sabine Schulte im Walde. 2024. The DURel annotation tool: Human and computational measurement of semantic proximity, sense clusters and semantic change. *Annual Meeting of the European Chapter of the Association for Computational Linguistics (EACL) 2024: System Demonstrations*. 137–149.

Sköldberg, Emma. 2017. Innehållet i Svensk ordbok utgiven av Svenska Akademien – eller kampen mellan norm och bruk. I Saga Bendegard, Ulla Melander Marttala & Maria Westman (utg.), *Språk och norm. Rapport från ASLA:s symposium*, 123–129. Uppsala: ASLA: Svenska föreningen för tillämpad språkvetenskap.

Sköldberg, Emma. 2022. Phraseological theory, evidence in corpora and lexicographical practice: On collocations in a monolingual dictionary of Swedish. I Kristian Blensenius (utg.), *Valency and constructions: Perspectives on combining words*, 155–182. Gothenburg: Meijerbergs institut för svensk etymologisk forskning.

Sköldberg, Emma. 2024. Andra upplagan av Svensk ordbok (SO): Förutsättningar, teoretiska överväganden, insatser och mottagande. I Denny Jansson, Ida Melander, Gustav Westberg & Daroon Yassin Falk (utg.), *Svenskans beskrivning 38: Förhandlingar vid trettioåttonde sammankomsten. Örebro 4–6 maj 2022, Del I*, 165–180. Örebro universitet.

Sköldberg, Emma, Shafqat Virk, Pauline Sander, Simon Hengchen & Dominik Schlechtweg. 2024. Revealing semantic variation in Swedish using computational models of semantic proximity: Results from lexicographical experiments. *Proceedings of the European Association for Lexicography (EURALEX) 2024*. 169–182.

SO. 2009. *Svensk ordbok utgiven av Svenska Akademien*. Stockholm: Svenska Akademien.
SO. 2021. *Svensk ordbok utgiven av Svenska Akademien*. 2. utg. Stockholm: Svenska Akademien.
Språkbanken Text. U. å. *SVT nyheter*. Dataset. DOI: 10.23695/vvd5-j848.
Svensén, Bo. 2009. *A handbook of lexicography: The theory and practice of dictionary-making*. Cambridge: Cambridge University Press.
Tarp, Sven. 2008. *Lexicography in the borderland between knowledge and non-knowledge: General lexicographical theory with particular focus on learner's lexicography*. Tübingen: Max Niemeyer Verlag.
Trap-Jensen, Lars. 2002. Normering og deskription i Den Danske Ordbog – mere eller mindre? *LexicoNordica* 9. 63–78.
Trap-Jensen, Lars. 2022. Svensk ordbok – anden og reviderte udgave. *LexicoNordica* 29. 197–213.

Korresponderande författare

Emma Sköldberg
Institutionen för svenska, flerspråkighet och språkteknologi
Göteborgs universitet
emma.skoldberg@svenska.gu.se

Kristian Blensenius
Institutionen för svenska, flerspråkighet och språkteknologi
Göteborgs universitet
kristian.blensenius@gu.se

Tools, methods and resources

6 The Word Rain visualisation technique applied to digital history: How to visualise, explore and compare texts using semantically structured word clouds

Maria Skeppstedt
Stockholm University

Magnus Ahltorp
Nakajima Koen Research Institute

Kostiantyn Kucher
Linköping University

Gijs Aangenendt
Uppsala University

Matts Lindström
Uppsala University

Ylva Söderfeldt
Uppsala University

The Word Rain text visualisation technique aims to retain the simplicity of the classic word cloud, while addressing some of its limitations. In particular, the Word Rain visualisation uses word embeddings to automatically give the visualised words a semantically meaningful position along the horizontal axis. In this handbook chapter, we showcase how this novel approach for word positioning makes the Word Rain technique suitable for exploring, analysing and comparing texts. More specifically, we show how the Word Rain Python module can be used to visualise longitudinal changes in periodicals published by the Swedish Diabetes Association, and how the Word Rain web service can be used to create visualisations that compare the patient organisation periodicals to journals published by the Swedish Medical Association.

1 *Introduction*

The classic word cloud[1] is a popular technique for providing a quick visual overview of the most important content of a text or text collection. The

1 The terms "word cloud" and "tag cloud" are often used interchangeably, but we will here consistently use the term "word cloud", since the main application scenario targeted is the visualisation of words automatically extracted from a running text, and not the visualisation of tags (Torres Parejo et al. 2021).

HUM INFRA Maria Skeppstedt, Magnus Ahltorp, Kostiantyn Kucher, Gijs Aangenendt, Matts Lindström & Ylva Söderfeldt. 2025. The Word Rain visualisation technique applied to digital history: How to visualise, explore and compare texts using semantically structured word clouds. In Gerlof Bouma, Dana Dannélls, Dimitrios Kokkinakis & Elena Volodina (eds.), *Huminfra handbook: Empowering digital and experimental humanities* (NEALT Proceedings Series 59), 147–182. University of Tartu Library. DOI: 10.58009/aere-perennius0175
© The authors, CC BY 4.0

technique is built on the idea of extracting the most frequent words from the text and displaying them in a graph with a font size that corresponds to their frequency (Cao & Cui 2016: 57–59).

While the result is a visually more appealing—and more spatially dense—alternative to a simple sorted word frequency list, it does not provide any additional analytical power. Neither the frequency list, nor the word cloud, assists you in the task of semantically sorting the most prominent words, or in the task of comparing the most prominent words in two different word clouds/word frequency lists. The layout of the word cloud makes it possible to display a large number of words in a small area by giving the less prominent ones a font size too small to read without zooming in. However, the technique does not provide any guidance to where in the cloud to zoom in to reach potentially interesting words, and therefore does not provide any advantage compared to reading a long word-frequency list.

There are many different types of more analytically powerful developments of the classic word cloud that build on interaction and/or animation (Collins et al. 2009, Koh et al. 2010, Lohmann et al. 2012, Liu et al. 2015, Wang et al. 2018, Xie et al. 2024), and, of course, even more examples of analytically powerful text visualisation techniques in general (Kucher & Kerren 2015), many of them relying on a dynamic and interactive user interface. None of them have, however, been close to achieving the same general level of popularity as the static word cloud.

We believe that one of the key factors for the popularity of the word cloud is the elegance that comes with simplicity. This hypothesis is, for instance, supported by the classic word cloud being recommended (in an information visualisation conference paper) as the first and most important text visualisation technique to teach within digital humanities (Jänicke 2022), as well as by the characterisation of the word cloud by Cao & Cui (2016): "There are reasons for word clouds being user friendly in general. Unlike other visualizations, word clouds need no additional visual artifacts and are self-explanatory. Users read the words and compare their sizes to know which one is more salient." We, therefore, set out to develop a text visualisation technique that—as far as possible—would retain this simplicity of the static word cloud, while also providing functionality that would be practically useful for analysing and comparing the content of text collections. This work resulted in the Word Rain[2] text visualisation technique.

We have previously introduced, explained and evaluated the Word Rain technique (Skeppstedt et al. 2024), and we have also exemplified how the technique can be used practically. We have previously focused on three

2 https://github.com/CDHUppsala/word-rain (last accessed: February 14, 2025)

different use cases: i) to compare texts from different text genres (Skeppstedt et al. 2024), ii) to study longitudinal corpora changes (Skeppstedt & Aangenendt 2024), and iii) to study the content of a text in relation to a dictionary/controlled vocabulary (Ahltorp et al. 2025). In this handbook chapter, we will further explore the first two of these use cases, by also providing concrete instructions—together with links to programming code—for how to generate and analyse the word rain visualisations. We will also provide a more advanced example, by adding a third use case, also for the task of exploring longitudinal corpora change. Each use case has an increasing level of complexity, both when it comes to generating the word rains and to interpreting the results. For the first case, we will use the Word Rain web service. This option provides less flexibility, but instead offers the large advantage of not requiring any programming knowledge at all. Instead, the texts are simply uploaded to the website in order to generate word rains. We will, thereafter, describe a use case that requires a minimum amount of programming code, and then the third one which requires more programming skills.

The chapter is organised as follows: We start by briefly discussing limitations associated with the classic word cloud, and we thereafter describe how these limitations are addressed by the Word Rain technique. In the section that then follows, we describe the concrete example use cases. We provide a short description of the corpora used in the examples, and then—for each example—we describe how the word rains are generated and how to read the resulting word rain graphs. Thereafter, we contrast the Word Rain visualisation technique with other similar techniques, by comparing it to a word frequency list and to the classic word cloud, as well as to previous efforts for improving the classic word cloud. The chapter is then concluded by a short summary.

2 *Limitations of the classic word cloud*

There are many versions of the classic word cloud. For instance, the words in the cloud could either be displayed in a random order or—with a small development of the algorithm—be positioned in alphabetical order (Cao & Cui 2016: 57–59). Instead of just displaying the most frequent *words*, it is also possible to include frequent n-grams (i.e., short word sequences) in the visualisation (Cao & Cui 2016: 43). Similarly, raw frequency is a standard measure used for determining prominence/font size, but there are also other options, such as TF-IDF, i.e., term frequency–inverse document frequency (Xu et al. 2016).

However, there are limitations associated with all these standard versions of the word cloud which decrease the usefulness of the word clouds for many analytical tasks. This is problematic, since the word cloud is a frequently used text visualisation technique (Brath & Banissi 2016).[3] Word clouds are not only used to provide a visually appealing illustration, but are also used for analytical purposes. For instance, Hicke et al. (2022) studied the purpose of including word clouds in 351 news articles and academic articles within digital humanities, and found that in 37% of the articles studied, the purpose of the word cloud was to enable an "analytic exploration".

When empirically comparing the efficiency of different standard design options for word cloud-like representations, it has been shown that the optimal design choice can be different depending on which analytical task is to be carried out (Felix et al. 2018). Design trade-offs can thereby be expected when aiming to support too many tasks. We have therefore mainly focused our attention on a few common analytical tasks. Hicke et al. (2022) provide a taxonomy for describing such tasks where word clouds are used today. From this taxonomy, we focused on the tasks of *finding the most prominent words* in a text, *finding "topics" or semantic word clusters*, and *gist-forming* (i.e. getting the gist of a text). Hicke et al. also found that users of word clouds were often expected to perform *inter-cloud comparison tasks*, i.e., tasks that included a comparison of two, or several, word clouds.

For performing these four tasks, we have identified three important limitations associated with the classic word cloud:

1. The first limitation regards the lack of a meaningful interpretation of how the words are positioned in the word cloud (Barth et al. 2014). The lack of a meaningful positioning might result in the reader incorrectly inferring a meaning to where the words are positioned in the graph. In addition, eye tracking studies have shown that a meaningful placement of the words increases the efficiency in how the graphs are used. More specifically, for finding prominent words, a position based on word prominence is most effective.[4] A thematically clustered layout was, instead, most effective for the aim of finding words belonging to a certain topic (Lohmann et al. 2009). Similarly, for an analytical task-based evaluation, word clouds that grouped words into semantic categories—either separated by different colours, or separated by space—made it easier to solve the task (Hearst et al. 2020). This possibility to employ

3 Among 106 peer reviewed text visualisations investigated, 39 were versions of the word cloud (Brath & Banissi 2016).

4 In the respective study, a circular layout with decreasing word prominence towards the edges was used.

word positioning to support analytical tasks is, however, not used in a classic word cloud.

2. Another limitation related to the lack of a semantically meaningful word positioning, is that the word cloud technique does not provide any guidance regarding where in the cloud to zoom in. This makes it very difficult to explore words rendered in a small font size. The use of a small font size even affects readability negatively when the font size is large enough to be *possible* to read without zooming in (Ren et al. 2024). Thereby, in practice only the most prominent words in the cloud can be used for forming a gist of the text content.

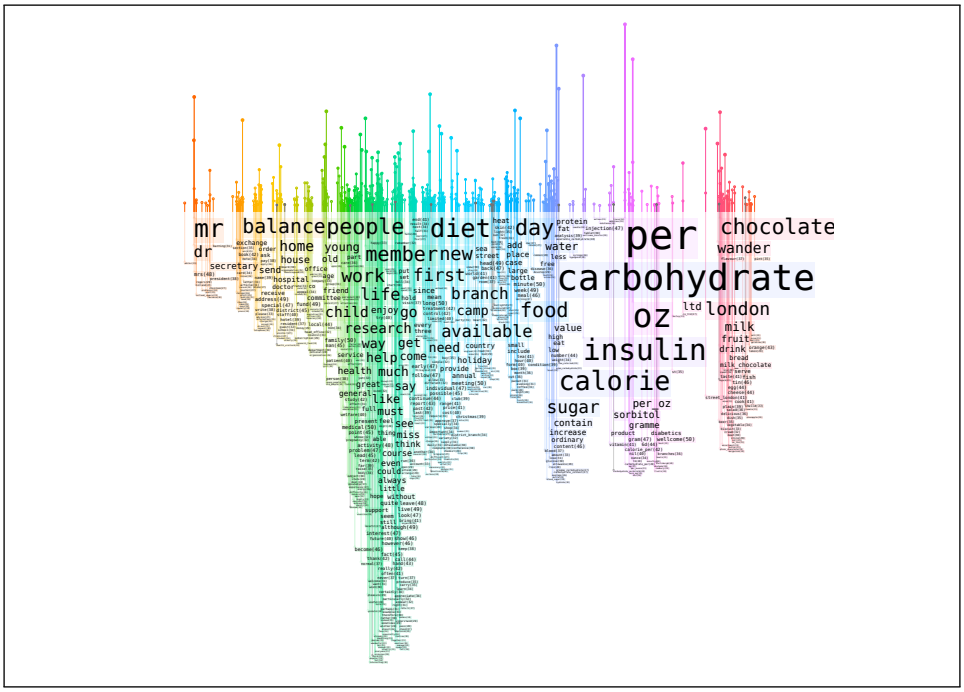
3. The third limitation regards the difficulty to compare word clouds, e.g., when tracking longitudinal changes in a series of word clouds or comparing two word clouds representing two text genres. As expressed by Cao & Cui (2016: 74): "[...] tracking words between different word clouds is the biggest pain for this type of methods". The main reason is the varying position of the same word between different word clouds (Cao & Cui 2016: 69).

Critique has also been directed towards the use of font size as a prominence indicator, since this might result in longer words incorrectly being perceived as more important, as they occupy a larger area in the graph (Viégas & Wattenberg 2008). However, a large font size is a self-explanatory signal for prominence and the use of a small font size enables the creation of dense graphs with many words. We will, therefore, address this limitation by alleviating it, not completely removing font size as a prominence indicator. In addition, results from Alexander et al. (2018)—where an accuracy of over 90% was achieved for a font size comparison task—show that there are situations in which users are able to compensate for length differences.

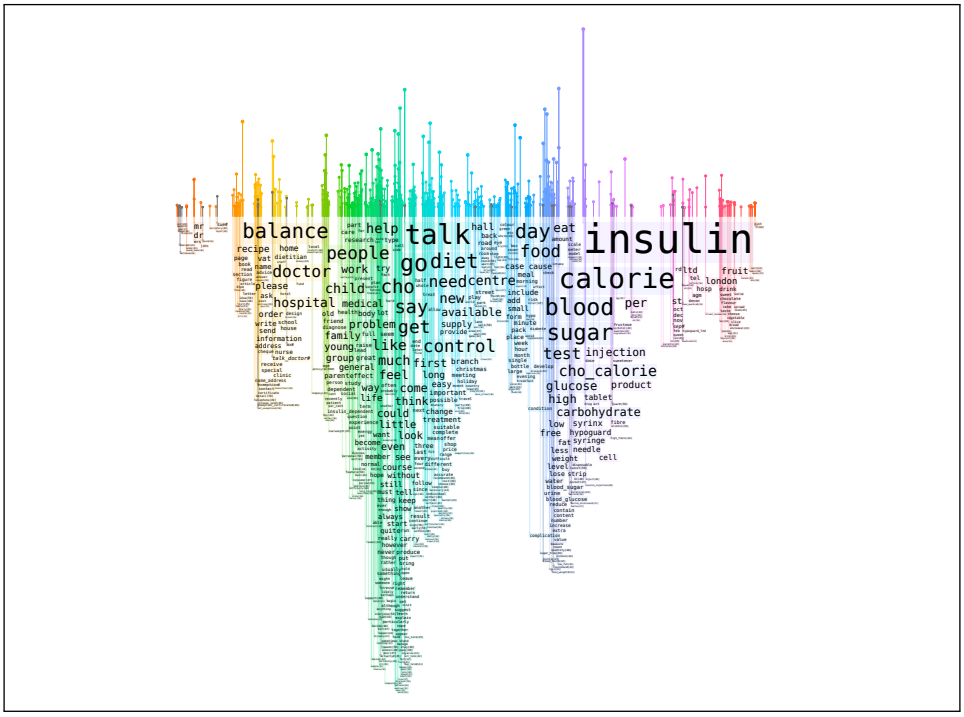
While there are previous extensions of the classic word cloud that aim to solve *one* of the limitations listed above (see Section 5.2), we are not aware of any text visualisation technique that—as the Word Rain technique—addresses all the described limitations.

3 *The Word Rain text visualisation technique*

The aim of the Word Rain visualisation remains the same as that of the classic word cloud, i.e., to create a static visualisation that quickly provides an understanding of text content by displaying the most prominent words in a text. However, by creating the Word Rain visualisation technique, we address important limitations associated with classic word cloud, thereby


mr
balance
people
diet
day
per
chocolate
dr
home
young
member
new
wander
carbohydrate
work
first
branch
food
oz
ltd
london
child
go
camp
research
available
milk
fruit
drink
way
get
need
insulin
help
come
bread
health
much
calorie
great
say
like
sugar
must
see
miss
think
course
even
could
always
little
without

a. *1965*


balance
talk
day
insulin
help
food
people
godiet
calorie
doctor
work
needcentre
cho
child
new
blood
per
hospital
medical
say
available
sugar
problem
get
supply
like
control
test
injection
much
first
cho_calorie
feel
long
glucose
product
could
come
think
high
carbohydrate
little
change
treatment
even
without

b. *1985*

◀ Figure 1: The top 600 most frequent words in patient periodicals from the British Diabetes Association.

making the visualisation more practically usable for analysing, exploring and comparing texts.

3.1 *The word positioning of the Word Rain technique*

The main method for achieving the additional analytical capability of the Word Rain visualisation is the way in which the words are positioned.

Particularly important is the feature of letting the *horizontal position* of a word be determined by the *meaning of the word*. Before the algorithm for positioning the words can be applied, each word to be visualised in a series of word rain graphs is assigned a fixed value on the horizontal axis (i.e., a fixed value on the x-axis), which represents its meaning. In the currently available implementation, this is achieved by projecting word embeddings—more specifically, multidimensional word-vectors from a word2vec-model—onto one dimension. Thereby, a word is assigned the point on the x-axis that is determined by projecting its word vector from a multi-dimensional space onto the one-dimensional x-axis.

This horizontal word positioning has the consequence that words with a similar meaning are positioned close to each other on the x-axis. Note that the x-axis does not represent a scale of meaning—it would be impossible to represent the general meaning of words on a scale. Instead, the word positioning could rather be viewed as depicting loosely defined clusters of meaning, and words with similar meaning—i.e., those belonging to the same semantic cluster—are positioned close to each other on the x-axis. For instance, in the upper graph in Figure 1, to the right (in violet), there are a number of words related to measurements, e.g., *oz*, *per*, *gramme*, and a bit to the left (in blue) there are many words related to *carbohydrate* and *sugar*.

When several word rains are generated as a series of graphs, as is the case for the two graphs in Figure 1, a word receives the same horizontal position in all word rains generated. That is, one and the same projection of word embeddings on the x-axis is used for every graph in a series. The words *insulin*, *calorie* and *balance*, which are prominent in both graphs, provide evident examples of this consistent word positioning.

While the horizontal position for a word is the same for all graphs included in a visualisation series, the vertical position (i.e., position on the y-axis) is not. The *vertical position* indicates *word prominence* in the text visualised, and it also adapts to the position of the other words in the graph.

The Word Rain algorithm positions the words at their pre-determined x-coordinate in a decreasing order of prominence. The algorithm starts by trying to position a word at the very top of the graph. If the extension of the word overlaps with that of a more prominent one, the word has to yield, and "rains down" in the graph until there is a free spot. In general, more prominent words are therefore positioned above less prominent ones. "Word prominence" can have different meanings (as will be exemplified by the use cases), but in Figure 1, it is equivalent to word frequency in the text visualised.

In the first graph in Figure 1, *per* is the most prominent word, and is therefore positioned first. The horizontal extension of the second most prominent word *carbohydrate* overlaps with *per*, and *carbohydrate* is therefore moved downwards until there is no longer an overlap. In contrast, the word *protein* does not overlap with a more prominent one, and is therefore positioned at the very top of the graph. A more compact graph can thereby be created by allowing the vertical position to not strictly follow word prominence.

The word positioning algorithm often results in vertical clusters of semantically similar words being created, where the more prominent ones generally (but not always) are positioned above the less prominent ones. The assigned x-coordinate of a word is the position where the word starts, i.e., in the example here, the point just to the left of the word.[5] In the first graph in Figure 1, a vertical cluster is, for instance, formed around the x-coordinate where the word *carbohydrate* starts. At almost the same x-coordinate, other words related to nutrition also start, i.e., the words *calorie* (below), and *fat* (above). When looking at a slightly larger portion of the x-axis, other related words can be found, such as *protein*, *eat*, and words related to the level of sugar (in blood/urine).

3.2 *The pinhead bars*

Similar to a classic word cloud, the font size used for the words correspond to the word's prominence. Above the area of the graph containing the words, there is a bar chart, consisting of a bar associated with each word in the graph. Their height also corresponds to prominence values of the words. These bars provide a prominence indication that is independent of the word lengths (thereby partly addressing the above-mentioned criticism of using font size as the sole prominence indicator). The pinhead at the top of each bar makes it easier to see where the bar ends.

5 For right-to-left languages, e.g., Yiddish, Arabic, Hebrew, this can be changed to the upper-right corner by a configuration option.

The bars also provide the additional functionality of making word clusters, as well as absences of words in certain semantic regions, more evident. For instance, when looking at the words in the right-hand side of the second graph in Figure 1, they might give the impression of populating the entire area. However, when looking at the bars, it is evident that there is a gap between the violet and the pink/reddish areas.

The graphs in Figure 1 show a colour configuration, where a gradient along the horizontal axis is formed by a wide spectrum of colours. It is, however, possible to configure the bars to be displayed using another colour scheme, for instance, to use a monochrome bar chart or to use fewer colours, as will be shown below. It is also possible to configure the word rain graphs to emphasise certain words, either by providing a pre-defined set of words to emphasise, or to configure the word rains to emphasise words occurring for the first time in a visualisation series (see use cases below). This is reflected in the bar chart by a grey colour. In the upper graph in Figure 1, there is for instance a cluster of emphasised words (e.g., *sunlight* and *exposure*) among the blue bars.

Each bar is connected to its corresponding word by a thin line. These thin vertical lines are drawn from the bottom of the bar to the starting position of the word with which the bar is connected. The length of these lines does not carry any meaning.

3.3 *Determining the horizontal word positions*

So far, we have used expressions such as "similar meaning" and "semantically close" rather carelessly. What we formally mean is the concept which is referred to as *paradigmatically related words* within distributional semantics, i.e., words that typically can be exchanged for each other in a sentence (Cuba Gyllensten 2023). Paradigmatically related words thus often appear in the same semantic contexts in a text. The words *fruit* and *bread* in the cluster of words to the very right (in red) in Figure 1 forms one example of paradigmatically related words. Both of them would, e.g., fit in a sentence such as "I eat a lot of (bread/fruit)".

To use a word2vec model is one, among many, methods to encode which words are paradigmatically related, and this is the method supported in the current implementation of the Word Rain algorithm. There are also many different methods available for reducing a multi-dimensional space to fewer dimensions. The default dimensionality reduction method supported in the current implementation of the Word Rain algorithm is t-SNE projection (van der Maaten & Hinton 2008).

When generating a word rain, it is either possible to use a pre-trained word2vec model, or to train a model on the same corpus that is visualised. The advantage of the latter option is that the word similarities in the word rain graphs then reflect paradigmatic similarities in the actual texts that are to be analysed. However, when the text corpus used for creating the word2vec model is too small, the quality of the model might not be high enough to be useful for the task of semantically sorting the words to visualise.

Also when the corpus used for training the word2vec model is relatively large, there might be words receiving a horizontal position in the word rain graphs that does not necessarily align with a human perception of paradigmatic similarity. This could, for instance, be the case for words that occur infrequently in the corpus used for training the model, or for words that occur in unexpected contexts in this corpus. For instance, the red food cluster in the first graph in Figure 1 contains the word *wander*, which does not seem to belong in this context.

It should also be noted that the effect of projecting the multi-dimensional information of a word2vec model onto just one dimension will result in information loss compared to using the original model. For the purposes of the Word Rain visualisation, it is enough to retain the information necessary for positioning similar words in close proximity, in order to let the user explore semantic neighbours or to identify clusters of similar words. However, in order to more closely study paradigmatic similarity distances—a task that is normally out of scope of the Word Rain visualisation—a method retaining more of the original multi-dimensional model is likely to be more suitable. Useful observations about paradigmatic similarity can, however, still be made based on the horizontal word positioning in a word rain, for instance observations that could form hypotheses to use for exploring semantic similarity in the original word2vec model. To summarise: the main aim of the Word Rain visualisation is *not* to convey detailed information on paradigmatic word similarity. Instead, the visualisation *employs* paradigmatic similarities in order to efficiently convey the most prominent words in a text.

4 *Use cases*

We present three use cases with an increasing level of implementation complexity. In the first one, the Word Rain web service is used and, thereby, no programming at all is needed. For the second use case, a minimum level of programming is required. The code for this use case does not contain any advanced programming logic, but is merely a way of setting different

parameters for the word rain. The third example, in contrast, shows how more complex configurations can be applied in order to tweak the word rain functionality. Also here, only a minimum amount of programming code is required, but the user needs to have some experience in applying programming logic. By offering—and here describing—different methods for generating word rains, we aim to target different kinds of users, both users who prefer an easy method for generating standard word rains without any need for coding, and users who want the flexibility to create word rains that fit specialised visualisation goals.[6]

4.1 *Corpora used for the experiments*

As example texts for describing the Word Rain technique, and for the three use cases, we used text corpora investigated in two projects connected to the Centre for Digital Humanities and Social Sciences Uppsala, CDHU, namely, the ActDisease and Swemper projects.

ActDisease[7] is a research project that studies the emergence, structure, and significance of diagnosis-specific patient organisations. The aim is to determine how certain diseases became subject to organisational efforts by people affected by them, and how this forming of organisations affected the understanding and management of those diseases. Around ten patient organisations from four European countries (Sweden, UK, France and Germany) are part of the study. As part of the project, their periodical publications (member newsletters and magazines) have been digitised, making up a corpus of approximately 125 000 pages from about 1890 to 1990 (Aangenendt et al. 2024). The project uses a combination of digital and traditional historiographic methods, with the digital text analysis aiming to provide insight into longitudinal changes in the texts, which helps guide the researchers in selecting case studies for a closer study of the texts and additional, archival sources. Furthermore, the project explores how to use digital text analysis methods to reveal layers of the text that are not readily visible to the human reader (Jänicke et al. 2017).

We here use two of the corpora digitised within the ActDisease project. For the use cases, periodicals issued by the patient organisation "The Swedish Diabetes Association" (Söderfeldt 2025) are used. This corpus consists of 8 891 pages, written in Swedish and published from 1949 to 1990 (except in 1950 and 1951). For the examples above, when explaining the Word Rain technique, we use periodicals from "The British Diabetes Association".

6 The code for generating the word rains showcased here is available at https://github.com/CDHUppsala/word-rain/tree/main/handbook (last accessed: February 14, 2025)

7 https://www.actdisease.org (last accessed: February 14, 2025)

The Swemper project[8] is a digital infrastructure currently under development at Uppsala University that aims to create a historical full-text database of Swedish medical periodicals produced during 1781–2011. By the end of the project around 550 000 pages (28 titles, ca. 1 100 physical volumes) will have been scanned, digitised and collected in a database. One of the goals of the project is to make this database available through a web interface (The Swedish Medical History Portal), which is also under development by the project. Another project goal is to explore and disseminate how state-of-the-art machine learning techniques can be used to improve the quality of layout analysis and OCR, and thereby provide better metadata to facilitate search and browsing of both images and text-data in the large database. The machine learning techniques used are described in Chapter 8 of this handbook (Ortiz Pablo et al. 2025).

Also from the corpora digitised in Swemper, we used one of the sub-corpora: "Läkartidningen". This is a Swedish medical journal published by the Swedish Medical Association. We here focus on content from the 1960s that had been digitalised and was available when this study was carried out.

For the first use case, the texts from the Diabetes periodical and from Läkartidningen were used as as-is, i.e., as provided from the digitisation process described by Aangenendt et al. (2024) and Ortiz Pablo et al. (2025), respectively. For use case 2 and 3, we used texts from the Diabetes periodical which had been lemmatised using Efselab (Östling 2018).[9]

4.2 *Comparing two text genres using the Word Rain web service*

The task for the first use case consists of creating an overview of—and comparison of—two corpora from two medical genres, both from the 1960s. The first corpus consists of periodicals from the Swedish Diabetes Association, and the second consists of texts from the journal Läkartidningen, as described above.

We created one text file consisting of pages from periodicals from the Swedish Diabetes Association published during the 1960s, and another text file consisting of the pages from Läkartidningen from the 1960s that so far have been digitised within the Swemper project. From the periodicals, we used all texts from the 1960s (2 239 pages), and from Läkartidningen we extracted pages containing either the substring *diabet* or the substring *sockersjuk* (1739 pages).

8 Communicating Science: Swedish Medical Periodicals, 1781–2011 (SweMPer) is a collaboration between Uppsala University Library, The Centre for Digital Humanities and Social Sciences (CDHU) and The Department of History of Science and Ideas at Uppsala University.

9 https://github.com/skogsgren/pefselab (last accessed: February 14, 2025)

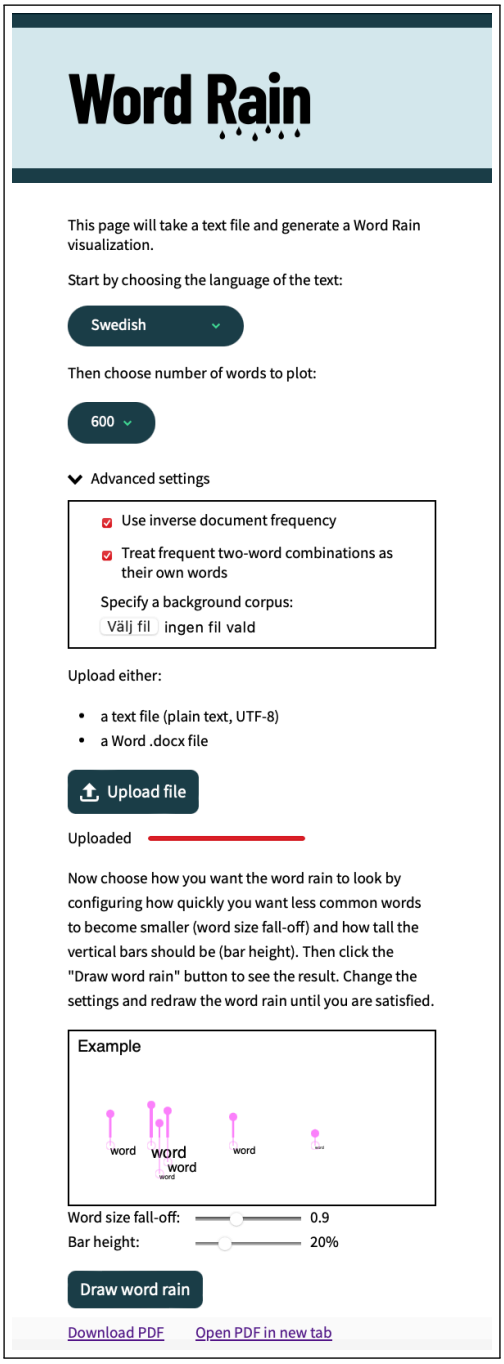


Figure 2: The Word Rain Web Service, i.e., a web page where texts can be uploaded to produce word rains.

The two text files were then uploaded to the Word Rain web service[10] (Ahltorp & Skeppstedt 2025), with parameters as shown in Figure 2. Parameters available include a slider for determining the height of the bars and how fast the font size should decrease for less prominent words. In the dialogue box that appears when "Upload file" is pressed, it is possible to select several files.

The web service produced the word rains shown in Figure 3, where the first graph shows the content from the patient periodical Diabetes, and the second graph shows the content from Läkartidningen.

When Swedish is chosen as the language of the corpora, words included in the standard Swedish stop word list from NLTK (Bird 2002) are used to filter which words are included in the visualisation. A pre-trained word2vec model, trained on Swedish text, is used for positioning the words. We configured the Word Rain web service to use TF-IDF as its prominence measure, i.e., a prominence measure that slightly exaggerates the differences between the corpora compared.

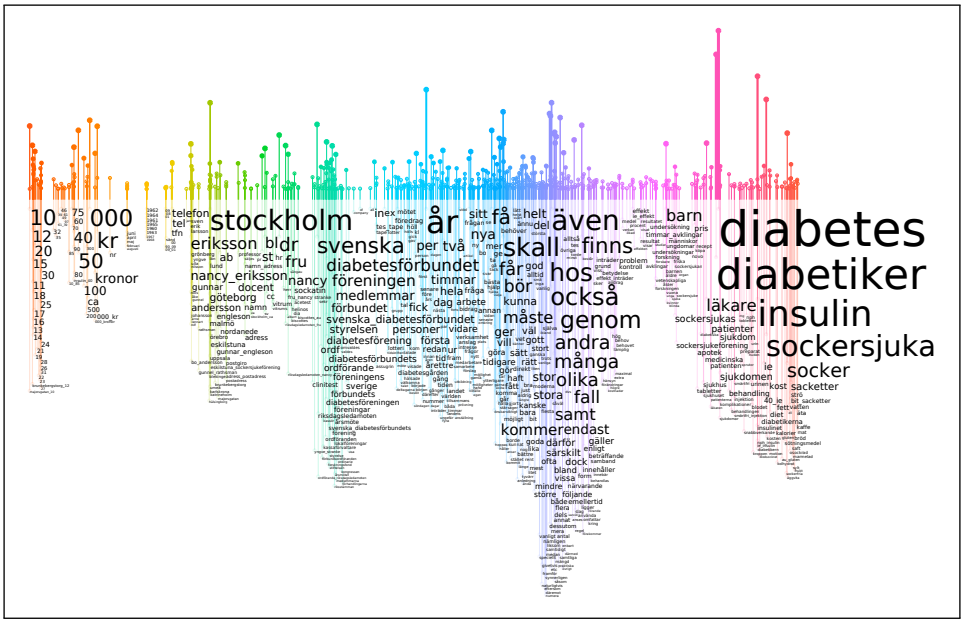
4.2.1 *Reading the word rains*

Despite using TF-IDF, it is still possible to observe many similarities between the two graphs in Figure 3. To the left, both graphs display different kinds of words and bigrams with numbers in orange/yellow, with more emphasis on years for Läkartidningen. To the right, in red/cerise, there are words closely related to diabetes for both texts, e.g., *diabetes* and *insulin*.

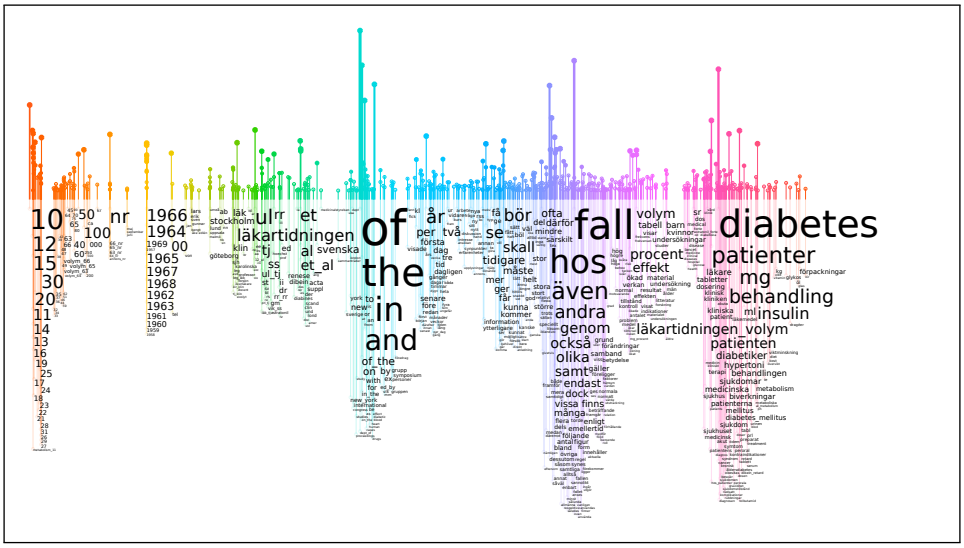
There are also differences between the two corpora; the most obvious being the vertical cluster of English words shown in turquoise in the Läkartidningen graph. While this cluster is very prominent in the Läkartidningen graph, the corresponding horizontal position is almost empty in the visualisation for the Diabetes periodical. That is, there are frequent English words in the Läkartidningen corpus, but not in the Diabetes corpus. Conversely, just to the left of the English words cluster, there is instead an empty segment in the Läkartidningen graph. At the same horizontal position in the Diabetes periodical graph, words related to the Swedish Diabetes Association, as well as words related to organisational matters in general, are located.

Another difference between the Diabetes periodical (directed towards patients) and the Läkartidningen journal (directed towards the medical profession) is the representation of places and people among the most common words. In the left part of both graphs in Figure 3, there is a vertical cluster of Swedish cities, more specifically at the horizontal position where the yellow colour turns into green. In Läkartidningen, only the three largest

10 https://wordrain.org (last accessed: October 7, 2025)



a. *Swedish Diabetes Association*



b. *Läkartidningen*

Figure 3: Comparing periodicals from the Swedish Diabetes Association (a), with content from Läkartidningen, a journal published by the Swedish Medical Association (b), for the first use case. Both corpora stem from texts on diabetes published during the 1960s.

cities in Sweden—as well as two university cities—have made it into the top 600 words on the pages discussing diabetes. In the Diabetes periodical, in contrast, there is a wider spread with some smaller cities included, as well as other types of places: two street names and a mention of the organisation's retreat *Nordanede*.

Names, and other words related to persons, are positioned directly to the left, as well as directly to the right, of the place cluster. In Läkartidningen, there is no full name with a frequency ranking high enough to be included among the top 600 most frequent words/bigrams. There is only one instance of a family name, and there are five instances of given names, all male. In Diabetes, in contrast, there are five different full names, mentioned frequently enough in the 1960s to be included among the most frequent words. Both female and male first names appear among the most prominent words. Frequently used titles in Diabetes include Mrs., Mr., doctor and professor, as well as titles in the form of functions within the patient organisation, such as president. In Läkartidningen, the titles are connected to the medical profession or academia, e.g., (associate) professor, licensed physician, chief physician and doctor.

These contrasts between the two word rains reflect the differences in content and audience of the corpora compared. However, they might not represent general differences between the two publications, since we used the entire content of the Diabetes periodical, while selecting a subset of Läkartidningen containing words related to diabetes. Nevertheless, the comparison of the two corpora form an illustrative example of how word rains can be read to detect differences between text collections.

The pink word cluster in the right-hand side of the graph contains words that could be described as "medical words". Here, there are differences related to vocabulary use. The word *diabetes* is frequent in both texts. The word *sockersjuka* 'sugar sickness', on the other hand, is very frequently used in the Diabetes periodical, but does not appear among the top 600 most frequent words in Läkartidningen. Conversely, in Läkartidningen the full medical term *diabetes mellitus* is used, but does not appear among the frequent words in the Diabetes periodical. Another difference regards what persons occur in the medical cluster. In both texts, the only medical profession frequently referred to in the 1960s (apart from references through titles) is *läkare* 'physician'. In the Diabetes periodical, the most important person in the medical word cluster is *diabetiker* 'diabetic', whereas *patienter* 'patients' is the most important person in Läkartidningen. Possibly, both words might often refer to a person with diabetes, but the vocabulary differs depending on the authors/target readers of the texts.

4.2.2 *Advantages/disadvantages of using the web service*

As previously mentioned, the large advantage of the Word Rain web service is that no programming is required to produce the word rains. The web service is therefore suitable both for those who do not have any need for the more detailed configuration options available when programming, as well as for those who first want to quickly try the visualisation technique on their corpus, to decide whether it would be useful to start using the Word Rain Python module.

A disadvantage of the web service, in addition to it offering fewer configuration options, is that not all corpora are possible to upload to an external organisation, e.g., corpora containing sensitive information. The code for the web service is, however, released as open source code.[11] It is therefore possible to set up the web service at any organisation where it is going to be used, either as a web service that can be accessed externally or as a local service within an organisation/at one specific computer.

4.3 *Visualising longitudinal trends*

The second use case focuses on only one of the corpora, the periodicals from the Swedish Diabetes Association. The task for this use case still includes comparison, but the aim of the word rains created is to visualise longitudinal trends. That is, instead of comparing two different text genres, as for the first use case, the aim is to make it possible to compare the content of each year to the content of the other years in the corpus. We therefore created a series of visualisations, with one word rain graph per year for the years 1949 and 1952–1990.[12]

The Word Rain Python module provides a number of configuration options for deciding which words to include in the graphs, for instance different types of word frequency cut-off configurations and different prominence measures. We have previously investigated and described the effects of varying the cut-off parameters with the aim of highlighting different aspects in the visualisation of longitudinal trends (Skeppstedt & Aangenendt 2024), and this will therefore not be the focus here. Instead, in use case two and three, we showcase two different *prominence measures* for visualising longitudinal trends.

As prominence measure in this second use case we employ raw word frequency. We also apply two different word frequency cut-offs: the word must occur at least ten times in the corpus and at least four times for the year

11 https://github.com/sprakradet/wordrain-service (last accessed: February 14, 2025)
12 There were no issues of the Diabetes periodical from the years 1950 and 1951.

visualised. For each year in the visualisation series, a graph is generated, which includes the top 600 most frequent words that fulfil the cut-off requirements. With these settings, we aim to showcase a longitudinal visualisation series with a configuration that is straight-forward to interpret.

Instead of using a pre-trained word2vec model, as in the first use case, a model was trained[13] on the same corpus that was visualised. Thereby, semantic similarity as expressed on closeness on the x-axis reflects semantic similarity in the corpus visualised, in contrast with general semantic similarity in the language, which was visualised in the first use case.

For the word rains in Figure 6, we have configured the graphs to include dashed vertical lines (and numbers associated with the lines). These lines have no other function than to offer a possibility to refer to different parts of the graphs, without having to rely on colour coding. We also used another colour gradient than in the previous example. The numbers in parentheses show the word frequency per 1000 words in the corpus.

We also configured the visualisation series to mark new words, i.e. words that have not appeared in the graphs for any of the previous years. These words are marked with a #-character and, when using the default font configuration, marked by an italic font. In addition, the bars corresponding to these new words are displayed in a grey colour.

4.3.1 *Stop words*

There are different ways to avoid very frequent words taking up a lot of space in the graphs, for instance, excluding words that occur all of the years in the series, or including general-language background corpora and using TF-IDF as the prominence measure. For this aim, it is also possible to use word rain configuration options for font size and for "word size fall-off", i.e., for how quickly the font size of less prominent words is to be decreased.

Another possibility, which is the method used here, is to employ a manually created stop word list, i.e., a list of words to exclude from the visualisations. We first employed the standard Swedish stop word list included in NLTK (Bird 2002), and then manually extended this list with words that took up a lot of space in the graphs. Most of these words were either semantically bleached in nature—e.g., *kunna* 'can' and *komma* 'come'—or were semantically rich words with a frequency that did not vary much over time—e.g., *insulin* and *diabetes*.[14]

13 For training the model, a window size of three and a vector size of 50 were used.

14 The use of *diabetes* actually *did* vary over time, as *sockersjuk* 'sugar-sick' was a more common term during the first six years visualised. However, if we treat these two terms as alternative ways to refer to the same concept, references to that concept were relatively stable over time.

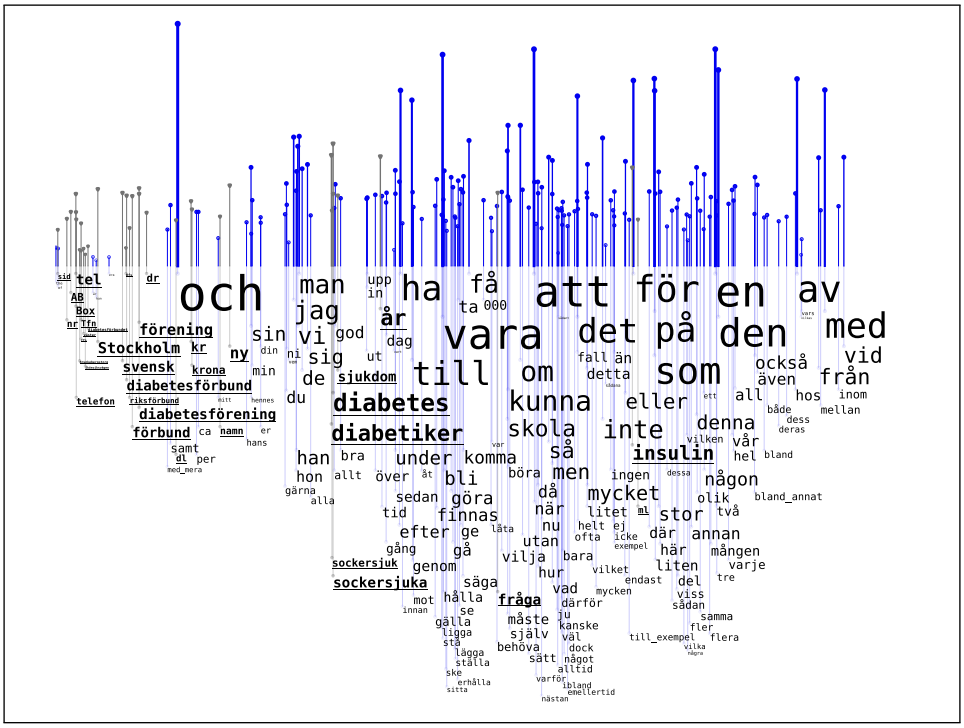
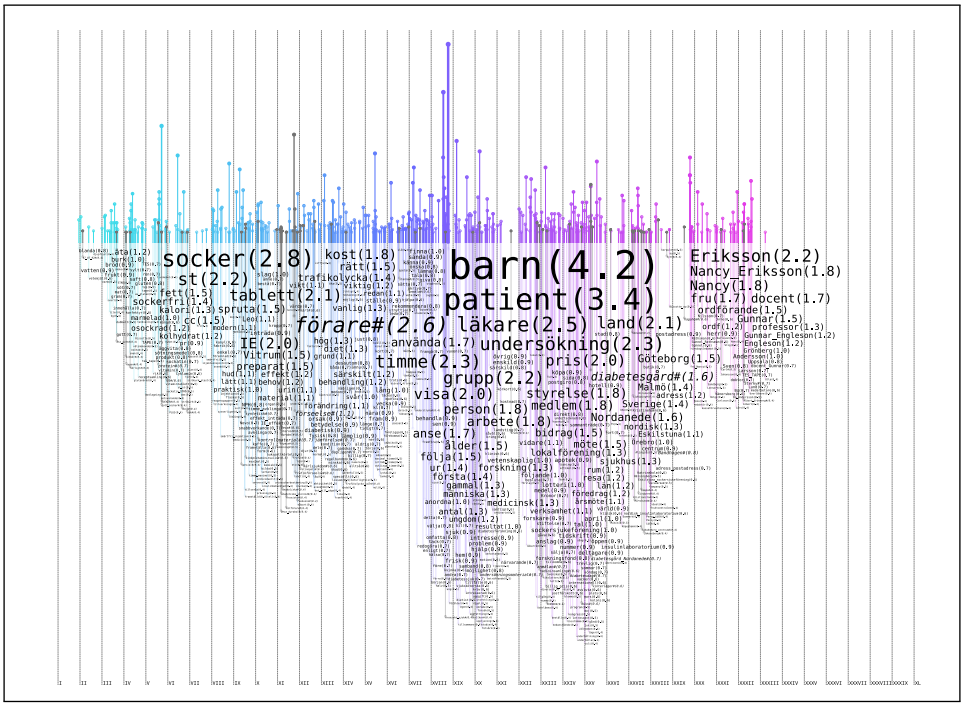


Figure 4: The stop words used in the second use case.

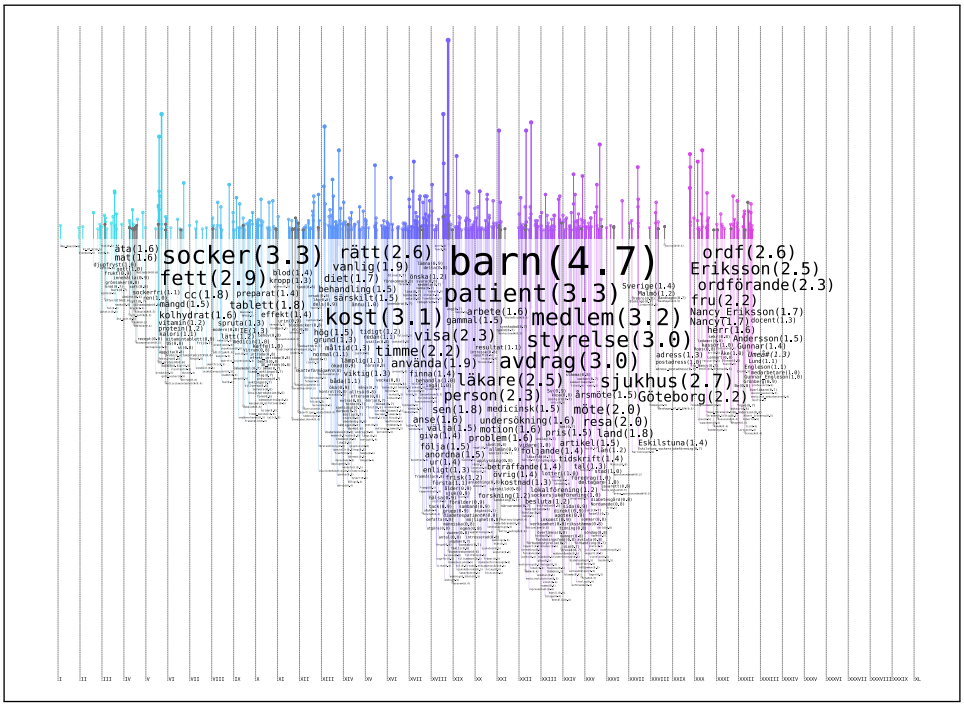
The stop words used are visualised in Figure 4, with the more semantically rich words underlined. Note that the visualisation of the stop words in Figure 4 is created as a completely separate graph from the series of graphs shown in Figures 5 and 6, with another projection of the words on the x-axis.

In addition to visualising the stop words, the graph in Figure 4 also showcases a configuration possibility of the Word Rain code, where the user can define a set of pre-defined words that are the only ones to be included in the graph. That is, instead of specifying a set of words to *exclude*, as for the stop words used for Figures 5 and 6, we here configure the word rain to *only include* a set of pre-defined words. The possibility to specify sets of words to exclude and include, thus, forms an option for easily tailoring the content of the word rains to specific visualisation needs. We have, for instance, previously experimented with compiling a list of named entities found in a text set, and both produced word rains that exclude the named entities (using the stop word functionality for excluding words), and word rains that only include the named entities (using the inclusion list).

In addition to showcasing the possibility to specify a pre-defined set


barn(4.2)
patient(3.4)
socker(2.8)
st(2.2)
fett(1.5)
tablett(2.1)
förare#(2.6)
läkare(2.5)
land(2.1)
undersökning(2.3)
IE(2.0)
timme(2.3)
grupp(2.2)
visa(2.0)
styrelse(1.8)
person(1.8)
arbete(1.8)
Eriksson(2.2)
Nancy Eriksson(1.8)
Nancy(1.8)
fru(1.7)
docent(1.7)
ordförande(1.5)
Gunnar(1.5)
professor(1.3)
Malmö(1.4)
Nordanede(1.6)
Sverige(1.4)
diabetesgård#(1.6)
pris(2.0)
Göteborg(1.5)
sjukhus(1.3)
bidrag(1.5)
möte(1.5)
forskning(1.3)
medicinsk(1.3)
medlem(1.8)

a. *1965*


barn(4.7)
patient(3.3)
socker(3.3)
fett(2.9)
kost(3.1)
medlem(3.2)
styrelse(3.0)
avdrag(3.0)
läkare(2.5)
sjukhus(2.7)
person(2.3)
Göteborg(2.2)
rätt(2.6)
visa(2.3)
timme(2.2)
arbete(1.6)
ordf(2.6)
Eriksson(2.5)
ordförande(2.3)
fru(2.2)
Nancy Eriksson(1.7)
NancyTi(1.7)
docent(1.3)
herr(1.6)
Andersson(1.5)
Gunnar(1.4)
Umeå(1.3)
Lund(1.1)
professor(1.1)
Nordanede(1.1)
Englesön(1.0)
Malmö(1.2)
Sverige(1.4)
Eskilstuna(1.4)
artikel(1.5)
tidskrift(1.4)
resa(2.0)
möte(2.0)
årsmöte(1.5)
land(1.8)
problem(1.6)
motion(1.6)
undersökning(1.6)
medicinsk(1.5)
sen(1.8)
person(2.3)
förfrågande(1.4)
betrfrande(1.4)
Övrig(1.4)
kostnad(1.3)
ålder(1.5)
förening(1.3)
lokalförening(1.3)
bidrag(1.5)
Nordanede(1.6)
nordisk(1.3)
Eskilstuna(1.2)
sjukhus(1.3)
rum(1.2)
resa(1.2)
förnedrag(1.2)
årsmöte(1.1)
lan(1.2)
verksamhet(1.0)
forskare(0.9)
sjukhus(0.9)
laboratorium(0.9)
ordförande(1.5)
Gunnar(1.5)
professor(1.3)
Englesön(1.2)
Gunnar_Englesön(1.2)
Eriksson(2.2)
Nancy_Eriksson(1.8)
Nancy(1.8)
fru(1.7)
docent(1.7)
ordförande(1.5)
Gunnar(1.5)
ordf(1.2)
professor(1.3)
Gunnar_Englesön(1.2)
Englesön(1.2)
Malmö(1.4)
Sverige(1.4)
Nordanede(1.6)
nordisk(1.3)
Eskilstuna(1.2)
sjukhus(1.3)
förnedrag(1.2)
lan(1.2)
verksamhet(1.0)
forskare(0.9)
sjukhus(0.9)
laboratorium(0.9)

b. *1966*

◀ Figure 5: Comparing two adjacent years, 1965 and 1966, for the second use case.

of words to include, Figure 4 also showcases i) the possibility to provide a user-defined list of words to underline, and ii) the possibility to create a monochrome word rain.

There is a trade-off between avoiding a very frequent, consistently occurring, word to take up a lot of space in the graph, and removing potentially interesting content from the visualisation. This is particularly the case when also removing semantically rich words, as for the example case presented here. Therefore, it might be relevant—as was done here—to also produce a visualisation of the content that was excluded.

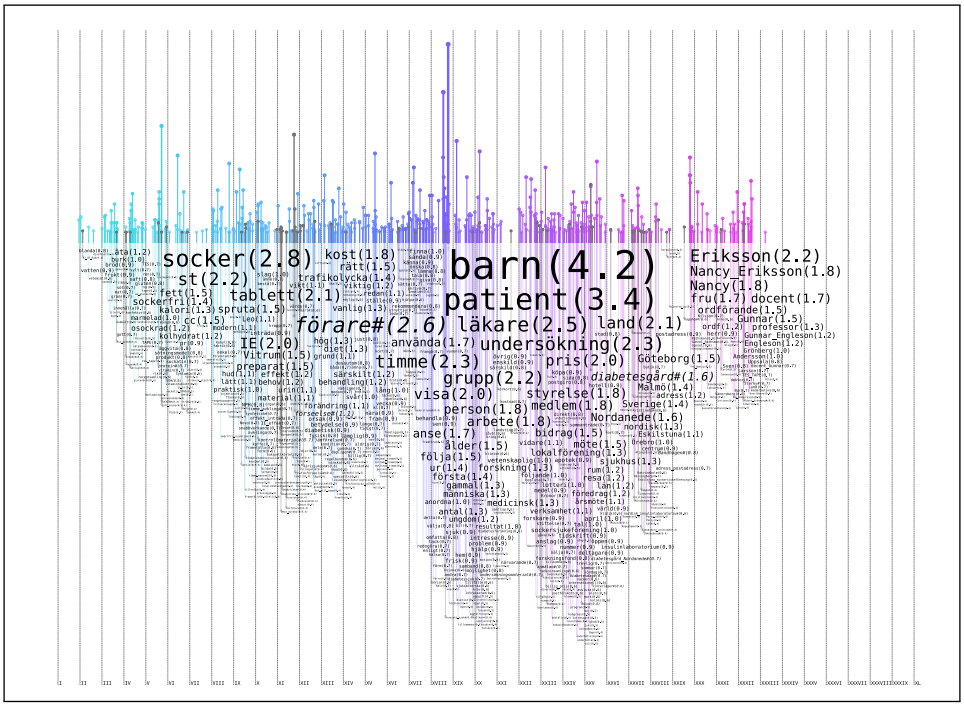
4.3.2 *Reading the word rains*

As examples from the visualisation series (which in total includes one graph per year, for the years 1949, 1952–1990), we compared graphs from the two adjacent years 1965 and 1966 (Figure 5), and graphs from two years far apart, 1965 and 1985 (Figure 6).

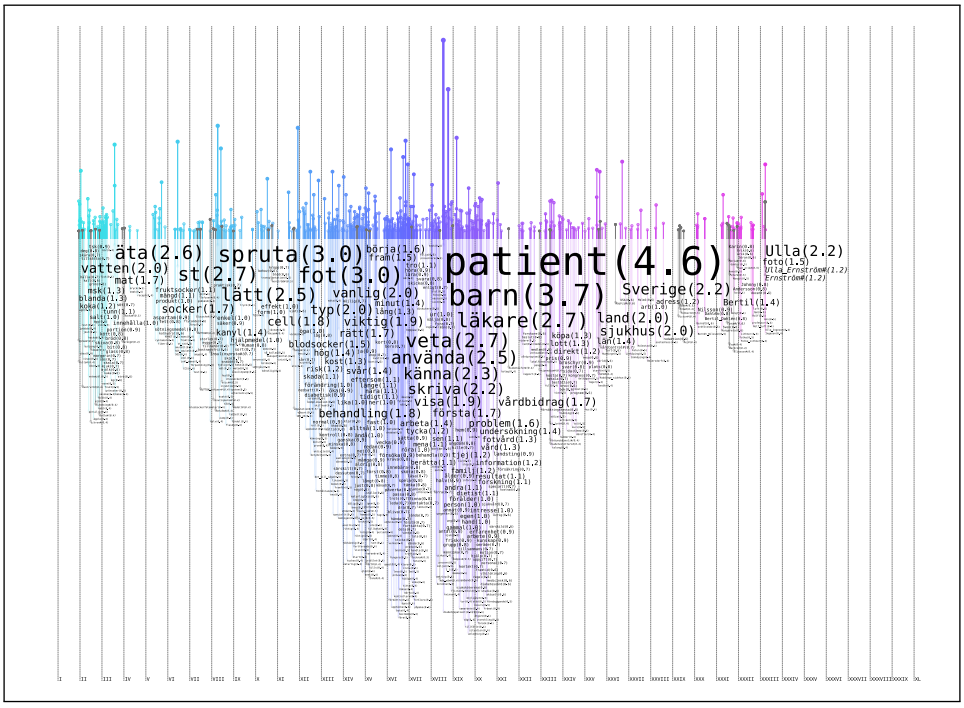
Table 1 contains a summary of the semantic word categories we were able to detect in the graphs in Figures 5 and 6. The roman numbers indicate an approximate position for the words in the graph. The categories within parentheses are only marginally represented in the graphs for the years 1965, 1966 or 1985, typically with only one or two words. However, these categories are prominent other years in the visualisation series—i.e., in the series created for the years 1949–1990—and we therefore included these categories in the table. The year in parentheses is the first year the semantic category appeared in the visualisation series.

The differences between the two adjacent years, 1965 and 1966, in Figure 5 are less obvious than between the two different corpora compared above. Instead, what stands out is rather the similarity between year 1965 and year 1966. There are, however, still examples of immediately evident differences. For instance, words from 1965 related to traffic and driving, e.g., *förare* 'driver' and *trafikolycka* 'traffic accident', are no longer present among the top 600 most frequent words in 1966. Instead, some words related to food have become more prominent in 1966, e.g., *kost* 'food/diet' and *fett* 'fat', as well as words related to organisation administration, e.g., *medlem* 'member' and *styrelse* 'board' and economics, e.g., *avdrag* 'deduction'.

With the grey colour marking, it is possible to spot semantic regions of new words in the graph. For instance, close to line XI in the graph for 1965,


barn(4.2)
patient(3.4)
socker(2.8)
förare#(2.6)
läkare(2.5)
undersökning(2.3)
timme(2.3)
grupp(2.2)
Eriksson(2.2)
Nancy_Eriksson(1.8)
Nancy(1.8)
fru(1.7)
docent(1.7)
ordförande(1.5)
Gunnar(1.5)
professor(1.3)
Gunnar_Engleson(1.2)
Engleson(1.2)
land(2.1)
använda(1.7)
IE(2.0)
Vitrum(1.5)
preparat(1.5)
kost(1.8)
rätt(1.5)
trafikolycka(1.4)
tablett(2.1)
spruta(1.5)
kalori(1.5)
kolhydrat(1.2)
effekt(1.2)
pris(2.0)
Göteborg(1.5)
diabetesgård#(1.6)
Malmö(1.4)
styrelse(1.8)
medlem(1.8)
Sverige(1.4)
visa(2.0)
person(1.8)
arbete(1.8)
Nordanede(1.6)
nordisk(1.3)
bidrag(1.5)
ålder(1.5)
följa(1.5)
forskning(1.3)
möte(1.5)
lokalförening(1.3)
sjukhus(1.3)
rum(1.2)
resa(1.2)
förstå(1.4)
gammal(1.3)
människa(1.3)
medicinsk(1.3)
förnedrag(1.2)
län(1.2)
ung(1.2)
antal(1.3)
ung(1.2)

a. *1965*


patient(4.6)
barn(3.7)
läkare(2.7)
äta(2.6)
spruta(3.0)
fot(3.0)
lätt(2.5)
veta(2.7)
använda(2.5)
känna(2.3)
skriva(2.2)
visa(1.9)
vårdbidrag(1.7)
behandling(1.8)
förstå(1.7)
problem(1.6)
undersökning(1.4)
vård(1.3)
land(2.0)
sjukhus(2.0)
Sverige(2.2)
Ulla(2.2)
Foto(1.5)
Bertil(1.4)
socker(1.7)
cell(1.8)
typ(2.0)
viktig(1.9)
vanlig(2.0)
vatten(2.0)
mat(1.7)
st(2.7)
börja(1.6)
från(1.5)
rätt(1.7)
kost(1.3)
svår(1.4)
hög(1.4)
blodsocker(1.5)
risk(1.2)
skada(1.1)
arbeta(1.4)
tycka(1.2)
tid(1.2)
direkt(1.2)
kopa(1.3)
kott(1.3)
lan(1.4)

b. *1985*

◀ Figure 6: The most frequent words/bigrams (after stop word removal) in periodicals from the Swedish Diabetes Association, from 1965 and 1985.

Table 1: Semantic categories in Figures 5 and 6.

Semantic category	Label	Appears
Food/recipes	II–V	1953
(Self tests for sugar)	V–VII	1954
Nutrition	VI	1953
Artificial sweeteners and types of sugar	VI	1954
Insulin and other medication	VIII–X	1949
Blood and urine	XI	1953
(Processes, body parts/substances involved in the cause of diabetes)	X–XI	1953
Medical complications, risks and related body parts	XII–XIII	1953
Traffic and driving	XIII	1965
Diet/weight/exercise/treatment	XIII–XIV	1949
Words for discussions and feelings	XVII–XVIII	1955
Children/youths, other people, professions	XIX	1949
Work (employment)	XX	1949
Research, advice, development, education	XX–XXI	1949
Economical matters	XXI	1949
Organisation words, e.g., member/board	XXII–XXIII	1949
Lottery, fundraising	XXIII	1957
Research or research funding	XXIII	1953
Organisational activities	XXIV–XXV	1954
Published/printed material	XXIV	1949
Summer camp/Nordanede/trips	XXV–XXVI	1949
Places where healthcare is taking place	XXVI	1954
Places outside Sweden	XXVI–XXVII	1954
Swedish cities and addresses	XXVII–XXIV	1949
Organisation functions, titles and names	XXX–XXXI	1949
Academic titles, names and university cities	XXXII–XXXIII	1949

there are a number of words not previously included in the series of graphs.

In Figure 6, which instead compares two years far apart, 1965 and 1985, the differences are larger. Many (but not all) semantic categories are present in both years, but they often differ in content. For instance, the insulin/medication category in 1985 reflects the spread of new technology around this time (Bradwel 2023: pp. 80–82, 85–87), with new words, such as *insulinpump* 'insulin pump' and *insulinpenna* 'insulin pen'. There is also more focus on the

Figure 7: Comparing two adjacent years, 1965 and 1966, with the frequency difference compared to previous year as the prominence measure for the third use case. ▶

body affected by diabetes, with *fot* 'foot' as one of the most prominent words in 1985, as well as *öga* 'eye' and *sår* 'wound' being prominent words. The food category is also more prominent, and it includes more verbs frequently used in recipes, such as *koka* 'boil' and *blanda* 'mix'.

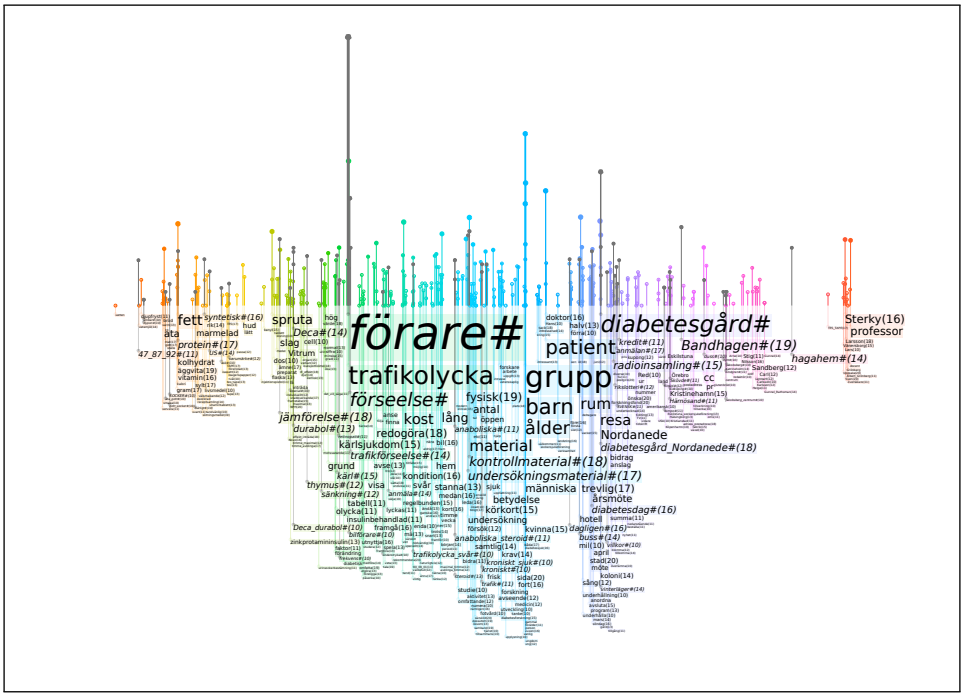
The graphs also show a change in common organisational activities, with many different kinds of activities in 1965—e.g., lectures, meetings, diabetes day, singing—while only congresses, national meetings and courses remain in 1985. Activities related to the diabetes retreat Nordanede, which was acquired in 1963 and closed in 1983 (Swedish Labour Movement's Archives and Library, Personal archives of Nancy Eriksson n.d.), are also not any more present in the graph for 1985. Organisation-related words in general receive much less attention in 1985, e.g., with words such as *styrelse* 'board' and *medlem* 'member' being very prominent in 1965, but not at all in 1985.

4.4 *Longitudinal trends, but with a focus on novelty*

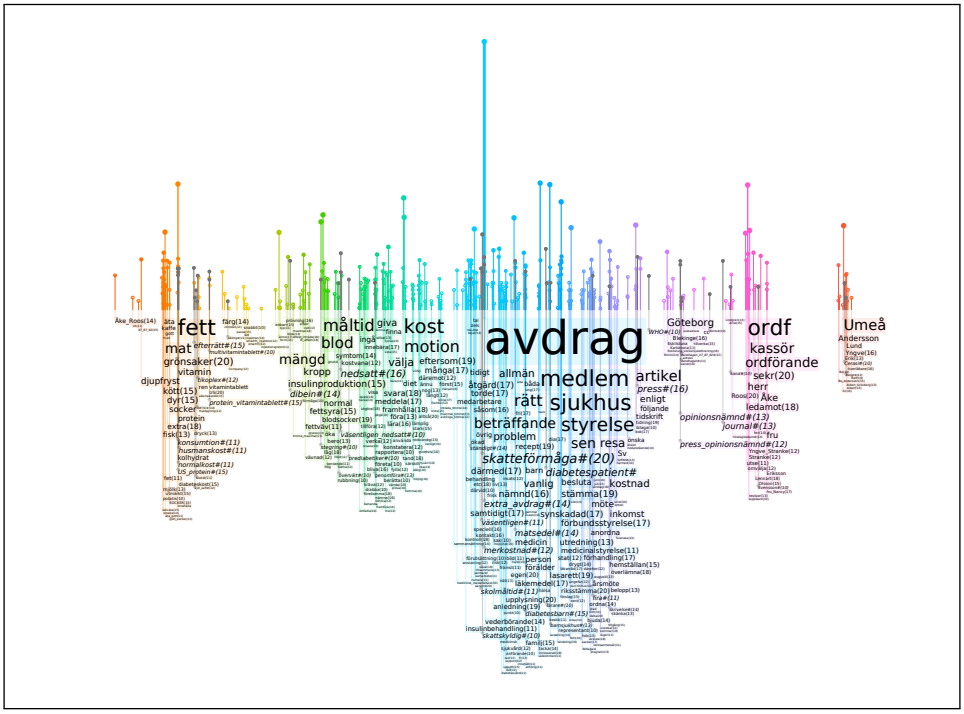
For the third use case, we also show visualisation examples from the two adjacent years, 1965 and 1966. Here, we do, however, not apply raw word frequency as the prominence measure for selecting which words to show. Instead, we use the word frequency *difference* between the year that is visualised and the previous year. That is, the graph for 1965 in Figure 7 shows the words for which the frequency has increased the most compared to 1964. Similarly, for the graph for 1966, the prominence measure applied is the increase in frequency from 1965 to 1966. Practically, we use the differences in relative frequency.

In contrast with the previous use case, there is no ready-made word rain configuration setting for using differences in relative frequency as the prominence measure. Instead, the Word Rain code offers a way for the user to submit their own Python function for defining word prominence. This functionality provides the user with a large degree of flexibility in defining the prominence measure, but, as already mentioned, also requires more programming skills.

The point of the prominence measure for the third use case is to make it possible to detect words that show a large increase in usage for the year visualised. This prominence measure is not only more difficult to technically implement, but also slightly more difficult to interpret than the standard fre-


förare#
trafikolycka
förseelse#
grupp
patient
diabetesgård#
Bandhagen#(19)
hagahem#(14)
Sterky(16)
professor
fett
spruta
kost
lång
ålder
barn
rum
resa
material
Nordanede
diabetesgård_Nordanede#(18)
kontrollmaterial#(18)
undersökningsmaterial#(17)
människa
trevlig(17)
årsmöte
diabetesdag#(16)
dagligen#(16)
anaboliska_steroid#(11)
trafikolycka_svår#(10)
sjuk#(10)
kronisk#(10)
kondition(16)
körkort(15)
undersökning
betydelse
kvinna(15)
jämförelse#(18)
durabol#(13)
trafikförseelse#(14)
grund
kärl#(15)
thymus#(12)
sänkning#(12)
Deca_durabol#(10)
olycka(11)
insulinbehandlad(11)
tabell(11)
visa
svår
stanna(13)
hem
redogöra(18)
kärlsjukdom(15)
Deca#(14)
protein#(17)
kolhydrat
fett
mat
grönsaker(20)
vitamin
djupfryst
kött(15)
dyr(15)
socker
protein
extra(18)
fisk(13)
konsumtion#(11)
husmanskost#(11)
kolhydrat
normalkost#(11)
protein#(15)
måltid
kost
motion
blod
mängd
kropp
normal
nedsatt#(16)
insulinproduktion(15)
dibein#(14)
protein_vitamintablett#(15)
avdrag
medlem
artikel
press#(16)
sjukhus
rätt
beträffande
styrelse
sen
resa
skatteförmåga#(20)
diabetespatient#
extra_avdrag#(14)
Göteborg
ordf
kassör
ordförande
sekr(20)
herr
Åke
ledamot(18)
journal#(13)
press_opinionsnämnd#(12)
Umeå
Andersson
Lund
Ystad(16)
Ängelholm(16)
Eskilstuna(16)

a. *1965*


avdrag
fett
måltid
kost
motion
blod
mängd
kropp
normal
nedsatt#(16)
insulinproduktion(15)
dibein#(14)
protein_vitamintablett#(15)
mat
grönsaker(20)
vitamin
djupfryst
kött(15)
dyr(15)
socker
protein
extra(18)
fisk(13)
konsumtion#(11)
husmanskost#(11)
kolhydrat
normalkost#(11)
protein#(15)
medlem
artikel
press#(16)
sjukhus
rätt
beträffande
styrelse
sen
resa
skatteförmåga#(20)
diabetespatient#
extra_avdrag#(14)
problem
allmän
åtgärd(17)
förslag(17)
rätt
sjukhus
Göteborg
ordf
kassör
ordförande
sekr(20)
herr
Åke
ledamot(18)
journal#(13)
press_opinionsnämnd#(12)
Umeå
Andersson
Lund
Ystad(16)
Ängelholm(16)
Eskilstuna(16)

b. *1966*

quency measures. When analysing these results, it is important to remember that novelty/frequency increase is the sole focus of the visualisation. When frequency for a word instead has *decreased* compared to the previous year, the word is therefore not included in the graph for that year, despite that this word might be a frequently occurring one also that year. When analysing visualisations using differences in relative frequency as the prominence measure, it is thereby useful to also look at word rain visualisations that use the raw word frequency measure. For instance, in Figure 7, the prominence measure focusing on novelty makes it easier to detect the large frequency increase for the two words *diabetesgård* 'diabetes retreat' and *Nordanede* 'name of a retreat place' (both shown in violet) in 1965, than to detect frequency changes by comparing two graphs that visualise raw word frequency. However, to study the general development of these two words, e.g., how much they then decreased in 1966, standard word frequency is a better prominence measure.

Apart from using a different prominence measure, we altered a few additional configuration settings compared to the previous use case, in order to showcase other possible word rain settings. For instance, the top 400 most prominent words are shown (instead of the top 600), the minimum word frequency in the text visualised is set to ten, the labelled lines are not included, and the number shown in parentheses is the word's absolute frequency in the text visualised (and this is only shown for a frequency of 20 and less).

For this use case, we also showcase a configuration option that allows the user to somewhat influence the result of the projection on the x-axis. This configuration was also used for the visualisation of the British Diabetes Association texts in Figure 1. With this configuration option, the user defines a k number of clusters, to use for clustering the words, and more distinct word clusters along the x-axis are thereby created. When comparing the graphs in Figures 5 and 6 to the graphs in Figures 1 and 7, it is possible to see that the words in the latter figures are positioned in more distinct clusters along the x-axis. In Figure 7, we configured the projection to use 50 clusters, and in Figure 1 to use 30 clusters.

The same wide colour scale as in Figure 1 was used for the colour gradient. For both Figure 1 and Figure 7, the colours and clusters were emphasised even further by using a configuration option which provides the words with a coloured background box. With the purpose of reminding about the possibility to vary the parameters for the word2vec model used, we here trained a new word2vec model on the corpus, slightly altering[15] the parameters used.

15 More specifically, increasing to a window size of five words and a vector size of 100.

4.4.1 *Reading the word rains*

In the graph for 1965, the previously mentioned prominent semantic categories are easier to detect when word frequency difference is used as prominence indicator. In green, under the word *förare* 'driver', there are a number of prominent words related to traffic and driving, mainly regarding traffic accidents. Similarly, there are many words related to the diabetes retreat Nordanede (between blue and violet), that have increased in 1965 compared to 1964. This is also the case for words related to nutrition (to the left in orange). There are also some individual words that stand out. For instance, among organisational activities (between blue and violet), there is the new word *diabetesdag* 'diabetes day', an annual campaign day with among other things a nationwide fundraising effort through the radio introduced in 1965 (Pehrson 1965). Consequently, among the lottery/fundraising words (in violet) *radioinsamling* 'fundraising through the radio' stands out.

In 1966, words related to nutrition continue to increase in frequency (to the left in orange), and organisational words also increase (in pink). Most prominent here, is the economic category (in blue, under the word *avdrag* 'deduction'). Also here, there are individual words that have increased, e.g., *kost* 'food/diet' and *motion* 'exercise'.

5 *Word Rain in relation to other text visualisation techniques*

We will end this chapter by positioning the Word Rain technique in the context of other text visualisation techniques. The content of this section is not required for understanding, generating or reading the word rains, but might be useful for comparing the technique to other possible choices for visualising or listing the most prominent words in a text or corpus (Felix et al. 2018).

We will start by comparing the Word Rain technique to the techniques it aims to replace, i.e., to word frequency lists and the classic word cloud. Thereafter, we will provide some other examples of word cloud extensions.

5.1 *Word Rain compared to a word frequency table and a word cloud*

By creating a word frequency list (Table 2) and a classic word cloud (Figure 8) for the same texts we illustrated above in Figure 6, we aim to emphasise the advantages of the Word Rain technique compared to the other two methods.

Similar to the word clouds, the word rains in Figure 6 also immediately show which are the most important words in the texts. However, when the Word Rain technique is used for visualising the texts, much more information

Table 2: A table showing the 30 most frequent words/bigrams (after stop word removal) in periodicals from the Swedish Diabetes Association, from the year 1965 and the year 1985. Frequency is shown by the number of occurrences per 1000 words (‰) in the periodicals from that year.

	Year 1965		Year 1985	
Rank	Word	‰	Word	‰
1	*barn*	4.2	*patient*	4.7
2	*patient*	3.4	*barn*	3.7
3	*socker*	2.8	*spruta*	3.0
4	*förare*	2.6	*fot*	3.0
5	*läkare*	2.5	*läkare*	2.8
6	*undersökning*	2.3	*veta*	2.7
7	*timme*	2.3	*st*	2.7
8	*st*	2.2	*äta*	2.6
9	*grupp*	2.2	*lätt*	2.5
10	*Eriksson*	2.2	*använda*	2.5
11	*land*	2.1	*känna*	2.3
12	*tablett*	2.1	*Sverige*	2.2
13	*visa*	2.0	*skriva*	2.2
14	*pris*	2.0	*Ulla*	2.2
15	*IE*	2.0	*land*	2.0
16	*kost*	1.8	*vatten*	2.0
17	*Nancy Eriksson*	1.8	*vanlig*	2.0
18	*Nancy*	1.8	*typ*	2.0
19	*styrelse*	1.8	*sjukhus*	2.0
20	*person*	1.8	*visa*	1.9
21	*arbete*	1.8	*viktig*	1.9
22	*medlem*	1.8	*cell*	1.8
23	*krona*	1.8	*behandling*	1.8
24	*docent*	1.7	*socker*	1.8
25	*fru*	1.7	*vårdbidrag*	1.7
26	*använda*	1.7	*första*	1.7
27	*anse*	1.7	*rätt*	1.7
28	*diabetesgård*	1.6	*mat*	1.7
29	*Nordanede*	1.6	*krona*	1.6
30	*ålder*	1.5	*problem*	1.6



a. *1965*



b. *1985*

Figure 8: Two word clouds showing the most frequent words/bigrams (after stop word removal) in periodicals from the Swedish Diabetes Association, from the year 1965 and the year 1985.

can be derived from the graphs. As exemplified in the use cases, the word rains make it possible to find semantic word categories, compare texts, and zoom in into interesting areas. We argue that it would have been very difficult to find prominent semantic word categories—i.e., to produce the content of Table 1—from the classic word clouds or from the word frequency lists (Table 2 and Figure 8). Similarly, the analyses of differences between

the graphs would also have been very difficult to produce based on the information provided by the classic word clouds or frequency lists.

5.2 *Previous improvements of the classic word cloud*

There are many other previously applied approaches, addressing some of the limitations of the word cloud. To address the absence of a meaningful word positioning, different types of semantically motivated word positioning algorithms for static word clouds have been proposed (Wu et al. 2011, Barth et al. 2014), including those based on word embeddings (Xu et al. 2016). The methods typically keep a layout similar to the classic word clouds, but position semantically related words close to each other, and emphasise semantic clusters/relations by colour coding the words.

Also the problem associated with the difficulties of comparing word clouds has been addressed through solutions visually very similar to a classic word cloud. For instance, word clouds more suitable for comparison can be achieved by assigning words a fixed position in a series of word cloud graphs, and only allowing sets of words that never occur in the same word cloud to occupy the same space (Herold et al. 2019). Another possible approach is to apply restraints on the word positioning to keep the contexts of the words similar in each graph produced in a series of word cloud graphs (Cui et al. 2010). A very different approach for addressing the difficulty of comparing word clouds is to produce a single word cloud for all the texts that are to be compared, and then use colours (Diakopoulos et al. 2015, Burch et al. 2014) or trend lines (Lee et al. 2010) to indicate in which texts the word occurs or how the word has varied over time.

There are also solutions to the problem of using font size as prominence indicator. For instance, Alexander et al. (2018) provided the idea of supplementing font size with a padded bounding box around the words, which has the effect of short words taking up as much space as longer ones.

The use of further typographic attributes—such as italic font or to underline text—has also been suggested, as a means beyond font size or colour to indicate word properties in word clouds (Brath & Banissi 2016).

There are thus many different approaches for addressing the limitations associated with the classic word cloud. As previously mentioned, however, these approaches typically focus on addressing only *one* limitation. With the Word Rain technique we, instead, aim to address *several* of the limitations that make the word cloud unsuitable as a tool for analysing text.

6 *Conclusion*

We have in this chapter described limitations of the classic word cloud, as well as how the Word Rain technique addresses these limitations. We have then provided three use case examples, with an increasing level of difficulty, which show how the word rains can be used for exploring, analysing and comparing the most prominent words in different texts. Finally, we have compared the Word Rain technique to using a word frequency list and a classic word cloud, as well as described previous approaches for improving the word cloud.

With this description of the Word Rain technique—and of how it can be practically applied—we hope to have convinced you to use a Word Rain visualisation, where you previously would have used a word frequency list or a word cloud. We also hope that our practical examples have shown you how a series of semantically structured word graphs very quickly can provide you with an overview understanding of the content of a text collection, as well as with an understanding of how this content varies over time. We would argue that after spending a short time analysing the visualisations generated here, even a person with no prior knowledge of the Swedish Diabetes Association would be able to possess an overview understanding of the content of its member publication, for the years discussed.

In many usage scenarios for the Word Rain technique, the word rains could be employed as a first step when approaching a new text collection to study. The technique provides a content overview of the texts, as well as lets the user explore and structure more specific, and potentially interesting, content. This overview understanding might be enough, but it might also be relevant to take the word rains as a point of departure, and dig deeper into the content elsewhere. For instance, to produce frequency trend lines for interesting words/word clusters found in the visualisations, or to use the word rain as a basis for selecting a subset of texts to read more closely. These further analyses are, however, out of scope for this handbook chapter. This is also the case for different types of text pre-processing techniques—e.g., named entity recognition or topic modelling—for further tailoring the visualisation to different user requirements. Also a discussion of how to best generate word rains to use as graphic design elements on printed material is left to future publications.

Through the three use cases with an increasing level of complexity, we have shown how the Word Rain technique can support both the user who wants to upload texts on a web page and create a standard word rain, as well as the user who wants to adapt the word rains to their specific visualisation requirements. The three use cases presented here are, however, just the

beginning. There are still many unexplored possibilities to further tweak the word rains, for instance, by using different text pre-processing techniques before generating the word rains, or by applying different word filtering options, different colour schemes and/or different prominence measures. Even larger are the possibilities in the form of different text collections on which it would be relevant to apply the Word Rain visualisation. We therefore look forward to following how the Word Rain web service—as well as the Word Rain Python module—will be applied by future users to generate novel types of visualisations for different kinds of text collections.

Acknowledgments

The work described in this handbook chapter has mainly been conducted within the ActDisease project, which is funded by the European Union (ERC ActDisease ERC-2021-STG 101040999). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

The development of the Word Rain Python module and the Word Rain web service is also funded by the Swedish Research Council: Huminfra (2021-00176, 2023-00171), InfraVis (2021-00181) and Swe-CLARIN/The National Language Bank of Sweden (2017-00626).

Finally, Communicating Science: Swedish Medical Periodicals, 1781–2011 (SweMPer) is funded by Riksbankens jubileumsfond.

References

Aangenendt, Gijs, Maria Skeppstedt & Ylva Söderfeldt. 2024. Curating a historical source corpus of 20th century patient organization periodicals. In *Proceedings of the Huminfra Conference (hic 2024)*, 76–82. DOI: 10.3384/ecp205011.

Ahltorp, Magnus, Jean Hessel, Gunnar Eriksson & Maria Skeppstedt. 2025. Visualisering av ett lexikons täckning av olika textgenrer: Experiment med en jiddischordbok [Visualising the coverage of dictionary for different text genres: Experiments with a Yiddish dictionary]. In *Nordiske studier i leksikografi*. Accepted for publication.

Ahltorp, Magnus & Maria Skeppstedt. 2025. Word Rain as a service: Making semantically structured word clouds available to everyone. In Vincent Vandeghinste & Thalassia Kontino (eds.), *Selected papers from the CLARIN annual conference 2024* (Linköping Electronic Conference Proceedings 216). DOI: https://doi.org/10.3384/ecp216.

Alexander, Eric, Chih-Ching Chang, Mariana Shimabukuro, Steven Franconeri, Christopher Collins & Michael Gleicher. 2018. Perceptual biases in font size as a data encoding. *IEEE Transactions on Visualization and Computer Graphics* 24(8). 2397–2410. DOI: 10.1109/TVCG.2017.2723397.

Barth, Lukas, Stephen G. Kobourov & Sergey Pupyrev. 2014. Experimental comparison of semantic word clouds. In *Experimental algorithms*, 247–258. Springer International Publishing. DOI: 10.1007/978-3-319-07959-2_21.

Bird, Steven. 2002. NLTK: The natural language toolkit. In *Proceedings of the ACL workshop on effective tools and methodologies for teaching natural language processing and computational linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics.

Bradwel, Stuart. 2023. *Insulin: A hundred-year history*. Cambridge: Polity.

Brath, Richard & Ebad Banissi. 2016. Using typography to expand the design space of data visualization. *She Ji: The Journal of Design, Economics, and Innovation* 2(1). 59–87. DOI: 10.1016/j.sheji.2016.05.003.

Burch, Michael, Steffen Lohmann, Fabian Beck, Nils Rodriguez, Lorenzo Di Silvestro & Daniel Weiskopf. 2014. RadCloud: Visualizing multiple texts with merged word clouds. In *Proceedings of the international conference on information visualisation* (IV '14), 108–113. DOI: 10.1109/IV.2014.72.

Cao, Nan & Weiwei Cui. 2016. *Introduction to text visualization*, vol. 1 (Atlantis Briefs in Artificial Intelligence). Atlantis Press. DOI: 10.2991/978-94-6239-186-4.

Collins, Christopher, Fernanda B. Viégas & Martin Wattenberg. 2009. Parallel tag clouds to explore and analyze faceted text corpora. In *Proceedings of the IEEE symposium on visual analytics science and technology*, 91–98. DOI: 10.1109/VAST.2009.5333443.

Cuba Gyllensten, Amaru. 2023. *Quantifying meaning*. KTH Royal Institute of Technology. (Doctoral dissertation).

Cui, Weiwei, Yingcai Wu, Shixia Liu, Furu Wei, Michelle Zhou & Huamin Qu. 2010. Context-preserving, dynamic word cloud visualization. *IEEE Computer Graphics and Applications* 30(6). 42–53. DOI: 10.1109/MCG.2010.102.

Diakopoulos, Nicholas, Dag Elgesem, Andrew Salway, Amy Zhang & Knut Hofland. 2015. Compare Clouds: Visualizing text corpora to compare media frames. In *Proceedings of the IUI workshop on visual text analytics*.

Felix, Cristian, Steven Franconeri & Enrico Bertini. 2018. Taking word clouds apart: An empirical investigation of the design space for keyword summaries. *IEEE Transactions on Visualization and Computer Graphics* 24(1). 657–666. DOI: 10.1109/TVCG.2017.2746018.

Hearst, Marti A., Emily Pedersen, Lekha Patil, Elsie Lee, Paul Laskowski & Steven Franconeri. 2020. An evaluation of semantically grouped word cloud designs. *IEEE Transactions on Visualization and Computer Graphics* 26(9). 2748–2761. DOI: 10.1109/TVCG.2019.2904683.

Herold, Elisa, Marcus Pöckelmann, Christian Berg, Jörg Ritter & Mark M. Hall. 2019. Stable word-clouds for visualising text-changes over time. In *Digital libraries for open knowledge*, 224–237. Springer International Publishing. DOI: 10.1007/978-3-030-30760-8_20.

Hicke, Rebecca M. M., Maanya Goenka & Eric Alexander. 2022. Word clouds in the wild. In *Proceedings of the IEEE workshop on visualization for the digital humanities*, 43–48. DOI: 10.1109/VIS4DH57440.2022.00015.

Jänicke, S., G. Franzini, M. F. Cheema & G. Scheuermann. 2017. Visual text analysis in digital humanities. *Computer Graphics Forum* 36(6). 226–250. DOI: 10.1111/cgf.12873.

Jänicke, Stefan. 2022. A research-teaching guide for visual data analysis in digital humanities. *Computer vision, imaging and computer graphics theory and applications*. 205–222. DOI: 10.1007/978-3-030-94893-1_9.

Koh, Kyle, Bongshin Lee, Bohyoung Kim & Jinwook Seo. 2010. ManiWordle: Providing flexible control over Wordle. *IEEE Transactions on Visualization and Computer Graphics* 16(6). 1190–1197. DOI: 10.1109/TVCG.2010.175.

Kucher, Kostiantyn & Andreas Kerren. 2015. Text visualization techniques: Taxonomy, visual survey, and community insights. In *Proceedings of the IEEE Pacific visualization symposium*, 117–121. DOI: 10.1109/PACIFICVIS.2015.7156366.

Lee, Bongshin, Nathalie Henry Riche, Amy K. Karlson & Sheelash Carpendale. 2010. SparkClouds: Visualizing trends in tag clouds. *IEEE Transactions on Visualization and Computer Graphics* 16(6). 1182–1189. DOI: 10.1109/TVCG.2010.194.

Liu, Xiaotong, Han-Wei Shen & Yifan Hu. 2015. Supporting multifaceted viewing of word clouds with focus+context display. *Information Visualization* 14(2). 168–180. DOI: 10.1177/1473871614534095.

Lohmann, Steffen, Michael Burch, Hansjörg Schmauder & Daniel Weiskopf. 2012. Visual analysis of microblog content using time-varying co-occurrence highlighting in tag clouds. In *Proceedings of the international working conference on advanced visual interfaces*, 753–756. DOI: 10.1145/2254556.2254701.

Lohmann, Steffen, Jürgen Ziegler & Lena Tetzlaff. 2009. Comparison of tag cloud layouts: Task-related performance and visual exploration. In *Human-computer interaction – interact 2009*, 392–404. DOI: 10.1007/978-3-642-03655-2_43.

Ortiz Pablo, Dalia, Sushruth Badri, Gijs Aangenendt, Mo von Bychelberg & Matts Lindström. 2025. A machine learning pipeline for digitalising historical printed materials – from data collection to a searchable database. In Gerlof Bouma, Dana Dannélls, Dimitrios Kokkinakis & Elena Volodina (eds.), *Huminfra handbook: Empowering digital and experimental humanities* (NEALT Proceedings Series 59), 207–250. University of Tartu Library. DOI: 10.58009/aere-perennius0177.

Östling, Robert. 2018. Part of speech tagging: Shallow or deep learning? *Northern European Journal of Language Technology* 5. 1–15. DOI: 10.3384/nejlt.2000-1533.1851.

Pehrson, Birger. 1965. 1964 blev ett händelserikt år. *Diabetes* (7).

Ren, Hui, Yuan Liu, Gaowa Naren & Junyi Lu. 2024. The impact of multidirectional text typography on text readability in word clouds. *Displays* 83. 102724. DOI: 10.1016/j.displa.2024.102724.

Skeppstedt, Maria & Gijs Aangenendt. 2024. Using the Word Rain technique to visualize longitudinal changes in periodicals from the Swedish Diabetes Association. In *Proceedings of the workshop on visualization for natural language processing*. DOI: 10.2312/vis4nlp.20241132.

Skeppstedt, Maria, Magnus Ahltorp, Kostiantyn Kucher & Matts Lindström. 2024. From word clouds to Word Rain: Revisiting the classic word cloud to visualize climate change texts. *Information Visualization* 23(3). 217–238. DOI: 10.1177/14738716241236188.

Söderfeldt, Ylva. 2025. Joint efforts in the Swedish model: The Swedish Diabetes Association under Nancy Eriksson (1956–1978). In Alexander Dunst, Chantal Marazie, Despo Kritsotaki, Matthew Smith & Nicolas Henckes (eds.), *Mobilising medicine: Health and social movements in global perspective, 1950s-2020s*. (In press). Manchester University Press.

Swedish Labour Movement's Archives and Library, Personal archives of Nancy Eriksson. N.d. *SE/ARAB/1611/4/1/4: Data om Diabetesgården i Nordanede; Diabetesgården i Nordanede.*

Torres Parejo, Úrsula, Jesús R Campaña, M Amparo Vila & Miguel Delgado. 2021. A survey of tag clouds as tools for information retrieval and content representation. *Information Visualization* 20(1). 83–97. DOI: 10.1177/1473871620966638.

van der Maaten, Laurens & Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9. 2579–2605.

Viégas, Fernanda B. & Martin Wattenberg. 2008. Tag clouds and the case for vernacular visualization. *Interactions* 15(4). 49–52. DOI: 10.1145/1374489.1374501.

Wang, Yunhai, Xiaowei Chu, Chen Bao, Lifeng Zhu, Oliver Deussen, Baoquan Chen & Michael Sedlmair. 2018. EdWordle: Consistency-preserving word cloud editing. *IEEE Transactions on Visualization and Computer Graphics* 24(1). 647–656. DOI: 10.1109/TVCG.2017.2745859.

Wu, Yingcai, Thomas Provan, Furu Wei, Shixia Liu & Kwan-Liu Ma. 2011. Semantic-preserving word clouds by seam carving. *Computer Graphics Forum* 30(3). 741–750. DOI: 10.1111/j.1467-8659.2011.01923.x.

Xie, Liwenhan, Xinhuan Shu, Jeon Cheol Su, Yun Wang, Siming Chen & Huamin Qu. 2024. Creating emordle: Animating word cloud for emotion expression. *IEEE Transactions on Visualization and Computer Graphics* 30(8). 5198–5211. DOI: 10.1109/TVCG.2023.3286392.

Xu, Jin, Yubo Tao & Hai Lin. 2016. Semantic word cloud generation based on word embeddings. In *Proceedings of the IEEE Pacific visualization symposium*, 239–243. DOI: 10.1109/PACIFICVIS.2016.7465278.

List of abbreviations

TF-IDF	Term frequency–inverse document frequency
t-SNE	t-distributed stochastic neighbor embedding
OCR	Optical character recognition
NLTK	Natural Language Toolkit

Corresponding authors

Maria Skeppstedt
Department of Linguistics
Stockholm University
maria.skeppstedt@ling.su.se

Magnus Ahltorp
Nakajima Koen Research Institute
magnus@nakajimakoen.org

Kostiantyn Kucher
Department of Science and Technology
Linköping University
kostiantyn.kucher@liu.se

7 Navigating Swedish Salafism
Large language model-augmented content detection and topic modeling using BERTopic with YouTube metadata

Jonas Svensson
Linnæus university

The chapter suggests and provides an example of a Large Language Model (LLM)-augmented method for gaining a quick overview of large sets of YouTube videos using metadata collected through the YouTube API. The case chosen is the Swedish Salafist YouTube channel *islam.nu* that houses 1680 videos. An LLM (GPT-4o mini) is given a prompt to guess the content of videos based on information given in their titles and descriptions. These guesses are then used in an LLM-augmented topic modeling process utilizing the Python library BERTopic and the HUMINFRA resource, the Swedish Royal Library's sentence-transformers model "sentence-bert-swedish-cased". The videos thus placed under topics are then again subjected to processing by an LLM, to produce easy-to-read representations of the topics. This method provides a convenient way to quickly understand the content of YouTube video sets and can serve as a first step in a purposive sampling procedure.

1 *Stating the problem: So much data, not enough time*

We live in a world with massive amounts of data potentially useful for humanistic research, given an understanding of humanistic research as exploring and explaining public expressions of the human mind (Bod 2013). The current point in history is exceptional in the extent to which such expressions are being recorded and archived. This development, linked to a larger process of digitalization of human communication, is simultaneously a blessing and a curse for scholars in the humanities. The blessing is that the material for

HUM INFRA Jonas Svensson. 2025. Navigating Swedish Salafism: Large Language Model-augmented content detection and topic modeling Using BERTopic with YouTube metadata. In Gerlof Bouma, Dana Dannélls, Dimitrios Kokkinakis & Elena Volodina (eds.), *Huminfra handbook: Empowering digital and experimental humanities* (NEALT Proceedings Series 59), 183–205. University of Tartu Library. DOI: 10.58009/aere-perennius0176
© The authors, CC BY 4.0

potential study is unprecedentedly large and diverse. The curse is that it is also *overwhelmingly* so.

In the face of this, there is a need for developing new methods for scholars to gain an overview of a vast material, and based on this select appropriate units for in-depth contextual and qualitative interpretation. Digital methods for collecting, organizing and analyzing large amounts of humanistic data have been available for some time now, but those are often reserved for a select few with specialized competences in how to collect and clean messy data, in transforming that data into computer readable forms, and in selecting, employing or constructing tools for further analysis.

In this article, I argue that the emergence of the technology of generative AI and Large Language Models (henceforth LLMs) opens up for a much wider application of computer-assisted methods in the humanities. The important change is that human-computer interaction increasingly utilizes every day, natural language. Some scholars have already discovered the power of the first-generation generative AI in the form of chatbots (such as ChatGPT) for tasks such as text summarization, image annotation, semantic search, and translation. A few have started to use them to augment their everyday workflow, performing tasks that are otherwise tedious and time-consuming.

2 *The current focus*

In this chapter I will showcase how a simple LLM-augmented process of topic modeling can be used to explore content collected from what is arguably one of the richest, diverse and at the same time overwhelmingly large archives of contemporary expressions of the human mind: the video sharing platform YouTube. While exact statistics are difficult to come by, one estimation from 2025 gives the number of 800 million videos on the platform (but no one really knows), with an additional 500 hours of material being uploaded every minute (Global Media Insight 2025).

From my own perspective, as a scholar within the Study of Religions, particularly interested in contemporary Islamic discourse, YouTube is a goldmine of naturalistic data (Potter & Shaw 2018). Here, individuals and religious groups engage in for example disseminating theology, recording actual religious practices and events, and confronting adversaries, all within a multimodal framework that opens for variety of analytical approaches. Furthermore, the YouTube platform, being an expression of Web 2.0 (O'Reilly 2007), also, to some extent, allows for assessment of impact and responses to the cultural expressions it houses, through statistics on views, likes and not least comments.

Given this, it is somewhat puzzling that YouTube has received relatively little attention within the field of the Study of Religions (Bekkering 2019). One of the reasons, perhaps, is that collecting and processing data through the web interface is cumbersome. YouTube's internal search algorithms are opaque. It is difficult to find appropriate material, and above all to assess how representative or illustrative a particular video is, in light of all videos that potentially *could* have been selected.

In the following, I will showcase how to collect and process metadata via the freely available YouTube API (Application Programming Interface) with the aim of gaining a bird's eye view of a set of videos on the platform to facilitate a purposive sampling of material for more extensive, in-depth analysis. For this I will use the power of generative AI in the form of LLMs. The case is limited to the videos in one YouTube channel that falls within my field of academic interest.

3 *Very short on islam.nu and Salafism*

The YouTube channel *islam.nu* forms part of a larger missionary endeavor of a small group of Muslim activists in a Stockholm suburban area. The group has an extensive online presence, with a website and a podcast. It is present on platforms such as Instagram and Facebook. The orientation is clearly and explicitly Salafist. Salafism, as a "new religious movement" (Meijer 2009) in Islam has been the object of much research worldwide. Generally, Salafism is characterized by a strong focus on a literalist, or fundamentalist, reading of scripture, the Qur'an and even more so the Hadith, i.e. records of (alleged) sayings and doings of the Prophet Muhammad. The word Salafi refers to the pious forefathers, *al-salaf al-salih*, the companions of the Prophet and the following two generations, whose beliefs and practices contemporary Salafis strive to emulate, often in minute detail. This involves shunning almost all religious development after this period in Islamic history as illegitimate innovations, *bida'*. Rejecting religious innovation does not, however, mean rejecting all things modern, including modern technology.

Through its various channels, the group *islam.nu* produces a large amount of material that is potentially available for scholarly, humanistic analysis. Analyses have also been conducted to some extent (Sorgenfrei 2021a,b, Olsson 2020, Olsson et al. 2022, Svensson 2022). The methods used often involve selecting instances of communication (e.g. audio sermons, videos, Instagram postings) and submitting these to in-depth study. So far, however, there is to my knowledge no attempt at providing a systematic overview of the material produced by the group, an overview that could assist in the selection process.

4 *The data*

The *islam.nu* YouTube channel is arguably the most influential Salafi oriented channel in Sweden (Svensson 2021). On September 12, 2024, I gathered metadata for all the 1 680 videos housed on the channel, using the freely available YouTube API (Google 2025), downloading information for each video on:

- Video ID
- Date and time of uploading
- Title
- Description
- Tags
- Number of views
- Number of likes
- Number of comments
- Duration.

While I did this programmatically, using a custom-built Python script,[1] there are several free services on the web offering the same functionality. The information gathered is available via YouTube's web interface, but collecting it manually is not particularly time efficient.

When the metadata was downloaded, the channel had 35 600 subscribers. The 1 680 videos had a grand total of 9.3 million views. The last video was published on September 10, 2024 and the first video in the channel was published nine years before that, on September 3, 2015. The mean number of views for the videos was 5 516, but the median was much lower, 1 019, which indicates a spread around the mean. Indeed, most views for a single video was 1.8 million. This video, along with the 14 other most viewed videos, featured Qur'an recitation. It is only on position 16 that we find the first video featuring material produced "in-house", published on July 2018, with a sermon on *adhan*, or call to prayer. That video has 37 737 views.

The total duration of all videos is approximately 885 hours. This is rather overwhelming as it corresponds to half a year full-time employment for a university researcher in Sweden. To assess the content through watching the videos is hence not viable, which is why a more time-efficient method, using metadata, is needed.

1 The script is available as a Colab from the Linnaeus University Language, Cognition and Culture Lab's resource page: https://moodle.lnu.se/course/view.php?id=65873.

There are three pieces of metadata that could potentially provide information on content: the title of the video, the description and the tags. A researcher could use this information to deduce possible content, and perhaps group the videos into thematic categories, or select a particular set of videos where the metadata points to particularly interesting content. Since the channel in this case consists of "only" 1680 videos, to do such an inventory manually is feasible, but still rather tedious, work. The automated process suggested below can provide some relief.

I have in a previous publication, before the mainstream availability of LLMs, made a computer-assisted attempt at finding thematic patterns in a larger set of Swedish Salafi YouTube videos, utilizing NLP methods of word frequency, word co-occurrence, and semantic network analysis (Svensson 2021). This was a rather complicated process, involving much data cleaning and preprocessing. While the methods applied produced results that did appear reasonable, given my domain knowledge in the field, these results were too general to really justify the effort.

Still, during this attempt, I did discover some characteristics of YouTube metadata that are important for the following. I discovered that overall, the most meaningful piece of metadata to use for computer-assisted processing was the video titles. These were, to a larger extent than the descriptions and tags, unique for each video. Descriptions and tags, and particularly the latter, were often generic for the channel, and to include them in a word frequency or word co-occurrence analysis obscured the results.

I also tried to employ some at the time available methods for topic modeling, but possibly due to the short snippets of text in the video titles, and the overall limited size of the corpus (i.e. all the titles), this did not produce any interpretable results. Such methods rely, ultimately, on word frequency and word co-occurrence metrics, and require larger amounts of text to identify patterns (Golub 2020).

When returning to the same problem a few years later, the technology has advanced. The shortcomings of earlier methods could, at least hypothetically, be overcome through:

1. Providing a label for each video indicating the content, based on *relevant* metadata.

2. Arranging these labels of content into meaningful groups, based on semantics, rather than on word frequencies or word co-occurrences.

For both tasks, I could use the power of LLMs, but in different ways.

5 *Short on BERTopic*

BERTopic is a tool for topic modeling constructed by Martin Grootendorst, and introduced in 2020, but gaining more widespread recognition from 2022 and onwards (Grootendorst 2022). Put in very simple terms, BERTopic takes a corpus of text units (called "documents") as input and uses a pre-trained LLM to arrange these text units in groups based on their *semantic* similarity. These groups, if distinct enough from one another, can be said to relate to a topic or theme in the full corpus. The power of BERTopic lies in the semantic embedding of text units. This embedding does not rely on the corpus itself, but on a huge number of other texts on which the model has been trained. This is like a manual process where the researcher invokes all their previous knowledge in attempting to categorize a particular text unit. Given that it is the semantic content of the document that is used as an arranging principle, the tedious tasks of preprocessing, tokenization, stop-word removal, stemming and/or lemmatization necessary for many previous forms of topic modeling are not a requirement. On the contrary, such manipulation of the text units may obscure the results.

BERTopic has become popular. To date, there are on Google Scholar more than 3 500 scientific papers citing the tool. Some of these contain comparisons with previous methods for topic modeling (Gan et al. 2024, Egger & Yu 2022, Turan et al. 2024), concluding that it is often superior. BERTopic is also, due to a modular structure, highly customizable, not least concerning the main operation, the sentence embedding. Within the framework of HUMINFRA, the KBLab at the Swedish Royal Library has developed a workshop in Google Colab using its own sentence embedder, "sentence-bert-swedish-cased", finetuned for Swedish on the Library's collections (KB Lab 2025). This is also the embedder used in the following.

6 *Preparing texts for BERTopic invoking GPT-4o mini*

For BERTopic to work at its best, the "documents" should then be in a form that makes sense semantically. This is not always the case with YouTube metadata. To give an example from the current set there is the following information on title and description for one video:

Title Surah Al-Alaq - Omar Hisham

Description

Koranen | Kapitel 96 - Al-'Alaq

Recitör: Omar Hisham

https://www.youtube.com/playlist?list=
PLaIDlec9j7J2o2Dy3yiMynYtZnPDIfkfZ

Here, the title alone does provide some information that could be interpreted as relating to the Qur'an, chapter 96. The name "Omar Hisham" does not provide any additional information. Combined with the information in the description, however, a qualified guess could be made. Here, Omar Hisham is also given a specific role, as a reciter (albeit the Swedish word is incorrect. It should be "recitatör" and not "recitör"). For someone with at least basic domain knowledge in Islam, a guess concerning the content would be something like: "A recitation of Sura al-'Alaq (96) in the Qur'an by Omar Hisham". Such a descriptive label could, potentially, facilitate the topic modeling process. Going through all the metadata manually, producing such guesses for all videos that could then be subjected to a topic modeling process would most probably defeat the purpose of using computer-assisted methods, as it could give a fairly good overview of recurring themes. And it would take time.

Summarizing content is an area in which current LLMs excel. Entering the information of title and description into the web application ChatGPT (using GPT-4o) with a prompt to guess the content produces the following:

> This YouTube video likely features a recitation of Surah Al-Alaq (Chapter 96 of the Qur'an), performed by Omar Hisham, who is identified as the reciter. The video may include a traditional Arabic recitation, possibly with subtitles or explanations in another language (like Swedish, given the term "Koranen" for "Qur'an" and the Swedish spelling of "Recitör").

This is impressive, but also contains perhaps a bit too much guessing. Still, entering each piece of metadata into a web based Chatbot is probably more time consuming than guessing the content manually.

To automate the process, I created a simple script in Python where I invoked OpenAI's recent model GPT-4o mini via the OpenAI API. There are of course several LLMs that could have been chosen for the task, including the full GPT-4o model that is the default model under the hood for ChatGPT, but GPT-4o mini is, at the time of writing, widely recognized as capable, fast, reliable and cost-efficient.

The task given to GPT-4o mini was to use the information in the title and description metadata for each video to guess the content. As an extra task, I also prompted the model to extract names of persons possibly featuring in the video, if mentioned (in the case above Omar Hisham). The latter was merely a test of using LLMs for Named Entity Recognition (NER) and it will not be further considered below.

The following was the prompt instruction. It was given in Swedish and has here been translated into English.

> For each object in an array of JSON objects, create two new key/value pairs.
>
> 1. A qualified guess about the content of a YouTube video based on the values of the video's title and description. The guess should be in Swedish, limited to the information in the title and description, without adding information that is not explicitly mentioned.
>
> 2. Information about who is speaking in the video. If the speaker cannot be identified, leave SpeakerGuess as an empty string.
>
> Return only the video Id and the two key/value pairs ContentGuess and SpeakerGuess.
>
> [Here, I included a list of further detailed instructions for the format of the output, and also an example of the desired output, a so-called "one-shot prompting"]
>
> Keep strictly to the example. Do not add the word JSON to the output. Make sure the output is in a strict JSON format. If you cannot guess the content, let the title be your guess.

GPT-4o mini delivered guesses for the content of all the 1680 videos. The titles and descriptions were presented to the model in batches of 15 videos at a time, to fit into the model's context window. The whole process took 11 minutes. For the example given above, the result was: *Recitation av Koranens kapitel 96, Al-'Alaq* (Recitation of the Qur'anic chapter 96, al-'Alaq).

7 *Topic modelling with BERTopic*

The result of the LLM-augmented content guessing was used as the "documents" for BERTopic to work with. The KBLab pretrained model for Swedish is showcased in the Colab workshop available via HUMINFRA (KB Lab 2025). The documents here are Swedish parliament petitions, perhaps an ideal test case, and the results returned are highly impressive. The data

Table 1: Comparison between BERTopic results with different minimal number of items in topic setting.

Min items per topic	Topics	Outliers
5	87	301
10	51	394

for the current chapter, i.e. the LLM-generated labels for YouTube videos might be considered less ideal. They are in Swedish, but also in a particular Swedish "religiolect" (Hary 2011) that mixes Swedish words with specific religious terms with Arabic origin, at times also using archaic Swedish words as well as idiosyncratic neologisms. It is not conceivable that there is enough material in the training data for "sentence-bert-swedish-cased" to cater for this particular "Islamic Swedish". I did try out the default BERTopic model in multilingual mode, but since this article focuses on the use one of the HUMINFRA resources, I will not follow up with a model comparison here.[2]

The output of the BERTopic topic modeling is in the form of a list of topics, based on a procedure that places *semantically* similar text units (i.e. the content guesses) in proximity to one another and identifies clusters of similar texts that are sufficiently distinct from other clusters of similar texts. Unlike other tools for topic modeling, BERTopic does not force all texts into such clusters. Texts that are difficult to assign to a cluster are instead considered to be outliers.

Besides choosing the pre-trained model to use for the embedding of text units, there are several other parameters that could have been adjusted. Here, I used the default settings in the KBLab workshop, except for one. Since the number of videos in the set is small, I decided to set the minimum number of items under each topic to five instead of ten. This was done to reduce the number of outliers (see Table 1)

The topic modeling process took 40 seconds. The result was 87 topics and a total of 301 videos as outliers. It is possible do further processing in BERTopic to reduce the number of outliers (Grootendorst 2025c), forcing them under topics, but for the following, I chose to disregard the outliers altogether.

2 The results from the two different models, using a minimum of five and ten items per topic, are available for comparison at https://github.com/jsnhum/huminfra_cookbook.

8 *What is under the topics?*

The default way in which BERTopic represents the topics is through a list of words particularly significant for the texts under that topic. Significance is calculated using a version of the established metric of TF-IDF or "Term Frequency-Inverse Document Frequency".[3] In simple terms, this procedure attempts at extracting terms significant for a particular text, in relation to all other texts in a corpus to which the text belongs. This is done by identifying terms that occur frequently in the text and weighing this against the number of other texts in the corpus in which the same term occurs. The same principle is used in BERTopic, but with the difference that here it is the frequency of a term in all texts under a topic that is weighted against the number of topics whose underlying texts contain that term. The procedure is referred to as "Class-based TF-IDF" (c-TF-IDF; Grootendorst 2025b).

In Figure 1, there is a hierarchical layout for all 87 topics with representations generated by c-TF-IDF, displaying the three most significant terms. The dendrogram, produced in BERTopic, also indicates calculated similarity between topics.

The results from c-TF-IDF may not be an ideal way to represent the topics. The content guesses are short snippets of text, which skews the metrics of term frequency. Even single words may have large relative frequency. Furthermore, the use of neologism and Arabic terms may also obscure the result.

Still, even in the c-TF-IDF-based representation one does find those that do make sense.

Hence, for example, the representation for topic number 45 is easy to interpret. It has the following representation (using the ten most significant terms): *the atheism*, *critical*, *atheism*, *analysis*, *simplified*, *counter arguments*, *trailer*, *society*, *against*, *focus*. An educated guess would be that this topic covers videos in which atheism is discussed critically.

Deciphering other topics may require some more domain specific knowledge. Number 55 has the representation *wudhu*, *instructions*, *guidelines*, *performance*, *learn*, *video*, *you*, *how*, *knowledge*, *lecture*. This most probably indicates videos with instructions on how to perform *wudu'*, a form of washing to achieve ritual purity.

Yet other topics are more difficult, if not impossible, to infer from the representation, such as topic 79, with the representation: *200*, *the belief dogma* [*trosläran*], *Islamic*, *questions*, *course*, *call to prayer*, *iqamah*, *inclusive*, *11*, *doubt*.

3 For a highly accessible introduction to TF-IDF, see Lavin (2019), who also provides some history for the technique, dating back to the 1970s.

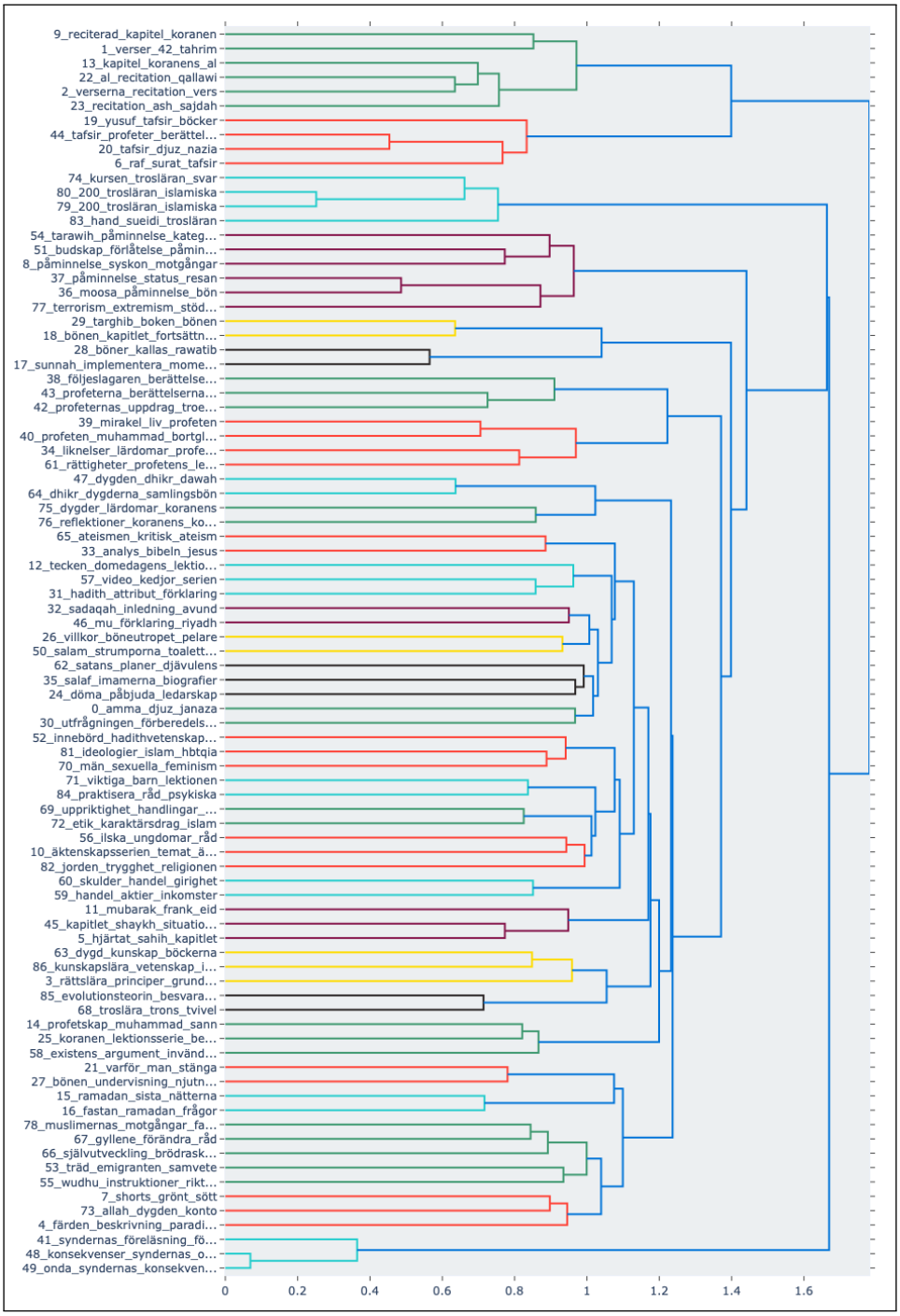


Figure 1: BERTopic hierarchical clustering

It should be remembered, however, that the grouping of videos under topics is done *before* the representation is extracted through c-TF-IDF. The grouping is based on semantics using a pretrained model, not on word prevalence or frequency neither in GPT-4o mini's content guess, nor in the corpus of all such guesses. This means that there are other ways to arrive at useful representations of the topics.

9 *GPT-4o mini again*

One such way is to again invoke the power of LLMs. This is suggested also in the BERTopic documentation, with instructions on how to implement this directly in the process of topic modeling (Grootendorst 2025a). Here, however, given the specific material at hand, I chose to do this representation on the side.

Each topic arrived at by the BERTopic modelling contains a set of short text (the content guesses). I provided GPT-4o mini with these sets with the following prompt (translated from Swedish):

> You are an assistant with great competence in finding summary labels for a list of short descriptions of YouTube videos from a Salafist YouTube channel. You receive a list of descriptions of YouTube videos and provide a summarizing title. The title should be short.
>
> [here follows an example]

The result can be seen in Appendix 1. The original labels in Swedish have here been translated into English.

10 *Interpreting the results*

Now then, we have arrived at a list of 87 topics covering 1 379 of the total 1 680 videos present in the *islam.nu* channel. BERTopic offers much additional functionality beyond identifying topics, in terms of further exploring and arranging those topics. For the purposes in this chapter we can now turn to a qualitative interpretation based on (1) the LLM generated topic representation (2) the hierarchical outline of topics as displayed in the dendrogram in Figure 1 and (3) domain knowledge to achieve what was presented as the goal in the outset, a bird's-eye characterization of the content of the *islam.nu* YouTube channel, and a guide for further purposive sampling.

Islam.nu hosts material that could be expected of a Salafi oriented YouTube channel: a strong focus on basic beliefs and rituals, the scriptures, the prophet

Muhammad and early Islam. Eighty-seven topics may appear a lot, and some of them could probably be merged, especially considering topic similarity as indicated in the dendrogram. One of the more obvious examples of this is the "topic" already mentioned several times: Qur'an recitations. It would appear as if these generally fall under topics 1, 2, 9, 13, 22, 23, all joined together at the top of the dendrogram. As mentioned above, Salafism is highly oriented towards the Islamic scriptures, the Qur'an and the hadith, and historically towards the early period in Islamic history. Both the Qur'an and the hadith recur often in the topics. When history is addressed, it is either the mythological salvation history of Qur'anic figures, or the Prophet Muhammad and his companions. There are few, if any, topics that concern the historical development of Islamic tradition. This is in line with the character of prototypical Salafism as a "deculturalized" form of Islam, where there is a discursive void stretching between the early history of Islam and the here and now.

There are other significant traits of Salafism noticeable in the topic representations. There is a specific topic on *hisba,* i.e. the ritualized practice of "informing" other Muslims about what God expects of them and what He forbids (topic 24), often from a somewhat elevated position of a select few (i.e. Salafis themselves). Such advice to other Muslims appears to be on basic beliefs, ethics and morals, sins and virtues, and obligatory ritual practices. It is striking that many of the topics appear to address what could be seen as "basic" religious beliefs and practices (fasting, almsgiving, beliefs in afterlife, paradise etc.), already known to most Muslims. This mirrors a common view among Salafis that although Muslims may have some knowledge and beliefs about these basics, that knowledge is incomplete or flawed and above all tainted by established cultural understandings and innovations in belief and practice. A general feature of Salafism is the stated need for beliefs and practices to be "purified" from anything that does not have a clear foundation in the scriptures. One can note the apparent strong focus on such a basic practice as prayer stretching over several topics (17–18, 21, 26–29). But "purification" is not only metaphorical. That there is a special topic on cleanliness and hygiene (50) does rhyme well with Salafism in general being strongly (some would claim obsessively) attentive also ritual purity, and whatever substances or actions that can jeopardize such purity.

While they appear rather scarce, there are also topics that goes a bit beyond the issue of correct individual faith and practice. Such include what appears to be comparisons between Islam and other forms of belief systems or systems of thought that challenge it, and I would assume, to the benefit of the former. Those other include atheism (65), science (85), Christianity (33) and political ideologies (81). There appear to be distinct topics that address

some issues on social relations, such as marriage and family relations (10, 56) and gender (70), as well as topics that address economy and commerce (59, 60). One topic could be mentioned in particular, i.e. number 77, represented as "Islam, Extremism, and Terrorism". Previous research on the group behind *islam.nu* has identified them as belonging to a particular category of "puritan" Salafism, not partaking in politics, at least not party politics, and particularly taking public stand against militant forms of Salafism (Ranstorp et al. 2019).

The above summarization provides a bird's-eye view. It gives some very broad outline of the *possible* content of the *islam.nu* YouTube channel. That this outline corresponds to what is already known, and somewhat expected, could be taken as an indication that the topic modeling was successful. Further probing into the LLM generated content guesses for videos housed under each topic would perhaps give a better basis for such evaluation. I cannot do a full-scale evaluation here, and it would defeat the purpose. But I would like to point to a couple of topics that can help illustrate why I deem BERTopic, and its reliance on pre-trained models, useful for the task at hand.

Topic 30 received the LLM-generated representation "Death and the Afterlife ", which could be seen as rather specific. The topic houses six videos. Four of these contain the word "death" in the title and the description, and in the LLM-generated content guess. However, two do not: one bears the title "Reflections from the funeral" and the second "The questioning in the grave", referring to a particular notion of partial punishment/rewards in the grave between death and final resurrection. From a semantic perspective, and judging from the titles, the placement of these videos under topic 30 does make sense.

Another example can be found in the topic "Bible Analyses and the Islamic View of Jesus" (33). Although Jesus, Islam and the Bible are words that occur in the LLM-generated content guesses of four of the eight videos under this topic, all are absent from the remaining four. BERTopic, however, judged (correctly in my judgment) the content guesses "Analysis of the Gospel of Mark", "Analysis of the Gospel of Luke", and "Analysis of the Gospel of John", to belong under this topic, resting on the inbuilt "knowledge" of the underlying LLM of that there is a connection between the words "Jesus", "Bible" and "the Gospels".

It should be stressed, albeit it is not an issue to be pursued further, that the topic modeling of YouTube videos using BERTopic showcased here makes it easy to combine the results with additional pieces of metadata, in order to asses for example if some topics receive more interaction than others (in terms of the number of views, likes and comments) or if certain topics are more prevalent in certain periods of time than in others. Such processing of

the results may be interesting, but they are perhaps not at the core of much humanistic research, with a stronger focus on detail, on careful, culturally contextualizing analysis of discourse and behavior, and a quest to interpret and understand individual expressions of the human mind, i.e. in this case the actual content of the videos.

This, then suggests topic modeling not as an end in itself, but as a means to an end, more specifically the possibility to find, in a large collection of data, specifically representative, or illustrative, cases for in-depth analysis. The task here might be identifying material that can help answer certain predetermined questions (such as *islam.nu* views on e.g. gender roles or terrorism), or to identify new, previously unattended to topics suitable for in-depth study.

11 *Working further*

As much as I would have liked to provide an example of such an in-depth exploration of one of the topics, it is not possible. The videos most certainly contain material where individuals express their religious beliefs. As such it becomes sensitive data which, according to both Swedish and European legislation, is forbidden to collect or process. An exception to this general prohibition could possibly apply in the current case, when the sensitive data has been manifestly made public by the research person. *Islam.nu* is an open missionary channel, reaching out to a general public, Muslim and non-Muslim, and the individuals that feature in videos are publicly professing their beliefs. However, whether this exception applies is not for me to decide. The decision lies with the Swedish national ethical board, to whom a lengthy application must be submitted, together with a fee of 5000 Swedish Crowns. This cumbersome procedure has (rightfully in my opinion) been the object of recent criticism in Swedish academia, particularly from the humanities and the social sciences where publicly available information on expressed beliefs, chosen ethnic affiliations and political stances of e.g. poets, novelists, opinion makers, other academics or religious and political leaders constitute important data. Nevertheless, non-compliance with the demand to have research on sensitive data preapproved is a criminal offence, punishable by fines or imprisonment.

There is, however, nothing hindering me from, as a closing segment, outlining a process for how a further probing into the results from the current topic modeling, for example on the videos categorized under the topic "Critique of Atheism and Arguments Against Atheism" (65). It is interesting, since it belongs to a set of topics in which *islam.nu* appears to direct attention

not only towards Islamic tradition, but also towards the larger society. The topic houses seven videos. A word search on the word stem *ateis-* 'atheis-' in the titles of all videos reveals three additional videos that may be of relevance. This is important, since it shows that there is reason to double check the results of the topic modeling. It does not, however, mean that BERTopic failed. It must be remembered that BERTopic places an item (here a content guess) under one, and only one, topic. Hence, a video guessed to feature an "Interview with a Muslim convert concerning his journey from Atheism to Islam", was placed under topic 84, "Guidance and Advice for Prospective Muslims", in which most videos apparently target conversion to Islam. Another video was labelled "Atheism | Analysis of Western ideologies and their relation to Islam", and placed under topic 81, "Political Ideologies and Beliefs". GPT-4o mini labelled the third and last as "A series of lectures on how to handle doubt and atheism", which BERTopic considered to be an outlier.

Adding these three videos to the sample, the total number of videos for a possible in-depth analysis is ten, with a total length of slightly more than seven hours. Assessing the content by actually watching is perhaps not the most time efficient way to approach the material. Some YouTube videos come with subtitles, and in some cases, the transcript can be accessed from within the YouTube platform. Not all videos have that option though, and the ten videos in this set do not.

Recent amendments to copyright law makes it possible for researchers in Sweden to copy and (securely and temporarily) store copyrighted material that they have legal access to for the purpose of datamining. This would allow for the downloading of the ten videos, as videofiles or soundfiles, and submit them to perhaps one of the more conspicuous advancements in research that the current LLM revolution has to offer: automatic transcription. The state-of-the-art tool for this is OpenAIs Whisper, a free, multilingual transcription (speech-to-text) tool, which, incidentally, is rumored to have been created for the purpose of harvesting data from recorded speech online, and in particularly on YouTube, to use for LLM training (Metz et al. 2024).

Running Whisper on Google Colab transcribes sound at a rate of approximately a third of the recorded time, i.e. a video of one hour transcribes in 20 minutes, with often next to perfect result.[4] To use this method for sensitive data, however, might not be advisable since it means sharing the data with Google. In the case of YouTube videos, however, this may be less of a problem, since they are already, at least partly, owned by Google.[5]

4 For a publicly available Google Colab page for transcription via OpenAI's Whisper (with instructions in Swedish), see Linnaeus University Language, Cognition and Culture Lab's resource page: https://moodle.lnu.se/course/view.php?id=65873.

One hour of speech amounts to around 15 pages of text, single spaced. So, the total number of pages of transcripts of the 10 videos in this example would be around 100.

Transcribed speech is generally not a pleasant read. If I were to assess further the content of the videos, I would probably first of all submit the texts to an LLM for summarization and information retrieval. While not totally reliable, I have found that an iterative process of oscillating between original files, and the LLM summarization facilitates the identification of segments of text relevant in-depth analysis. In the current case, for example, conceivable questions to ask would be for the LLM to extract segments from the texts that provide argumentations against atheism, discussions on how atheism relates to Islam and definitions of atheism.

12 *Concluding remarks*

The example given here of an LLM augmented topic modeling in search of patterns and themes in sets of YouTube videos, using easily accessible metadata, does seem to hold some promise for anyone interested in systematically exploring this enormous, and continuously expanding, archive of expressions of the human mind. The metadata collected through the YouTube API is here limited to one channel and 1680 videos. The method can easily be scaled up, to process information contained in much larger datasets (see e.g. Svensson, in press; Svensson, forthcoming). Here it can be mentioned that the YouTube API also allows for collection of data via searches that can circumvent the search algorithms applied in the web interface.

However, some words of caution are warranted. The use of an LLM to produce more semantically meaningful "guesses" of the content in videos, using titles, descriptions and potentially tags, to gain better results from the BERTopic modeling, is risky. LLMs are still prone to hallucinating, providing responses that are not in line with the expected, in unpredictable ways. Although I have not systematically assessed the quality of GPT-4o mini's guesses for the current article, I have found no examples of hallucinations. I did, however, find that the results generated both in content guesses and in the final LLM-generated representations where not fully consistent over consecutive attempts. There were slight variations in wording and focus.

I have above limited myself to text content in the metadata (and hypo-

5 To cater for the bulk transcription of sensitive data, Linnaeus University has, within the framework of the Huminfra infrastructure, invested in a custom assembled, stand-alone computer to run Whisper. It currently transcribes sound at a speed of approximately 1:1 and is housed in Linnaeus University Language, Cognition and Culture Lab (LiLa).

thetically in transcripts). This is a limitation, given the multimodal character of YouTube data. BERTopic is not (yet) multimodal. However, there are ways in which image material can be processed, again invoking the power of LLMs, or perhaps rather LMMs, "Large Multimodal Models". Besides textual metadata, the YouTube API allows for downloading image material, or more specifically thumbnails of the videos. Such image material can become an additional basis for the content guesses, using a multimodal model that can perform image annotation.

A more accurate topic modeling than the one above would be to transcribe all the videos, alternatively download subtitles when available, and use these texts as the "documents" submitted to BERTopic. However, BERTopic has a limit to the length of texts that it can process, around 300 words. Transcripts would have to be summarized, a process that could be done using LLMs. However, this would incur significant costs, at least if commercial models were to be used. Just as there are LLM-powered techniques for (relatively) quick transcription of spoken content in the videos, there are also methods to summarize the image content, using an LLM with "vision" capabilities (for an example, see OpenAI 2025).

Such expansions to increase the quality of the topic modeling would, however, beat the purpose of the procedure outlined in this chapter. The aim has been to introduce a quick and cost-efficient method to arrive at a bird's eye view of YouTube material and facilitate purposive sampling. The whole process, from downloading the metadata from the YouTube API to arriving at a dataset with videos assigned to a topic with an LLM generated representation, takes around 20 minutes. It is only considering how much time and effort the same procedure would demand if done fully manually that the LLM-augmented method can be evaluated, considering also the loss of the researcher's full control over the processes that it involves. Still, it is a method that does not challenge or replace the specific competences associated with humanistic research, in terms of an informed culturally contextualizing interpretation of expressions of the human mind. It is merely a method for finding such expressions and perhaps justifying that they are made into objects for in-depth analysis in the first place.

References

Bekkering, Denis J. 2019. Studying religion and YouTube. In Alphia Possamai-Inesedy & Alan Nixon (eds.), *The digital social: Religion and belief*, 49–60. Berlin: De Gruyter. DOI: 10.1515/9783110497892-003.

Bod, Rens. 2013. *A new history of the humanities: The search for principles and patterns from antiquity to the present*. Oxford University Press, USA. DOI: 10.1093/acprof:oso/9780199665211.001.0001.

Egger, Roman & Joanne Yu. 2022. A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify Twitter posts. *Frontiers in Sociology* 7. DOI: 10.3389/fsoc.2022.886498.

Gan, Lin, Tao Yang, Yifan Huang, Boxiong Yang, Yami Yanwen Luo, Lui Wing Cheung Richard & Dabo Guo. 2024. Experimental comparison of three topic modeling methods with LDA, Top2Vec and BERTopic. In Huimin Lu & Jintong Cai (eds.), *Artificial Intelligence and robotics*, 376–391. Singapore: Springer.

Global Media Insight. 2025. *YouTube statistics 2025* (*demographics, users by country & more*). https://www.globalmediainsight.com/blog/youtube-users-statistics.

Golub, Koraljka. 2020. Automatic identification of topics: Applications and challenges. In Joacim Hansson & Jonas Svensson (eds.), *Doing digital humanities: Concepts, approaches, cases*, 5–26. Växjö: Linnaeus University Press.

Google. 2025. *Add YouTube functionality to your app*. https://developers.google.com/youtube/v3.

Grootendorst, Maarten. 2022. *BERTopic: Neural topic modeling with a class-based tf-idf procedure*. https://arxiv.org/abs/2203.05794.

Grootendorst, Maarten. 2025a. *6B. LLM & generative AI*. https://maartengr.github.io/BERTopic/getting_started/representation/llm.html.

Grootendorst, Maarten. 2025b. *C-TF-IDF*. https://maartengr.github.io/BERTopic/api/ctfidf.html.

Grootendorst, Maarten. 2025c. *Outlier reduction*. https://maartengr.github.io/BERTopic/getting_started/outlier_reduction/outlier_reduction.html.

Hary, Benjamin. 2011. Religiolect. In Deborah D. Moore, Anita Norich & Joshua L. Miller (eds.), *Critical terms in Jewish language studies*, 43–53. Ann Arbour: Frankel Institute for Advanced Judaic Studies.

KB Lab. 2025. *BERTopic workshop: analyzing Swedish parliamentary motions*. https://colab.research.google.com/drive/10kB3wfoHSfZE48vEKmznIw-ff36uR8gs?usp=sharing.

Lavin, Matthew J. 2019. Analysing documents with TF-IDF. *Programming historian* 8. DOI: 10.46430/phen0082.

Meijer, Roel (ed.). 2009. *Global Salafism: Islam's new religious movement*. London: Hurst & Company.

Metz, Cade, Cecilia Kang, Sheera Frenkel, Stuart A. Thompson & Nico Grant. 2024. How tech giants cut corners to harvest data for A.I. *The New York*

Times. https://www.nytimes.com/2024/04/06/technology/tech-giants-harvest-data-artificial-intelligence.html.

O'Reilly, Tim. 2007. What is Web 2.0: Design patterns and business models for the next generation of software. *Communications & Strategies* (65). 17–37.

Olsson, Susanne. 2020. Advising and warning the people: Swedish Salafis on violence, renunciation and life in the suburbs. In Simon Stjernholm & Elisabeth Özdalga (eds.), *Muslim preaching in the Middle East and beyond: Historical and contemporary case studies*. 155–172. Edinburgh: Edinburgh University Press.

Olsson, Susanne, Simon Sorgenfrei & Jonas Svensson. 2022. Puritan Salafis in a liberal democratic context. In Magnus Ranstorp, Linda Ahlerup & Filip Ahlin (eds.), *Salafi-jihadism and digital media*, 92–112. Routledge. DOI: 10.4324/9781003261315.

OpenAI. 2025. *Processing and narrating a video with GPT's visual capabilities and the TTS API*. https://cookbook.openai.com/examples/gpt_with_vision_for_video_understanding.

Potter, Jonathan & Chloe Shaw. 2018. The virtues of naturalistic data. In Uwe Flick (ed.), *The SAGE handbook of qualitative data collection*, 182–199. London: SAGE.

Ranstorp, Magnus, Filip Ahlin, Peder Hyllengren & Magnus Normark. 2019. *Between Salafism and Salafi-Jihadism: Influence and challenges for Swedish society*. Swedish Defence University, Center for Asymmetric Threat Studies. URN: urn:nbn:se:fhs:diva-8534.

Sorgenfrei, Simon. 2021a. "Perhaps we see it in negative terms, but, ultimately, it is positive.": The responses of Swedish Salafis to COVID-19. *Tidskrift for Islamforskning* 15(2). 40–62. DOI: 10.7146/tifo.v15i2.125959.

Sorgenfrei, Simon. 2021b. Crowdfunding Salafism. Crowdfunding as a Salafi missionising method. *Religions* 12(3). 209. DOI: 10.3390/rel12030209.

Svensson, Jonas. 2021. Mönster i svensk YouTube-salafism. Datorassisterad metadataanalys som urvalsmetod i religionsvetenskapliga studier av videomaterial online. *Religionsvidenskabelig tidskrift* 73. 60–83. DOI: 10.7146/rt.vi73.129550.

Svensson, Jonas. 2022. Mönster i rappakaljan: Ett test med datorassisterad, urvalsgenererande fjärrläsning av automatiskt transkriberade salafistiska predikningar på svenska. *Tidsskrift for islamforskning* 16(1). 134–155. DOI: 10.7146/tifo.v16i1.132561.

Svensson, Jonas. In press. "Allah Says Beat Them". An Analysis of Non-Muslim Islams Justifying Domestic Violence. In J. Petersen & A. Ackfeldt (eds.), *Non-muslim islams*. Edinburgh: Edinburgh University Press.

Svensson, Jonas. Forthcoming. Muhammed och monoteisternas fiende. *Chaos*.

Turan, Salih Can, Kazım Yildiz & Büşra Büyüktanir. 2024. Comparison of LDA, NMF and BERTopic topic modeling techniques on amazon product review dataset: A case study. In Fausto Pedro García Márquez, Akhtar Jamil, Isaac Segovia Ramirez, Süleyman Eken & Alaa Ali Hameed (eds.), *Computing, Internet of Things and data analytics*, 23–31. Cham: Springer. DOI: 10.1007/978-3-031-53717-2_3.

Appendix 1 *LLM generated representations*

	Label		Label
0	Djuz 'Amma and Janaza (Funeral)	14	Proofs for Muhammad's Prophethood and God's Existence
1	Quran Recitations	15	Ramadan: Preparation, Practice, and Consequences
2	Quran Recitations of Various Surahs	16	Fasting and Ramadan: Rules, Questions, and Advice
3	Islamic Jurisprudence: Basic Principles	17	Sunnah Prayers and Prayer Rituals
4	Paths to Paradise	18	Chapter on Prayer and the Virtue of the Friday Prayer
5	Sahih al-Bukhari and the State of the Heart	19	Tafsir of Surah Yusuf and Djuz 'Amma
6	Tafsir of Surah al-A'raf	20	Tafsir of Various Surahs and Djuz 'Amma
7	Brief Inspirational Messages #Shorts	21	The Importance of Prayer in Islam
8	Reminders and Teachings on Islamic Virtues and Practice	22	Quran Recitation and Memorization
9	Recitations of Surahs	23	Quran Recitations – Various Surahs
10	Marriage and Family Bonds in Islam	24	Encouraging Good, Forbidding Evil, and Leadership
11	Eid Mubarak, Friday Prayer, and Religious Advice from Shaykh Abdulwadod Frank	25	The Quran as the Revelation of God
12	Signs of the Day of Judgment and Belief in Fate	26	The Importance and Practice of Prayer
13	Quran Recitations and Reflections on Content		

(*Continues on next page*)

(*Continued from previous page*)

	Label
27	Prayer: Focus, Engagement, and Direction
28	Prayer in Islam and Sunnah Prayers
29	Chapter on Prayer
30	Death and the Afterlife
31	Aqidah, Hadith Jibril, and Supplication/Invocation
32	Zakat, Sadaqah, and Islamic Rules
33	Bible Analyses and the Islamic View of Jesus
34	The Prophet's Sunnah and Hadith
35	Biographies of Imams and Theology
36	Advice and Reminders with Moosa Assal
37	Tawhid, the Prophet's Status, and the Quran
38	Stories of the Companions and the Prophets
39	The Life and Qualities of the Prophet Muhammad
40	The Life and Teachings of the Prophet Muhammad ﷺ
41	Sins, Their Consequences, and Forgiveness
42	The Stories and Qualities of the Prophets
43	Stories of the Prophets and Sacred Figures
44	Tafsir of Surah Al-Anbiya and Djuz Tabarak, along with Quran Memorization
45	Sahih al-Bukhari, Biographies, and Reflections on Surahs and Hadith
46	Explanations of Surahs, Hadith Literature, and Islamic Books
47	Virtues and Dawah
49	The Evil Consequences of Sins
50	Salam, Cleanliness, and Hygiene in Islam
51	Reminders on Faith and the Day of Judgment
52	Islamic Values and Rights
53	Motivation and Reminders of Good Deeds
54	Tarawih Reminders and Tawhid
55	Wudhu (Ablution) and Umrah
56	Respect for Parents, Anger, and Social Problems
57	Islamic Teaching and Reflections
58	Arguments and Proofs for God's Existence
59	Islamic Trade and Investments
60	Economy and Financial Ethics in Islam
61	Goodness, Parents, and the Prophet's Rights
62	Satan's Plans and Evil
63	Wisdom and Knowledge
64	Hajj, Dhikr, and Hadith
65	Critique of Atheism and Arguments Against Atheism
66	Personal Development and Relationships
67	Life Guidance and Advice in Islam
68	Faith, Doubt, and Aqidah
69	Faith, Sincerity, and Good Deeds
70	The Role and Status of Men and Women in Islam
71	Muslim Education and Important Lessons
72	Ethics, Morality, and Character in Islam
73	Faith and Love of Allah
74	200 Questions on Islamic Theology
75	Virtues of the Quran and Reflections
76	The Quran: Reflections and Analyses
77	Islam, Extremism, and Terrorism

(*Continues on next page*)

(*Continued from previous page*)

	Label
78	Muslim Behavior, Unity, and Handling of Trials
79	Islamic Theology and Practice
80	200 Questions on Islamic Theology
81	Political Ideologies and Beliefs
82	Islam and Muslims: Fajr Reminders and the Importance of Religion
83	Islamic Theology and Evidential Arguments
84	Guidance and Advice for Prospective Muslims
85	Theology, Science, and Handling Doubts in Islam
86	Islamic Knowledge and Worship

Corresponding author

Jonas Svensson
Department of Cultural Sciences
Linnæus University
jonas.svensson@lnu.se

8 A machine learning pipeline for digitalising historical printed materials – from data collection to a searchable database

Dalia Ortiz Pablo
Uppsala University

Sushruth Badri
Uppsala University

Gijs Aangenendt
Uppsala University

Mo von Bychelberg
Uppsala University

Matts Lindström
Uppsala University

Recent developments in the fields of machine learning and computer vision have created new opportunities for the digitalisation of printed historical materials. However, successful integration of machine learning models requires interdisciplinary collaboration between computer- and data scientists, researchers, librarians and/or archivists, and digitisation experts. This chapter describes a comprehensive pipeline designed to address the challenges of digitalising printed historical materials, from document-scanning best practices to incorporating state-of-the-art machine learning techniques. It aims to streamline the management and processing of historical data, making the digitalised materials accessible and searchable through the application of machine learning techniques. The content of this chapter encompasses scanning best practices, annotation approaches, model training, and deployment. This chapter presents a collection of useful tools for each stage of building a machine learning model, step-by-step instructions and example notebooks designed to be easily adapted to other cases.

1 *Introduction*

Availability and accessibility of historical data has always been a condition for Humanistic research, long before the term "Digital Humanities" was coined. When European Renaissance humanism emerged in the 1400s, this also meant the establishment of a new mode of knowledge production that was focused on the uncovering and reconstitution of historical (classical) remnants and records – thus forming the basis for the scholarly endeavour that would eventually, over the centuries and across various specialised

HUM INFRA Dalia Ortiz Pablo, Sushruth Badri, Gijs Aangenendt, Mo von Bychelberg & Matts Lindström. 2025. A machine learning pipeline for digitalising historical printed materials – from data collection to a searchable database. In Gerlof Bouma, Dana Dannélls, Dimitrios Kokkinakis & Elena Volodina (eds.), *Huminfra handbook: Empowering digital and experimental humanities* (NEALT Proceedings Series 59), 207–250. University of Tartu Library. DOI: 10.58009/aere-perennius0177 © The authors, CC BY 4.0

university disciplines, become configured as modern "Humanities". In the early- to mid-1800s, at the same time as History was formed as a professional discipline with new ambitions of rigour and objectivity modelled on the natural sciences, this impulse was further stimulated by a growing abundance of archival sources that became available in post-revolutionary (1789) Europe and its emerging nation states. The archives of Kings and other dynastic rulers suddenly became properties of nations and institutions such as national archives and national libraries were established. In the wake of this new archival and documentary landscape, influential historiographers such as Leopold von Ranke expressed a belief that History, finally, would be truthfully uncovered – "wie es eigentlich gewesen" (eng. "how it really was"; Von Ranke [1824] 1885: VIII).

Today, as the archival landscape is yet again changing through digitalisation, Digital Humanities similarly promises new modes of Humanistic knowledge production, often in conjunction with the application of new machine learning and artificial intelligence techniques. However – whether researchers are looking to use such new possibilities to further glean from posterity unnoticed clues and details (Ginzburg 1979), or to discover patterns at a macro-scale that would otherwise be impossible to observe (Jockers 2013, Moretti 2005) – any such endeavour presupposes the basic operation of converting analog, historical documents to digital representations that can be further processed and analysed.

Integrating state-of-the-art machine learning requires technical skills and knowledge about how to construct a digitalisation pipeline and implement machine learning models. These skills are not typically associated with the average humanist, but rather with computer- and data scientists. At the same time, successful digitalisation requires intimate knowledge of the physical collections, usually held by archivists or librarians working in cultural heritage institutions, with support from digitisation experts (Smith & Whearty 2023). The possibilities and challenges presented by artificial intelligence applied to digitalisation thus extends to and includes the cultural heritage institutions that hold the historical collections. Without scanning and digitalisation of the physical sources, there can be no sets of historical data to peruse in the process of training machine learning models.

Interdisciplinary collaboration is therefore a necessary component in any successful digitalisation project, ideally involving professional experts from all disciplines relevant to the entire digitalisation pipeline. Awareness of this need to crossing traditional and disciplinary boundaries is growing; the archival discipline, for instance, is currently facing major theoretical and practical changes due to digitalisation and the introduction of artificial intelligence (Colavizza et al. 2022), leading to calls for "a blend of computational

and archival thinking" (Marciano et al. 2018: 181), including participation of archival professionals in the creation of datasets for training, testing, and controlling the output of algorithms (Jaillant & Caputo 2022).

By drawing on practical experience from a specific real-world case that involves such interdisciplinary collaboration and researcher-driven digitalisation, this chapter provides hands-on advice and perspectives, more broadly applicable to the problem of digitalising printed historical documents and applying machine learning techniques. We describe a general comprehensive methodology and pipeline for digitalising printed historical materials using machine learning, including data acquisition (scanning of printed text). Most of this description is based on the authors' experience with the digital infrastructure project *Communicating Medicine: Digitalisation of Swedish Medical Periodicals, 1781–2011* (SweMPer) hosted by the Department of History of Science and Ideas at Uppsala University in collaboration with the Centre for Digital Humanities and Social Sciences (CDHU) and Uppsala University Library. SweMPer aims to build a searchable database of a wide range of Swedish medical periodicals created by medical professionals, patients, and medical students. The pipeline used in SweMPer utilises machine learning models for object detection to perform layout recognition and optical character recognition (OCR) to extract detailed metadata for increased searchability. At the end of the chapter, notebooks for specific parts of the pipeline are provided.

2 *Research context*

As digitised materials are increasingly being used by historical researchers for Natural Language Processing (NLP) and text mining tasks, the demands on digitalisation quality are rising. High quality digitalisation – including accurate character and word recognition, proper layout recognition, and correct reading order – is needed for successfully performing popular NLP tasks such as topic modelling, named entity recognition, and training word embeddings (Hill & Hengchen 2019, Neudecker et al. 2021, van Strien et al. 2020). Historical print often includes heterogeneous and complex layouts, as well as font- and spelling variations, causing standard layout detection and OCR software to fall short (Jarlbrink & Snickars 2017). Application of machine learning methods can offer significant advantages for digitalisation by improving quality, and it may also open up new research venues and searchability possibilities that are not available with more basic procedures.

With a digitalisation pipeline that produces rich and detailed metadata, materials can be filtered on a more granular level based on the type of

content. For example, in the case of newspapers and periodicals it could be valuable to separate main article text from other text-types that occur on the printed pages such as advertisements, tables, lists, and crossword puzzles. In addition to the textual layer, researchers may also want to explore visual elements – such as illustrations, photographs, diagrams, etc. – something that can be facilitated by machine learning-based computer vision tasks, such as image classification and clustering (Wevers & Smits 2020) and the automatic generation of image tags and descriptions (Thomas & Testini 2024, Villaespesa & Crider 2021).

Designing a customised digitisation pipeline, in comparison to off-the-shelf solutions, often provides greater flexibility for integrating and adapting components that are optimally tailored to the material of interest. Machine learning models for computer vision offer new opportunities for addressing the quality of the digitisation, in terms of the accuracy of the layout detection, optical character recognition (OCR), and metadata extraction (Thomas & Testini 2024). Achieving higher quality and better, enriched metadata through machine learning, however, requires annotating training data for fine-tuning layout recognition and OCR models. This is particularly the case with historical materials where pre-existing annotated data usually does not exist.

While there is great potential in the application of machine learning to digitalisation, ethical issues also need to be taken into account in relation to the annotation and training process. One well-known problem is the issue of bias in machine learning models. Here, we refer to bias in the broader, humanities-centered sense of misrepresentation in various forms and at various levels of Artificial intelligence (AI) usage, for example under- and misrepresenting certain groups of people such as women and minorities (Jaillant & Caputo 2022, Nicoletti & Bass 2023). Machine learning models are trained based on datasets that have been constructed by humans; therefore, they also pick up any implicit and explicit biases contained in the data (Ntoutsi et al. 2020). Due to institutionalised racism, sexism, and various cultural stereotypes, certain kinds of people may also be overrepresented in the data and subsequently also in the AI algorithm. As Jaillant & Caputo (2022: pp. 824-5) have put it: "for example, since White men are over-represented at the most senior levels of government and other sectors, documents by women or ethnic minorities could be flagged as less important". While using OCR when digitising historical documents is a technical exercise that does not in itself necessarily introduce or perpetuate bias, it can still be error-prone and it is important to be aware of the affordances of written language and limits of any particular technology; for example, Swedish or Spanish has certain letters that OCR models trained on American

English may not recognize. Choices made in subsequent efforts of "cleaning" and "sanitising" problematic OCR may, yet again, potentially introduce bias (Arroyo-Ramirez 2016). At later stages, during computational analysis and post-processing, further biases may also be introduced – for instance, textual bias may occur when producing word embeddings (Bolukbasi et al. 2016). Being aware of and addressing such problems is important to all phases of the construction and implementation of a digitalisation pipeline.

Another important aspect to take into consideration is the priorities made and selection of sources when digitalising. Digitalisation efforts are commonly driven by institutional actors and custodians of historical collections, i.e. libraries and archives. One recent example is a digitisation program initiated at the The National Library of Scotland (NLS) that is expected to provide digital versions of as much as a third of its collection by 2025 (Jaillant & Caputo 2022). Many larger scale and well-known projects have been concluded in the past decades, most prominently perhaps the Google Books Scanning project, primarily in collaboration with many of the major American university libraries[1]. In the Swedish context one impactful effort concerns the The Swedish Newspapers digitalisation project (Svenska dagstidningar[2]). This project, initiated and hosted by the National Library of Sweden, makes it possible to search and read digital historical newspapers from 1645 up to today, and has in many ways proven critical to research.

While there is undeniable value in such general, institutional initiatives, there may be a further benefit in researcher-driven efforts tailored to the needs of individual projects and specific research questions. This is important in order for researchers to get access to sources that reflect their particular research interests, making them less dependent on what is available and what is being prioritised by cultural heritage institutions. From the perspective of preservation, for instance, sources from the twentieth century might be considered too 'modern' to be prioritised for digitisation over rare objects, or not consulted enough by users to warrant inclusion. Furthermore, they might simply be excluded because of legal problems concerning copyright status. In the case of The Swedish Newspapers mentioned above, the database currently limits access to anything published after 1924 for copyright reasons (unless physically situated at the National Library or certain university and public libraries) – and there are no options for bulk downloading of data, required for further and more advanced text mining or computational analysis. In other cases, metadata might be lacking or the data could be organised in ways that make it less conducive to efficient research.

1 https://books.google.com; see also, https://www.hathitrust.org/
2 http://tidningar.kb.se/

Researcher-driven digitalisation allows for control over the selection of sources and produces high quality data and metadata adapted to specific aims, priorities and research questions. The SweMPer-project provides one example of this approach to digitalisation, but other projects can also be mentioned to further exemplify a growing trend of research- and researcher-driven digitalisation efforts. One such project is ActDisease[3], which aims to study the history of twentieth century patient organisations. The study includes patient organisations from four European countries (Sweden, United Kingdom, France and Germany). ActDisease incorporates digitalisation workflows customised to and driven directly by the project's aims. To meet the project's research goals, member newsletters and magazines have been digitalised, creating a corpus of sources unlikely to be digitised without any research initiative (Aangenendt et al. 2024). Another example is the research infrastructure project Swedish Riksdag 1867-2022: An Ecosystem of Linked Open Data (SWERIK)[4] which illustrates the need for high quality digitalisation and descriptive metadata. The Swedish parliamentary proceedings, motions, and reports have previously been digitised by the National Library of Sweden, but the quality has proven insufficient for many research demands. The aim of SWERIK is to enrich the Swedish parliamentary documents by correcting digitisation errors and providing linked metadata for parliamentary documents and members of parliament, opening up new ways for researchers to explore the materials both quantitatively and qualitatively.

In conclusion, digitalisation efforts initiated and driven directly by researchers and their perspectives are crucial to create conditions for high quality historical research. By interdisciplinary collaboration that involves not only humanities researchers, but also machine learning experts and professionals at relevant libraries and archives, digitalisation workflows that are highly customised and tailored to the researchers' needs can be achieved within the framework of a specific project. The general and project-specific digitalisation pipeline presented below can hopefully help facilitate future researcher-driven digitalisation projects.

3 *Description of methods and tools*

Training a machine learning (ML) model—such as one designed to assist in the digitization of historical materials—typically involves four key stages: data acquisition, annotation and processing, model training, and deployment. Each stage benefits from collaboration among archivists, librarians,

3 https://www.actdisease.org, see also chapter NN in this book
4 https://swerik-project.github.io/

researchers, and engineers, ensuring that issues are addressed, project quality is enhanced, and transitions between stages are seamless. In the context of digitization, and specifically for the case study presented in this chapter, the data acquisition stage focuses primarily on scanning. Accordingly, data acquisition will be addressed in Section 3.1, annotation and processing in Section 3.2, model training in Section 3.3, and deployment in Section 3.4.

3.1 *Scanning and data cleaning preparation*

The first step in developing a machine learning based digitalisation pipeline is acquiring the data, which in the context of this handbook chapter means acquiring digital copies of the materials of interest. This can be done in several ways depending on the availability of the data. This section focuses in particular on scanning, which is the methodology employed in the SweMPer-project and presented in section 4. The scanning of historical materials is best done by digitisation experts at libraries or archives. However, if you want to digitalise material on your own there are some general guidelines that are important to follow. These guidelines can also be helpful when discussing the details of digitisation with libraries and archives that are often necessary collaborating partners in such undertakings.

Before initiating the scanning process, it is crucial to establish a consistent and descriptive file-naming convention and folder structure that captures key metadata. This decision is usually made by the lead researcher, ideally in collaboration with the engineer who will process the data in subsequent steps. One example of such an organization is incorporating metadata such as year of publication, ascending and "zero-padded" file sequence numbering (i.e. 0001, 0002, 0003 ... 0999), and any other information that may be of relevance. This helps keep the digitised materials organised at the file system level. Alternatively, a unique identifier can be used for each publication or collection that will be digitised, which can later be replaced with a more descriptive name. Common practice here is to use underscore, dashes, or to capitalise the first letter of each section, as computers struggle handling files with blank spaces [5].

Next, the specifications of the digitisation need to be decided upon, often in collaboration with the digitisation experts at cultural heritage institutions. This includes deciding on a suitable output format (tiff, jpg, or pdf), required image resolution (300 ppi or higher for printed materials with small font), colour settings, and what should be captured on each image (single page, double page).

5 https://datamanagement.hms.harvard.edu/plan-design/file-naming-conventions

When performing the scanning, it is crucial to keep the page and the text as straight as possible. Try to avoid scanning pages warped or skewed, as this can negatively affect the recognition of layout and text elements in later stages of the pipeline. Another recommendation is to avoid adding additional 'noise' on the image, in the form of spots, marks, and lines that are not part of the original document. This 'noise' could later on be interpreted as characters or lead to misinterpreted characters in the OCR process. A regular quality control can help identify any issues. Lastly, having a high contrast makes it easier to segment between the background and page elements. There are post-processing methods to deskew the page, straighten text lines, and remove noise but in general the better the image quality is from the outset, the better the final outcome.

3.2 *Data annotation and preparation for training*

Once the scans are available in digital image format, following the steps outlined in section 3.1, the next step is data annotation. This process involves several key considerations, including selecting the elements to annotate, determining appropriate labels or names, choosing an annotation tool, and deciding on the output format. Ideally, these decisions are made collaboratively by the lead researcher, the engineer, and the annotators to ensure consistency, efficiency, and to identify and address any potential biases. In this section, we aim to clarify the following questions:

- What is data annotation in the context of image-based collection digitalisation? What does an annotation look like?
- What annotation format is best for your task?
- What are the best practices for data annotation?
- What are the best tools for easy, efficient, and secure data annotation?
- How do we prepare the data to feed it into our ML model?

3.2.1 *Data annotation and how it looks*

Before annotating any data, it is essential to understand what data annotation entails in the context of computer vision. *Data annotation* involves identifying and labelling elements of interest within images, videos, or other forms of visual media. During this process, the annotations are converted into tags and coordinates that a computer can interpret and utilise (OpenCV 2024). There are different forms of annotation in computer vision, including:

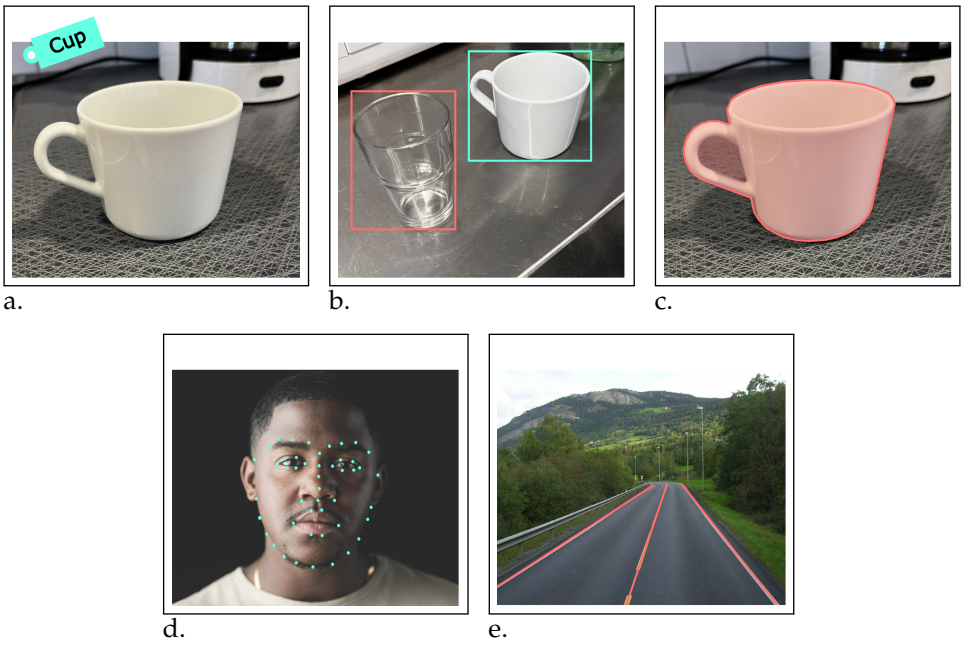


Figure 1: Types of data annotation in computer vision: (a) full image classification, (b) 2D bounding boxes, (c) segmentation, (d) key point and landmark, (e) lines and splines (image sourced from Małecki 2006).

- *Full image tagging*, such as labelling an image as "cup" or "glass" (Figure 1a), which is primarily used in classification tasks.
- *2D bounding boxes* (Figure 1b), which are rectangles used to identify and determine the location of objects, commonly utilised in object detection and localisation tasks.
- *Polygonal segmentation* (Figure 1c), which goes beyond rectangles by surrounding objects with more complex polygons to obtain a more precise shape and localisation.
- *Key point and landmark* (Figure 1d) involves drawing points on an object within an image, which is useful for pose detection and facial recognition.
- *Lines and splines* annotation (Figure 1e) is created by using lines to mark clear paths, such as roads, which is particularly useful in navigation tasks (OpenCV 2024, Xailient 2024).

Linking the types of data annotations to the use case presented in this

chapter, we will focus on 2D bounding box annotations. Bounding boxes are one of the most common types of annotation for object detection, their aim is to determine the location and boundaries of target objects by surrounding them with tight rectangles. The annotations provide metadata for each box in an image, such as the coordinates of the corners or centres, the widths and heights relative to the image, and the class of each box (Xailient 2024).

3.2.2 *Bounding boxes label formats*

As mentioned earlier, annotating data with 2D bounding boxes generates information that can be fed into a ML model to teach it to detect elements in an image. How these data are organised and stored varies based on the object detection ML model one plans to use and the data annotation tool employed. It is possible to convert between different formats, as will be demonstrated in the last section "Hands-on Practical Introduction".

COCO and *YOLO* are conventional annotation formats used by the most popular object detection ML models in computer vision. The choice between the two relies on the specific requirement of the ML model. The COCO format stores information for all images and all bounding boxes in a single *JSON* file. Each bounding box is formatted as $[x,\ y,\ width,\ height]$, where the tuple (x, y) corresponds to the coordinates of the upper-left corner, and *width* and *height* are the dimensions of the box from (x, y). The file also contains the IDs for each class and each image (Medium 2024). Figure 2 (left side) illustrates an example of this. The single JSON file associates an integer ID with each category, image, and box per image. The details of each box are stored in the "annotations" section with a *bbox* entry per bounding box. In the *bbox* entry, the values 0 and 1 correspond to x and y, respectively, and 2 and 3 correspond to the width and height. In contrast, YOLO uses *TXT* files, generating one file per image. If an image has no target objects, a TXT file is not needed. Each TXT file contains one row per annotated object, formatted as: *class x y width height* (see Figure 2 right side bottom rectangle), where (x, y) correspond to the center of the box (Ultralytics 2023), as opposed to the COCO format, where (x, y) correspond to the upper-left corner.

Two important differences to highlight are image re-scaling and the way information is stored. In the COCO format, the image size information is not re-scaled, so the size of the bounding box matches the image's scale. This means that if the image has dimensions (W, H), then the box widths and heights will fall within the intervals $[0, W]$ and $[0, H]$, respectively. In contrast, the YOLO format re-scales the image, resulting in width and height values that are always between the interval $[0, 1]$ (Medium 2024,

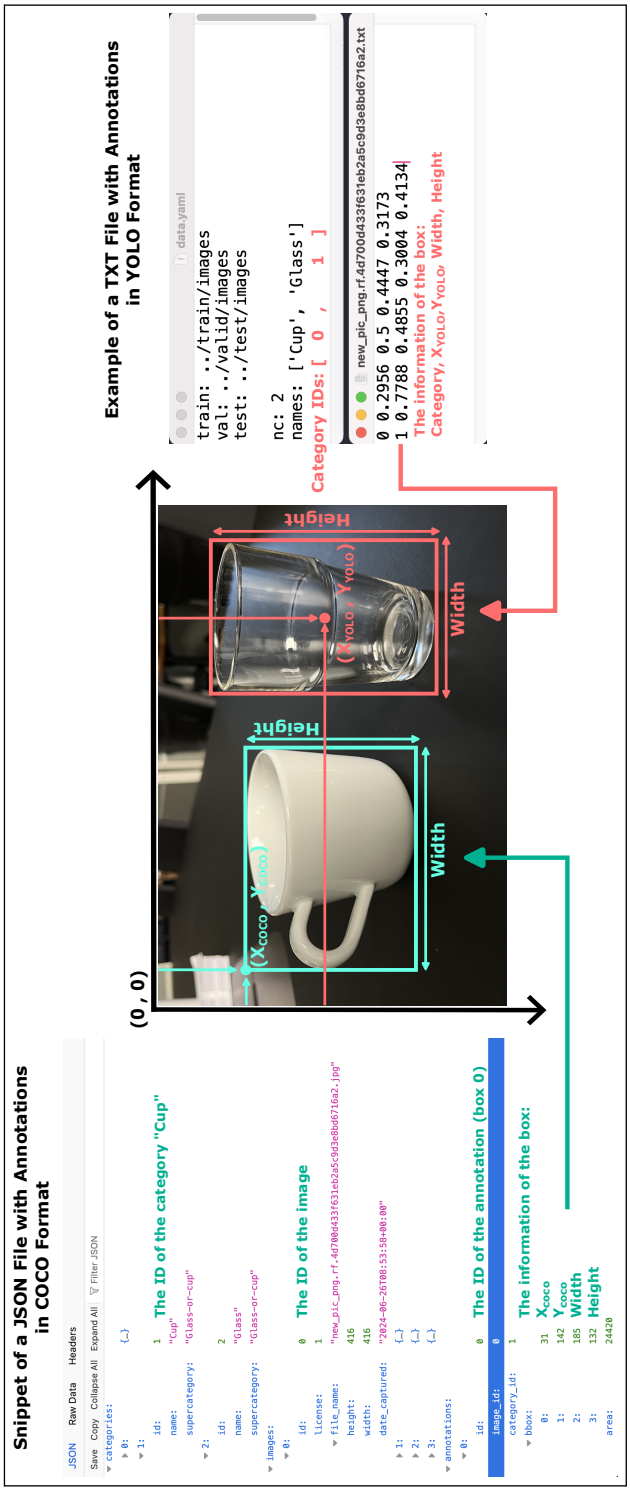


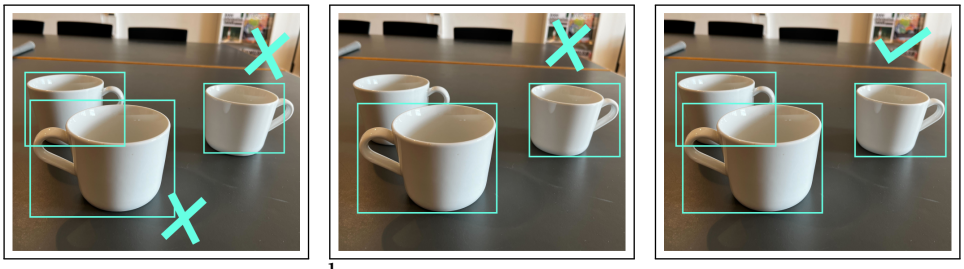
Figure 2: Comparison of COCO and YOLO annotation formats. The left side shows a COCO JSON snippet with bounding boxes in image scale. The right side shows a YOLO TXT snippet with normalised bounding boxes.

Ultralytics 2023). This re-scaling process involves normalising the bounding box coordinates by the dimensions of the image, so the box coordinates are relative to the image size. For instance, if an image has a width of 500 pixels and a bounding box is 250 pixels wide, the YOLO format will store this width as 0.5. Regarding file organisation, the COCO format stores all the information for a set of images in a single file (such as the one in Figure 2). Conversely, the YOLO format creates one file for each image, along with a *YAML* file (Figure 2, top rectangle on the right side) that contains links to the folders for each subset of images (we discuss these subsets in more detail in Section 3.2.5).

3.2.3 *Best practices for data annotation*

While various best practices for data annotation have been proposed, there is no universal consensus. The approach to annotating data largely depends on the specific task at hand. In some instances, it is essential to annotate every object within the category of interest in an image, whereas in other cases, this level of detail may not be necessary. Regardless of the annotation process, consistency and clear guidelines are crucial, especially when collaborating with others or using crowd-sourced labelling. Consistency in naming conventions is particularly important to prevent future confusion (Labellerr 2023). For instance, in the cup-or-glass scenario, it is essential to use ”cup” uniformly rather than allowing variations such as ”mug” or ”Cup.” Such variations could lead to inconsistencies, as the label for the same object would shift from one—”cup”—to three—”Cup,” ”mug,” and ”cup.” Inconsistent labels could confuse a ML model when trying to distinguish between objects in these classes. Similarly, it is crucial to agree on the number of classes of interest. For example, if one is annotating images of pets, one should decide whether to classify them into just two categories—dogs and cats—or to include additional classes, such as specific breeds (Nanonets n.d., Roboflow 2024b).

Tight bounding boxes are another key practice in data annotation. It is recommended to draw boxes that tightly encompass the entirety of the object without cutting off any portions (Nanonets n.d., Roboflow 2024b). In Figure 3a, an example is shown where the front most cup is annotated with a box that is too large, which is not recommended, sub-figure (c) shows a box that fits the cup correctly. Related to this, it is advisable to ensure that the entire object is covered without missing any borders. This practice is crucial for helping the ML model accurately learn the contours of the object (Roboflow 2024b). Figure 3a shows a box that is too small to fit the object (rightmost cup), which goes against best practices, and a correct annotation



a. b. c.

Figure 3: Examples of the implementation of good and bad practices in data annotation: (a) This is an incorrect annotation because two of the boxes do not properly fit the objects. (b) This is incorrect because one of the objects is not annotated. (c) This image is correctly annotated.

is shown in sub-figure (c). It has also been noted to be beneficial to label occluded objects if they are identifiable, even if this results in overlapping boxes. Additionally, these objects should be labelled as if they were in full view (Roboflow 2024b). In Figure 3c, we can see that the cup at the left-back is annotated as it was in full view.

Finally, it is recommended to label every instance of the object in each image. Failing to do so can introduce false negative predictions, i.e. incorrect or no detection when in fact there was an object of interest, which would negatively impact the ML model's performance (Nanonets n.d., Roboflow 2024b). Figure 3b shows an example of an incorrect annotation where the left-back cup is not surrounded by a rectangle, which goes against the recommendations. However, in some situations, it might not be necessary to recognize all instances of an object, even within the same class. For instance, when developing a ML model specifically for close-up stop signs, it might be valuable to consider whether to annotate distant stop signs. Annotating every stop sign would train the ML model to recognize all stop signs, regardless of their distance. On the other hand, choosing not to annotate distant objects could emphasize the importance of proximity, though this approach may have its own drawbacks such as introducing false negatives, as mentioned previously.

3.2.4 *Tools for data annotation*

There are several tools for data annotation. These range from cloud based services to locally hosted ones and paid or open source. In this section, we will mention some of the tools that we have used and found accessible.

```
{
  "version": "5.4.1",
  "flags": {},
  "shapes": [
    {
      "label": "cup",
      "points": [
        [
          368.0952380952381,
          178.88888888888889
        ],
        [
          779.2063492063493,
          675.7142857142857
        ]
      ],
      "group_id": null,
      "description": "",
      "shape_type": "rectangle",
      "flags": {},
      "mask": null
    }
  ],
  "imagePath": "../photo_2023-07-06 14.58.47.jpeg",
  "imageData": "/9j/4AAQSkZJRgABAQAAAQABAAD/
```


Figure 4: Example of the LabelMe annotation format for an image containing a single rectangle.

LabelMe

In 2007, MIT released a dataset alongside a data annotation tool called *LabelMe*. This tool was designed to be web-based, allowing a large number of users to access and annotate the material. The emphasis was on simplicity and ease of use, with the added benefit of being open-source (Russell et al. 2008). Although access to the original web-based platform has since been closed, as it was specifically developed for annotating their dataset, the popularity of LabelMe led to the creation of a new open-source version. This updated version can be installed on most common operating systems and provides an interface for image annotation using polygons, rectangles, circles, lines, points, line strips, and more[6].

Annotations generated with this tool are typically stored on the same computer where the tool is run and follow a specific annotation format. LabelMe produces a JSON file for each image, which includes key information for each shape, such as labels and points (for example, the top-left and bottom-right corners of a rectangle, see Figure 4). Thus, when using LabelMe for annotation, a set of JSON files —one per image— with annotations in the LabelMe format will be generated. Then, converting this format into one that better suits the requirements of the ML model intended to be used

6 https://github.com/wkentaro/labelme

may be needed. An example of how to perform this conversion is provided the Hands-on Practical Introduction section.

The biggest advantage we have found with this tool are that it is open-source and hosted on a local server, providing a secure way to annotate sensitive data. However, one disadvantage is that it is a very simple and lightweight tool, limited to certain types of image data annotation and not suitable for other data types. Moreover, this tool does not facilitate easy collaboration, and it is difficult to track what has been annotated and what has not; thus, it is recommended for small datasets where one main person is performing the annotation. Finally, while this tool is simple, its interface may not be the most user-friendly for all users.

Roboflow

Roboflow[7] is another image annotation tool primarily focused on computer vision. It provides a platform where users can easily train ML models, although with limited customisation options. Further, it can be used for tasks such as segmentation, image captioning, and keypoint detection, among others. It offers a web-based free version, as well as premium versions that include features like access to premium GPUs for training, additional credits for ML model training and inference, and, most importantly, private datasets and trained ML models. This tool supports working with images in the most common formats, as well as videos. It also allows users to upload annotations generated with other tools. Roboflow offers various output formats, which can be further specified according to the ML model version, such as YOLO and COCO.

Roboflow has two main advantages. First, its simplicity and user-friendly interface make it easy to navigate, which is especially suitable for beginners in data annotation. Second, the ability to train ML models on the same platform is valuable for testing purposes. The main disadvantage of Roboflow is that, in the free version, datasets and ML models created on the platform are open to the community, making it unsuitable for sensitive data or data with copyright concerns.

Label Studio

Label Studio is a versatile data labelling and annotation tool that supports multiple projects and users. It offers an open-source version that can be hosted on a local computer, as well as premium versions that provide additional

7 https://roboflow.com/

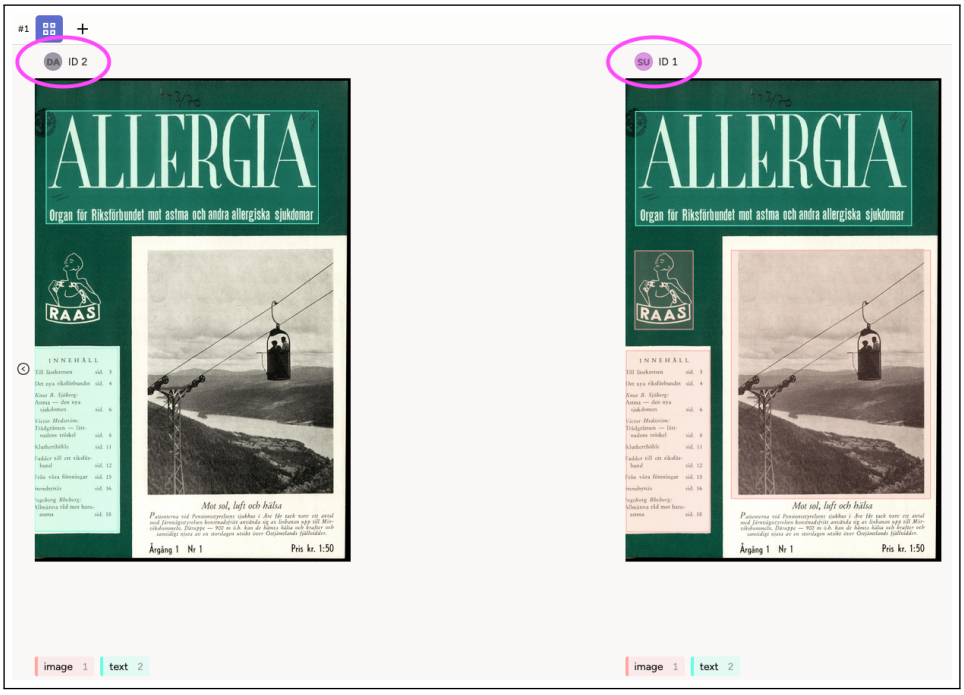


Figure 5: Example of collaborative annotation in Label Studio. The pink ovals highlight two different collaborators who have annotated separate copies of the same image, producing distinct annotation versions for comparison.

features such as cloud-based usage and data storage, user management, activity logs, and AI-assisted labelling, among other capabilities.[8] Label Studio focuses on three main areas: simplicity, configuration, and integration. The tool features a user-friendly interface that is easy to navigate and set up. Additionally, users can configure it to suit their specific annotation needs, styles, and data types. This flexibility is made possible by its multi-type data capabilities, which allow users to annotate, for example, both images and text simultaneously. Finally, Label Studio can be integrated with a variety of ML frameworks and models (Medium 2020).

Based on our experience, some key advantages of this tool include: (1) a highly capable open-source option that allows multiple users to collaborate and annotate the same dataset. This can be done in two ways: either in parallel, where users add annotations to the same file without seeing who created each annotation, or by annotating separate copies of the same file, generating multiple versions of the annotation. The latter approach, for

8 https://docs.humansignal.com/guide/get_started

example, enables comparison to detect discrepancies in labels (see Figure 5); (2) the ability to host the tool locally, ensuring that both the service and data remain on a local computer, shared only through a secure protocol; and (3) support for annotating data in various formats beyond traditional ones, enabling tasks such as visual question answering, image captioning, text summarisation, speaker segmentation, response generation, and more—all within a simple and user-friendly interface. Regarding output formats, Label Studio allows users to export annotations in different formats, including COCO, and YOLO.

3.2.5 *Getting the data ready for training*

Proper data organisation and processing are essential, though sometimes cumbersome, steps in the machine learning pipeline. This process typically involves modifying the dataset and splitting it into three subsets: training, validation, and testing. In some cases, only two splits—training and testing—are used. The choice of splits and data processing if any is ultimately up to the engineer responsible for training the machine learning model.

The *training dataset*, which contains the largest number of images among the three subsets, is used by the ML model to learn features and patterns. This dataset plays a pivotal role in the training process as it provides the raw data from which the ML model extracts and learns to recognise various features (Goodfellow 2016: 129). To enhance the effectiveness of this learning process, various transformations, also know as *data augmentation*, can be applied to the training dataset, including operations such as rotation, scaling, cropping, and colour adjustments (Mikołajczyk & Grochowski 2018, Shorten & Khoshgoftaar 2019). By introducing such variations, the dataset is artificially increased in size and diversity, helping the ML model to improve its ability to generalise to new, unseen data, and preventing it from learning the data too wall (this phenomenon is also known as *overfitting*). This approach ensures that the ML model does not become overly specialised to the specific examples it was trained on but rather learns to recognise patterns and features that are more broadly applicable (Obeso et al. 2017, Bahrami & Albadvi 2024). Additionally, these transformations can be used to ensure other important aspects of the training data, such as balance and representativeness. Data augmentation can expand certain classes, potentially preventing biased results (Pastaltzidis et al. 2022, Zhang et al. 2022). The *validation dataset* is used to evaluate the ML model's performance during training (Goodfellow 2016: 129). Finally, the *test dataset* is employed to evaluate the fully trained ML model "in the wild" using images the ML model has never encountered before (Google n.d., Xie et al. 2011), which is a crucial step for assessing how

the ML model will perform on new, unseen data. Usually, the ML model is evaluated on the validation data several times before it is evaluated on the test data.

As previously discussed, the choice of a ML model, which is intrinsically linked to the annotation format, along with the tool employed for data annotation, plays a critical role in how data must be organised for training. The choice of ML model is often dictated by the specific requirements of the task, which in turn determines the appropriate annotation format and the necessary tools for data preparation. For instance, the YOLO family of ML models (see Section 3.3.2 for more details) requires the YOLO annotation format, a specific file structure, and a .YAML configuration file. To prepare data for input into a YOLO ML model, it needs to be divided into three directories—train, validation, and test—each containing two subdirectories: images and labels. A configuration file is also required to specify the locations of these subdirectories, as well as the number and names of the classes (see Figure 2). Another example is the Layout Parser family of ML models (see Section 3.3.2 for more details), which utilises the COCO annotation format and a specific file structure during the training process. In this case, the data must be organised into three directories—train, validation, and test—with each containing a JSON file that includes annotations in COCO format.

The process of preparing and organising data for training can be cumbersome, which underscores the importance of the data annotation tool. In Roboflow, for example, a dataset can be exported in a format that is ready for training. This includes automatically dividing the data into the required file structure for popular ML model families and generating any necessary configuration files. In contrast, tools like Label Studio and Labelme do not provide automated file organization capabilities, requiring this step to be performed manually or through alternative methods. Manual data organization can be error-prone and time-consuming, requiring careful attention to file structure and format to ensure compatibility with the model. This can introduce additional complexity, especially for large datasets or when working with multiple ML models.

When discussing data splitting, creating the training, validation, and testing subsets is a process in itself. The method described in this section for splitting the data is known as *hold-out cross-validation* and has become popular due to its simplicity and effectiveness (Reitermanova 2010). Further, the specific ratio in which to split the data can heavily affect a ML model's performance. For this reason, alternative methods have been proposed in the literature that depend on the type and quantity of data at hand (Reitermanova 2010, Xu & Goodacre 2018).

In conclusion, the process of data preparation, including careful consid-

eration of splitting, annotation formats, and tools, is vital for successful ML model training. Choosing the right methods and tools as well as documenting these choices can streamline this process, reduce potential errors, and improve the overall performance of the ML model.

3.3 *Model training*

3.3.1 *What "training a model" really means*

In the preceding sections, the concept of training a machine learning model has been mentioned multiple times. It is therefore pertinent to examine what this process entails. However, before embarking on this, it is relevant to consider two related concepts: *mathematical model* and *machine learning model*. A mathematical model is an abstract mathematical representation of a system's dynamics (Bender 2000). Such models can be derived using differential equations, graphs theory, boolean relations, among other methods (Bunge & Bunge 1973, Pavlopoulos et al. 2011, Wang et al. 2012). In contrast, a machine learning model is a type of mathematical model that integrates this mathematical framework with an algorithm designed to identify patterns in a training dataset. These patterns are then encoded into the mathematical model's parameters. In the upcoming sections, the terms model and machine learning model will be used indistinguishable to refer to machine learning models.

Training a machine learning model involves feeding it a training dataset, performing computations, and adjusting the model's parameters to better fit the data. During this training process, a validation dataset is used to assess the model's performance (Goodfellow 2016: 110). This iterative process continues until a predefined stopping criterion is met. There are two main criteria for stopping model training: a fixed number of iterations, known as *epochs*, or the use of *early stopping*. In the first approach, a predetermined number of epochs, such as 300, is specified. This means that the process of learning from the entire dataset, adjusting the parameters, and evaluating on the validation dataset is repeated for, for example, 300 times. While this method is straightforward, it is not highly effective, as the model may overfit to the training dataset. To address this issue, an alternative approach known as early stopping has been proposed. In early stopping, the learning process is terminated when the model's parameters show no improvement, based on the validation metric, for a pre-specified number of iterations (Goodfellow 2016: 238).

A model can be trained in two ways: *from scratch* or using *transfer learning*. A model is said to be trained from scratch when its parameters, before learn-

ing anything from the training data, are initialised randomly. In contrast, transfer learning involves using a pre-trained network that has been fully trained on a different dataset, similar to the target dataset, and task. The parameters learned from this initial task can then be repurposed or transferred to a second network to be trained on the target dataset and task (Yosinski et al. 2014). In this context, all the parameters learned from the initial task can be used to initialise the new machine learning model, or only a subset of them may be utilised.

3.3.2 *Common machine learning models and tools in document digitalisation*

In document digitalisation, the choice of machine learning models varies depending on the task, and the assesment of which model is best is primaritly done by the engineer. In the context of digitalising large-scale printed collections, object detection, together with layout analysis, and optical character recognition (tasks in the realm of computer vision) are often involved (Aangenendt et al. 2024, Gruber et al. 2020, Matschak et al. 2022). Moreover, these models can be used in a chain to obtain and analyse different components in an image. In section 4.1, an example of how such models are used in a digitalisation project is presented.

Object detection models

Object detection is fundamental to understanding the structure and layout of a document. Machine learning models for object detection can be trained to locate and classify specific elements within a page, such as text blocks, tables, and figures. These models can be further refined for more specific tasks, such as distinguishing between photographs and advertisements or separating body text from titles. Currently, several pre-trained machine learning models for object detection are openly available, each offering varying levels of performance depending on the task. These models do not need to be trained from scratch, as transfer learning can be applied. Furthermore, these convolutional neural network-based models can be modified and combined with others to enhance functionality.

According to the HuggingFace leaderboard, some of the best-performing models or model families are (Rafael Padilla & the Hugging Face Team 2023):

- The YOLO Family: This family of models, developed by Redmon et al. (2016), is designed for real-time object detection and include a series of models from YOLOv2 to YOLOv11. These models have been proven to be both fast and accurate (Kaur & Singh 2023) and have been applied

in various tasks, such as automatic person detection in thermal images (Krišto et al. 2020) and layout analysis (Naiman et al. 2023).

- Deformable DETR (Deformable detection transformer): DETR, introduced by Zhu et al. (2020), is a recent object detection model that eliminates the need for anchor boxes or points (Zou et al. 2023). As a transformer-based model, it requires different data annotation and training processes compared to traditional object detection models. Deformable DETR builds on DETR, offering improvements in speed and performance, particularly for detecting small objects (Zhu et al. 2020).

In relation to the content of this chapter, as mentioned earlier, object detection models can also be employed to learn page layouts. For instance, by training models on labeled page layouts, the system can infer and categorise different structural elements across various document types. This is particularly useful when working with historical documents, which often exhibit complex and inconsistent layouts.

Layout analysis toolkits

Layout ML models are a specialized subset of object detection models that focus on learning the arrangement of objects on a page. One such toolkit is Layout Parser, which simplifies the application of layout detection models. It includes a model zoo with pre-trained models, ssuch as frcnn and mask-rcnn networks trained on PubLayNet (Zhong et al. 2019), a modern scientific documents dataset and newspaper navigator (Lee et al. 2020), a U.S. newspaper dataset from the 20th century, among others (Shen et al. 2021).

Optical character recognition toolkits

Another crucial aspect of document digitalisation is converting scanned images of text into machine-readable form, which is achieved through optical character recognition (OCR) models. Some of the most popular OCR models include:

- EasyOCR: It is known for providing cost-efficient results while maintaining a competitive accuracy (Roboflow 2024a). EasyOCR is a lightweight and user-friendly OCR model that supports over 80 languages. It excels in handling complex documents and works well with various fonts

and handwriting, making it a versatile tool for a wide range of OCR tasks[9].

- Tesseract: Developed by Google, Tesseract is an open-source OCR engine extensively used for converting printed text into digital form. It supports up to 127 languages, depending on the version, and can process various text formats (e.g., single or multiple columns) and fonts. A list of supported fonts is available in the tool's documentation[10]). Tesseract is particularly popular for its integration with different tools and its ability to be customized for specific tasks (Smith 2007).

3.4 *Model deployment*

Once a model (or a series of models) is fully trained, the next step to make predictions or extract results on the complete dataset; this process is also called *inference*. This step involves extracting metadata from the objects of interest and text from digital images, if an OCR tool was utilised. The results can be stores in files, such as one file per image processed, for future export to a database. The output files may contain different attributes depending on the model's objective. For instance, a metadata file might include information about the locations of titles, images, subtitles, and other elements of a digital image of a newspaper page. These files are typically stored in formats like JSON or XML, which allow for easy integration with databases and other systems. Thus, these files can be stored in a database, which, in turn, can be connected to a web portal with a searchable system [11].

4 *A use case scenario*

4.1 *Introduction to SweMPer*

The Swedish Medical Periodicals (SweMPer) project aims to digitize, curate, and make accessible a vast collection of printed Swedish medical journals spanning over two centuries. These periodicals, which have been crucial in documenting medical discourse from the 18th century to the present day, provide an invaluable resource for historians, researchers, and the general public. By the end of the project around 550 000 pages (28 titles, ca 1100 physical volumes) will have been scanned, digitalised and collected in a

9 https://github.com/JaidedAI/EasyOCR
10 https://github.com/tesseract-ocr/tesseract/blob/4.1/src/training/language-specific.sh
11 See for example https://books.google.com; see also, https://www.hathitrust.org/.

database. This data will be made available through a web interface (The Swedish Medical History Portal), also under development by the project. However, the complexity of these historical documents ranging from varied layouts and formats to different languages and degraded print quality poses significant challenges for digital processing and analysis. While similar difficulties exist in other large-scale digitization efforts, historical medical periodicals present additional hurdles due to their specialized terminology, evolving typographic conventions, and domain-specific elements, making automated processing more challenging. To address these obstacles, the SweMPer project employs advanced machine learning and computer vision techniques, with a particular focus on object detection, and optical character recognition (OCR).

The primary objective of SweMPer is to create a comprehensive digital archive of these journals, enabling researchers to access and study them in a structured and searchable format. To allow for long-term historical perspectives, the SweMPer-database includes medical journals published in Sweden in the period 1781–2011. The starting point is the year when the earliest periodical was first published; the end point marks the year when the availability of existing digital editions makes further digitisation redundant. During this long era of emerging and changing media and technologies, including print, photography and digital media, scientific journals became an inextricable part of medical culture (Podolsky et al. 2012) and evolved into an extensive knowledge infrastructure that has enabled and redistributed medical authority, influence and power across different media.

In Sweden and elsewhere the earliest collegial publications, produced by and for doctors, were soon followed by commercial periodicals more broadly aimed at communicating medical and pharmaceutical knowledge to the public, nationally as well as locally. By the 1930s self-governed patient organisations began publishing independent journals, enabling patients and laymen to collect their own data and articulate knowledge claims, while making demands on the professional agents of medical care. Medical students also started to issue journals that from the 1960s increasingly participated in the political turmoil of the day.

Thus, to ensure broad representativity, the full SweMPer-database includes the major Swedish journals produced by doctors, public health and health care professionals, by student and patient organisations, as well as publishers in the commercial market. This enables study of themes ranging from professional debates to popular discourse and the experiences of patients.

Crucially, the journals included in SweMPer are heterogeneous in terms of content and genres, including editorials, professional debates, readers'

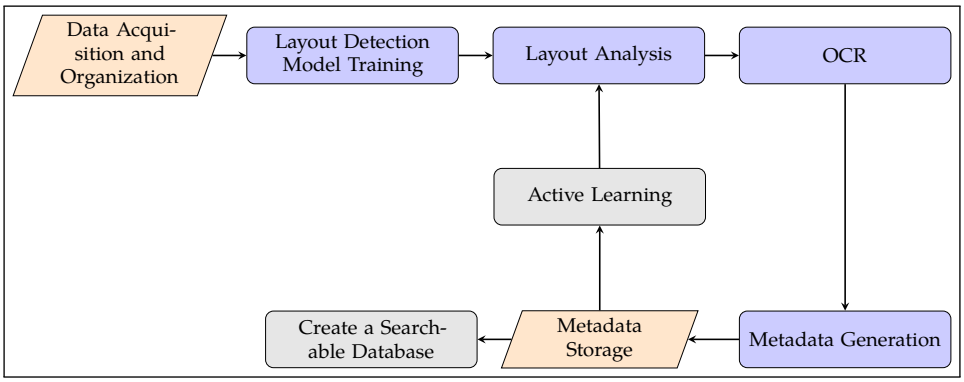


Figure 6: SweMPer pipeline overview

letters and advertisements for drugs and treatments. They also contain large amounts of non-textual media such as photographs, tables, x-ray images and illustrations, mirroring the fact that medical history is as richly visual as it is textual. While this variety is challenging in terms of automated metadata extraction, it also means that the finished SweMPer-database will support not only rich textual but also visual analyses of medical history.

By automating the extraction of textual and visual content from these documents, SweMPer not only preserves important historical records but also enhances their accessibility and usability for a wider audience. This use case scenario outlines the development, implementation, and outcomes of the current SweMPer digitisation pipeline.

4.2 *SweMPer pipeline overview*

The SweMPer pipeline is designed to handle the entire process of digitizing historical documents, from image acquisition to the creation of a searchable platform to access the images and metadata generated in the SweMPer project. The pipeline begins with the collection of high-resolution images of the journals, which are then processed using machine learning models for layout detection and OCR tools, and create a searchable platform. Consequently, the extracted data is organised and stored, ensuring that each document is easily retrievable and its content is searchable.

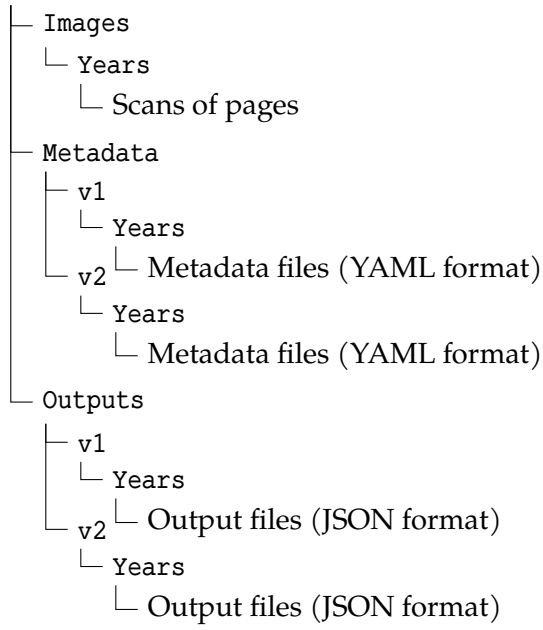
A visual abstract of the pipeline in Figure 6 illustrates its key components and the flow of data through the system. The components in gray will be implemented in the future. A detailed explanation of each part in the pipeline is presented in the following sections.

4.3 *Explanation of pipeline components*

4.3.1 *Data acquisition and organization*

The first step in the SweMPer pipeline is the acquisition of images from the historical journals. The data for this step was obtained by scanning physical copies of journals in-house at the Center for Digital Humanities and Social Sciences at Uppsala University, and from the Uppsala University Library. The scans are stored in .jpg format and organized into a structured folder hierarchy to facilitate efficient processing in later stages.

Before continuing with any other processing, a plan for organising the data, intermediate and final results was also necessary. A thorough organisation not only simplifies the storage and retrieval of images, metadata, and output files, but also plays a critical role in the metadata extraction process in later stages. By embedding key information into the filenames and organising the scans and associated files in a structured manner, it is ensured that metadata and layout data can be accurately and efficiently extracted and used during subsequent processing. This level of detail is particularly important when working with large datasets where manual intervention would be impractical. In this manner, the folder structure was designed as follows:



In this structure:

- `Images` directory: Contains scans of historical journal pages, organized by year, with each scan renamed to include essential metadata such as the journal name, year, volume, and scan number. For example:

  `journalname_year_numberofthescan.jpg`

  Additional information such as the volume and issue numbers may be included depending on availability, for example:

  `status_1947_vol010_nr006_0017.jpg`

- `Metadata` directory: This directory holds metadata files generated during the layout analysis and processing stages. Each version of the metadata extraction is stored in separate folders (e.g., `v1`, `v2`), representing different iterations or improvements in the metadata extraction process. The files are organized by year and named according to the corresponding image, with the same naming convention as the images but in YAML format:

  `status_1947_vol010_nr006_0017.yaml`

- `Outputs` directory: This directory stores the results of the layout detection inference, with the output files saved in JSON format. As with the metadata directory, the output files are organized by version (e.g., v1, v2) and year, with each file representing the layout detected in the corresponding image:

  `status_1947_vol010_nr006_0017_layout.json`

Each scan and its associated metadata and output files are named consistently, allowing for easy cross-referencing between the image files, the extracted metadata, and the layout detection results. For example:

- A typical file in the `Images` directory might be named: `allergia_1957_0054.jpg`.
- If the volume and issue numbers are available, the filename might be: `status_1947_vol010_nr006_0017.jpg`.
- The corresponding metadata file: `status_1947_vol010_nr006_0017.yaml`.
- The corresponding output file: `status_1947_vol010_nr006_0017_layout.json`.

4.3.2 *Layout detection model training*

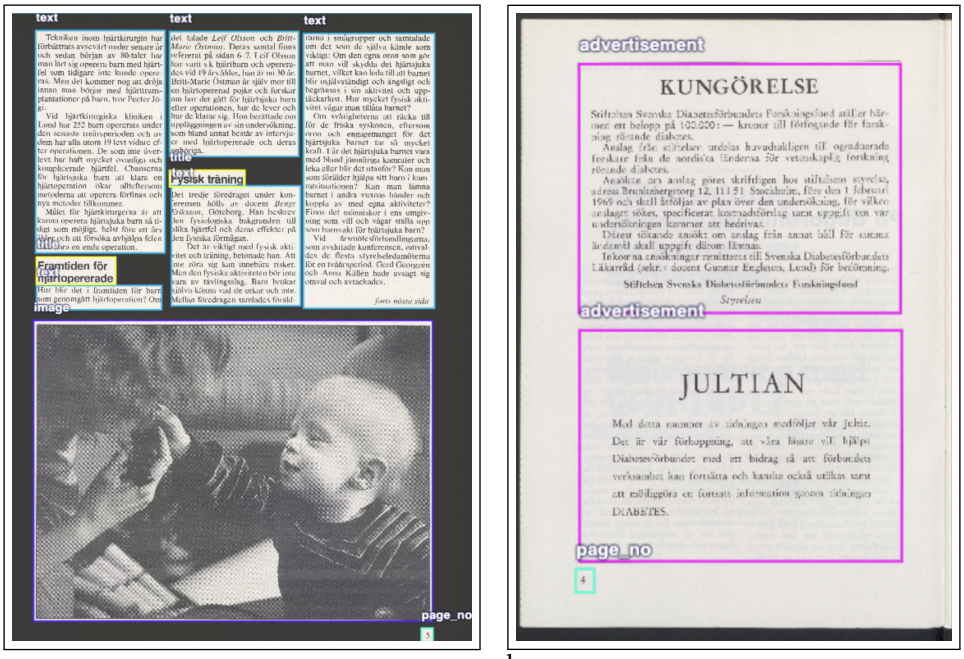
After obtaining and organising the scans of each journal, the next step in the SweMPer pipeline involves detecting the layout of historical journal pages. Detecting layout elements such as authors, tables, lists, and titles is crucial because it allows us to create a structured, searchable database from the digitized journals. This structure enables users to easily locate specific sections, such as an article's title or a table of medical data, thus improving both the research experience and the accessibility of the documents. To do this, we first selected a subset of data for manual data annotation, followed by model training using the annotated data.

Data Selection and Annotation: For data selection, at first, we randomly selected an equal number of images from all the available journals expecting to obtain a balanced dataset and store the details like filenames, which journal they come from, etc., to track which images are selected for training. This helps us to avoid selecting duplicate images for training. At this step, we had a total of 3351 images, of which 2701 images were used for training and 650 images were reserved for testing. The chosen images were manually annotated using the Roboflow platform. This annotation provided bounding box information for each layout element. After the annotation process, we can easily export the annotations in the required format for model training. Some of the examples can be seen in Figure 7

Moreover, we chose to have the following classes which will allow us to maximise the metadata collection. These classes capture a wide range of elements commonly found in historical documents, maximizing the usability of the metadata for future analysis and search functionality:

- advertisement
- author
- header_or_footer
- image
- list
- page_no
- table
- text
- title

During the initial phase of annotation and data selection, we observed a significant underrepresentation of certain layout classes, such as tables and lists. This means that these types of layout elements appeared much less frequently compared to other classes, such as text or images, in the annotated

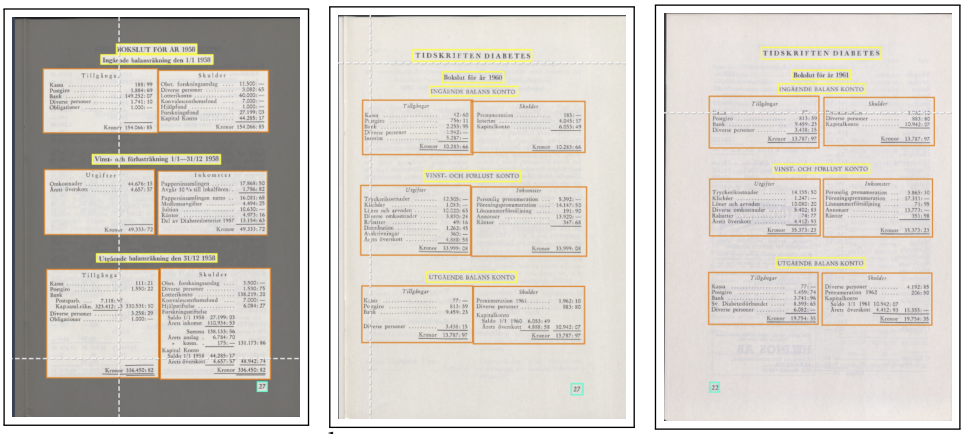


a. b.

Figure 7: Examples of annotations for the model training process.

data. This underrepresentation posed a challenge for training a robust model, as having too few examples of certain classes can limit the model's ability to learn and accurately predict those categories. As a result, we needed to take additional steps to balance the dataset to ensure that the model could learn to accurately detect these underrepresented elements. To overcome this imbalance, we employed a emperical approach using image feature vectors, which are numerical representations of the essential features of an image. These feature vectors encapsulate key visual information, such as shapes, textures, and patterns. By leveraging these feature vectors, we were able to identify images in the dataset that were visually similar to those containing the underrepresented classes. This allowed us to increase the presence of these underrepresented classes in our training data and improve the model's performance.

Using ResNet18 for Feature Extraction: To efficiently select more images with similar layouts (containing tables and lists), we used a ResNet18 (He et al. 2016) model as a feature extractor. ResNet18 is a pre-trained convolutional neural network (CNN) that is widely used for feature extraction due to its efficient architecture. The model generates a 512-dimensional vector (a



a. b. c.

Figure 8: Examples of selected images from the SweMPer dataset, which contain the underrepresented class of tables/lists. After calculating cosine similarity between the annotated images with underrepresented classes and the rest of the dataset, visually similar images were selected and added to the training set. These images helped increase the representation of tables in the training data and improving the model's ability to generalize across different layouts. Images (b) and (c) are examples of similar images retrieved using cosine similarity with (a).

feature vector) for each image, which essentially acts as a compact representation of that image's visual content. The 512-dimensional vectors that ResNet18 produces are rich enough to capture meaningful features of the images without being computationally expensive. These feature vectors can then be used to measure similarity between images.

Cosine Similarity for Similarity Search: After obtaining the 512-dimensional feature vectors for all images in the dataset, we needed a way to find images that were visually similar to those containing underrepresented classes (tables and lists). Examples of these can be seen in Figure 8. To do this, we used cosine similarity, which measures the cosine of the angle between two vectors in a high-dimensional space. Cosine similarity is a widely-used metric for comparing vectors because two images with similar visual content will have vectors pointing in similar directions, resulting in a high cosine similarity score (closer to 1).

Finding Similar Images to Expand Training Data: We focused on the images in the annotated set that contained the underrepresented classes. For each of these images, we extracted their 512-dimensional feature vectors using ResNet18. Then, we computed the cosine similarity between the selected

images (with underrepresented classes) and all other images in the dataset after which, for each image, we sorted all the similarity scores and selected the top 3 most similar images (excluding the image itself) based on this score. This process allowed us to find additional images that were visually similar to those containing the underrepresented classes. These extra images were then added to the training set, boosting the representation of those classes. By adding these similar images to our dataset, we were able to increase the total number of images in the dataset from 3351 to 3506 (2805 for training). This enriched dataset was more balanced, ensuring that the model would be exposed to a wider variety of tables and lists during training.

Selecting the Layout Detection Model: Once the data has been annotated and prepared, the next step in the SweMPer pipeline is training the machine learning model for layout detection. This process requires careful selection of the model, training data, and hyperparameters used during training. For layout detection, we experimented with two different models: the LayoutParser (Shen et al. 2021)'s Mask R-CNN model (He et al. 2017) (pre-trained on the PubLayNet dataset (Zhong et al. 2019)) and YOLO-NAS (AI 2023) (a more recent object detection model). The goal was to identify a model that not only provided high accuracy but also performed well on the specific task of detecting various layout elements in historical documents.

LayoutParser's Mask R-CNN: This model uses the Mask R-CNN architecture, which is well-regarded for its ability to perform both object detection (bounding boxes) and instance segmentation (pixel-level masks). LayoutParser provides pre-trained models specifically tailored for document analysis tasks, with the PubLayNet dataset serving as its training base. PubLayNet consists of over 360,000 document pages, making it highly suitable for tasks like detecting layout elements in historical journals. Mask R-CNN requires annotations in the COCO format.

YOLO-NAS: YOLO-NAS is a more recent advancement in the YOLO family of models, known for its speed, accuracy, and architectural improvements that leverage neural architecture search (NAS). YOLO-NAS is optimized for object detection and provides state-of-the-art performance in terms of both speed and accuracy. YOLO-NAS excels at detecting objects quickly, making it ideal for certain real-time layout analysis tasks. YOLO-NAS, like other YOLO models, uses a different annotation format. Each image is accompanied by a separate TXT file for its annotations called the YOLO format.

Early in the training process, we conducted test training runs on a subset of the dataset to evaluate the performance of both the Mask R-CNN model

and YOLO-NAS. During these tests, we closely monitored key performance metrics such as mean Average Precision (mAP), as well as the quality of the model's outputs, especially in complex layout scenarios. After manually inspecting the outputs of both models, we found that LayoutParser's Mask R-CNN model provided more robust inferences for our dataset, even though it is based on an older architecture. Despite the comparable mAP scores (around 90% for both models), the Mask R-CNN model demonstrated superior and more robust performance when it came to detecting the more complex documents within the SweMPer dataset. Thus, although YOLO-NAS is faster and more modern, we decided to proceed with LayoutParser's Mask R-CNN for its superior performance in our specific task.

Segmentation masks: Bounding boxes are a simple but effective way of representing object locations within an image, as they enclose each object with a rectangle. For layout detection, bounding boxes were used to annotate various elements on the page. However, since we chose the Mask R-CNN model, which requires segmentation masks instead of bounding boxes, we had to convert the bounding box annotations into segmentation masks. Segmentation masks provide a pixel-level representation of objects, making them more precise than bounding boxes, which are limited to rectangular shapes. To handle this, we used a script to convert the bounding boxes into pixel masks, ensuring the annotations were suitable for the Mask R-CNN model.

Transfer Learning and Fine-Tuning: Instead of training the Mask R-CNN model (He et al. 2017) from scratch, we used transfer learning to fine-tune a pre-trained model on our annotated dataset. Transfer learning is particularly useful when dealing with limited datasets because it allows us to leverage a model that has already learned a wide range of features from a large, general dataset and fine-tune it to perform well on a specific task. Instead of training a model from scratch, transfer learning allows us to build on an already powerful feature extractor.

In our case, the LayoutParser Mask R-CNN model was pre-trained on the PubLayNet dataset, which contains over 360,000 document pages. This pre-training gave the model a strong foundation for understanding document layouts, making it well-suited for our task of layout detection in historical Swedish medical journals. The classes in the PubLayNet dataset are: text, title, list, table, figure. The variant of the model used is the Mask R-CNN X-101-32x8d-FPN-3x Network.

The Mask R-CNN X-101-32x8d-FPN-3x network is a highly versatile model for both object detection and instance segmentation, composed of several key components:

- **Backbone** (**ResNeXt-101, Xie et al. 2017**): This extracts the main features from images.
- **FPN** (**Feature Pyramid Network**): Helps process images at different scales.
- **RPN** (**Region Proposal Network**): Suggests regions in the image that might contain objects.
- **RoI Align** (**Region of Interest Align**): Warps these regions of interest into a uniform size for further processing.
- **Bounding Box and Mask Heads**: Identifies the class of each object and draws bounding boxes around them, while also generating detailed masks.

When refining a pre-trained model, we freeze the early layers so they are not retrained. These layers focus on general features, like edges and textures, which are common across many types of images. For our model, we tested freezing 2, 3, and 4 sections and found that freezing the first 3 sections resulted in the best fine-tuning performance. This choice provided a balance between retaining generalizable features learned from PubLayNet and allowing the later layers to adapt to the more specific elements in our dataset.

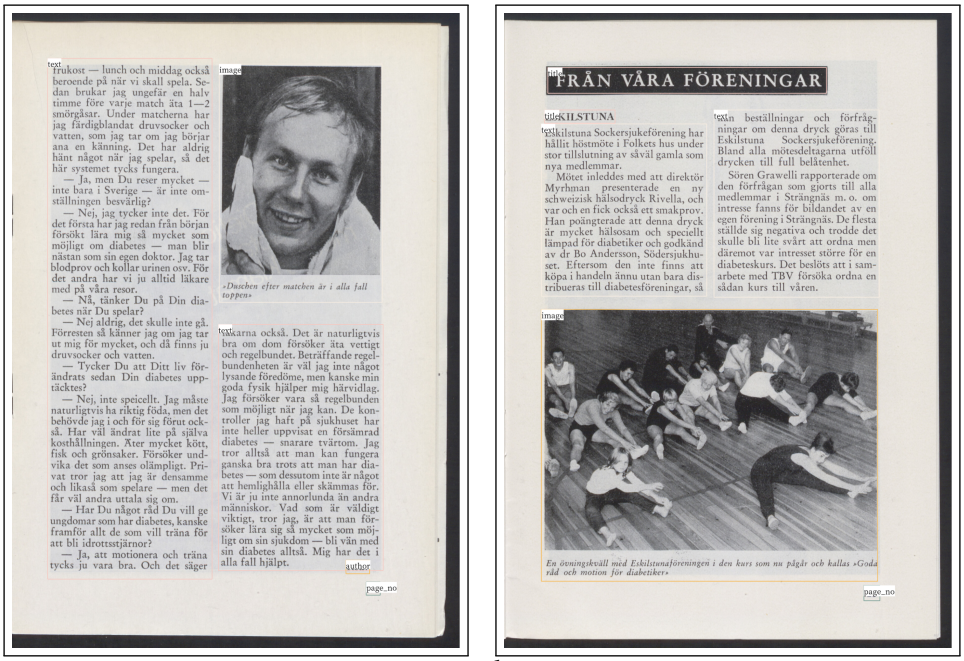
Through this approach, we maximize the benefit of transfer learning by maintaining the general document layout features learned from PubLayNet, while adapting the model to the specific challenges of historical Swedish medical journals. This ensures that the model performs well on layout detection tasks, as seen in Figure 9, and produces high-quality results in later steps, such as metadata extraction.

4.3.3 *Layout analysis*

Once the layout detection model has been trained, the next step in the SweMPer pipeline is layout analysis using the transfer learned model from above. The goal of layout analysis is to identify and classify different layout elements in historical journals. The images in Figure 10 shows the inference results using our trained layout detection model.

The table from Figure 10 contains information about the bounding boxes detected by the layout detection model:

- **x_1, y_1**: The top-left corner of the bounding box.
- **x_2, y_2**: The bottom-right corner of the bounding box.
- **block_type**: Describes the geometric shape of the layout element.



a. b.

Figure 9: Examples of layout detection results generated by our transfer learned LayoutParser model. The model accurately identifies and labels various layout elements such as “text”, “image”, “author”, “title”, and “page number”, preparing them for further steps like text extraction using OCR.

	x_1	y_1	x_2	y_2	block_type	type	score
0	144.569962	184.140015	811.390625	2298.991943	rectangle	text	0.999926
1	837.126099	1264.826416	1505.461792	2263.205811	rectangle	text	0.999638
2	840.967163	207.490494	1501.823608	1174.445557	rectangle	image	0.998965
3	1354.159546	2221.210449	1450.243896	2276.750732	rectangle	author	0.993870
4	1436.580078	2310.358398	1493.492188	2363.374756	rectangle	page_no	0.987107

Figure 10: An example output from the layout analysis model in a tabular format.

- **type**: The predicted class of the detected layout element.
- **score**: The confidence score of the predicted class, indicating how certain the model is in its classification.

This structured output provides an understanding of the layout elements present on the page. For instance, the first row in the dataframe indicates that a text block is detected within the coordinates defined by x_1, y_1, x_2, and

y_2 with a confidence score of 0.999926, meaning the model is almost certain of its prediction. By capturing these layout elements as bounding boxes, we can efficiently map the page structure and prepare it for subsequent steps, such as OCR and metadata extraction.

4.3.4 *OCR*

After the layout analysis step, the next task in the pipeline is to extract the textual content from the identified layout elements. For this, we use Optical Character Recognition (OCR) with tesseract (Smith 2007) to transform the visual information into machine-readable text. Our OCR process is efficiently integrated with the results from the layout detection model, ensuring that we only perform OCR on relevant portions of the page.

The key steps in this process are as follows:

1. **Filtering the Output by Element Type**: Once we obtain the bounding boxes and class predictions from the LayoutParser model, we filter the output DataFrame 10 to retain only the layout elements that contain textual content. Specifically, we focus on the following classes:

   - `text`
   - `title`
   - `page_no`
   - `header_or_footer`
   - `author`

2. **Cropping the Image Sections**: For each identified layout element in the filtered DataFrame, we crop the corresponding region of the original image. The bounding box coordinates (`x_1, y_1, x_2, y_2`) provided by the layout model are used to isolate the relevant portion of the page that contains text.

3. **Running Tesseract OCR**: The cropped image sections are then passed through the Swediah version of Tesseract OCR engine. Tesseract is configured to work with the Swedish language (`--lang swe`). This configuration allows Tesseract to accurately recognize the text in these cropped regions.

4. **Text Extraction**: Tesseract processes each cropped section and returns the corresponding text for each layout element. The extracted text is then stored along with the bounding box coordinates and the type of the layout element in the output dataframe, facilitating further analysis and metadata generation.

4.3.5 *Metadata generation*

In the SweMPer project, generating metadata for each processed page is a critical step that consolidates information from multiple stages of the pipeline, including layout detection, OCR, and the inherent metadata embedded within the image filenames. The goal of metadata generation is to create a structured YAML file that captures essential details about each page, enabling the creation of a comprehensive, searchable database of journal content.

Once the text elements have been extracted through OCR and combined with the layout model's output, we proceed to generate metadata in a structured YAML format. This metadata contains key information such as:

1. **File Information**: Metadata about the file itself, including its name, size, creation date, and MD5 checksum. This helps maintain data integrity and track the origin of each file.

2. **Document-Specific Information**: Information from the file names is parsed and used to fill in fields like the journal name, year of publication, volume number, and issue number.

3. **Layout Elements**: Detected layout elements such as text blocks, tables, images, page numbers, titles, etc., are described with their bounding box coordinates, ocr content (for text elements), and confidence scores.

This structured metadata not only allows for easy indexing and retrieval of journal pages but also facilitates more advanced queries. For example, users could search for pages with a specific title, or for advertisements published in a particular year. Furthermore, by capturing the OCR text within the metadata, it becomes possible to perform full-text searches across the entire corpus, making the journal collection an invaluable resource for researchers and historians.

The metadata is saved as a YAML file (as seen in Figure 11), which is a human-readable format that also supports complex data structures. This makes it easy to parse programmatically for building databases or other digital collections. With such metadata in place, a well-structured, searchable digital library can be created, offering powerful tools for exploring and analyzing Swedish medical periodicals over time.

```
Swemper-page-descriptor:
  ContainerVolumeFile: '"Diabetes_1963_vol013_nr003_0012.yaml"'
  CreationDate: '"2024-08-01"'
  DocumentTopic: '"Diabetes Research"'
  FileSize: '"0.84MB"'
  FileType: '"JPEG"'
  IdentifiedLayoutElements:
    Advertisements:
    - BoundingBox:
      - 147.48545837402344
      - 1315.8365478515625
      - 1346.800552368164
      - 972.1468505859375
      Content: ''
      Score: '"0.9977514147758484"'
      Type: '"advertisement"'
    AuthorNames: []
    HeaderOrFooters: []
    Images: []
    Lists: []
    Pagination:
    - BoundingBox:
      - 1445.08544921875
      - 2315.408447265625
      - 54.815673828125
      - 52.37255859375
      Content: '"11"'
      ElementID: '"3"'
      Score: '"0.9865828156471252"'
      Type: '"page_no"'
    Tables: []
    TextBlocks:
    - BoundingBox:
      - 131.93858337402344
      - 809.3839111328125
      - 673.6935577392578
      - 444.4246826171875
      Content: '"Det saknas en rimlig generositet

        från stat, kommun och arbetsgiva-

        re när det gäller de sockersjuka

        ungdomarna, påpekar docent Ber-

        til Söderling i en artikel i Svenska

        läkartidningen. Man spekulerar i

        alltför hög grad i de tillstötande

        s. k. sensymtomen. Det tillkommer

        knappast ett handikappvänligt"'
      ElementID: '"1"'
      Score: '"0.9994248151779175"'
      Type: '"text"'
    - BoundingBox:
      - 830.3582763671875
      - 200.9127655029297
      - 672.113525390625
      - 1048.6465606689453
      Content: '"samhälle att härvidlag uppträda

        som ett försäkringsbolag, vilket

        inte önskar ta några risker för

        framtiden.

        Om och när senkomplikationer

        uppträder och om de medför in-

        skränkningar i arbetsförmågan, det

        är frågor som bör lösas vid den

        aktuella tidpunkten och inte vid

        anställningens eller utbildningens

        början, hävdar docent Söderling.

        Vi får inte glömma att vi också har

        en skara högt kvalificerade diabe-

        tiska män och kvinnor i olika yr-

        ken, personer som gjort och allt-

        jämt gör synnerligen betydelseful-

        la insatser. Det är ytterligt motbju-

        dande att en välståndskultur som

        vår inte fattat klarare ståndpunkt

        beträffande diabetisk och annan

        handikappad ungdoms = utbild-

        ene och anställningsförhållan-

        en."'
      ElementID: '"2"'
      Score: '"0.9997593760490417"'
      Type: '"text"'
    Titles:
    - BoundingBox:
      - 133.75193786621094
      - 198.21221923828125
      - 527.6947174072266
      - 406.5040283203125
      Content: '"Arbetsgivarna

        restriktiva

        mot diabetes-

        ungdomar"'
      ElementID: '"0"'
      Score: '"0.9932644367218018"'
      Type: '"title"'
  JournalName: '"Diabetes"'
  Language: '"Swedish"'
  NumberIndex: '"nr003"'
  PageFilename: '"Diabetes_1963_vol013_nr003_0012.jpg"'
  PageIndex: '"0012"'
  Source: '"diabetes(in-house scanned)"'
  Timestamp: '"2024-08-01T07:01:24.435795"'
  TotalPages: '"209"'
  VolumeIndex: '"vol013"'
  Year: '"1963"'
  md5_checksum: '"04e9d6f562eb98bde52513d821cc2b21"'
```

Figure 11: Example of a YAML file showcasing the metadata output from our layout detection model and OCR processing. Each file contains detailed information, including the bounding boxes of layout elements (e.g., text blocks, advertisements, page numbers), confidence scores for detected elements, extracted text content, and additional document metadata like file size, journal name, and timestamp.

4.4 *Future work*

4.4.1 *Active Learning*

In order to further improve the performance of the layout detection model, we will incorporated an Active Learning (Settles 2009) approach. Specifically, we will utilize uncertainty sampling (Settles 2009: 11–12). Active learning is a process where we focus on the most challenging cases for the model—those where the model is least confident in its prediction. By having humans review these difficult cases, the model learns from its mistakes, making it more accurate in future predictions.

One important consideration in this process is the potential for class imbalance and bias in model predictions. If certain layout elements are consistently misclassified or have lower confidence scores due to under-representation in the dataset, active learning allows us to correct this by prioritizing uncertain cases for annotation. This ensures that less frequent or more complex elements, such as tables and lists, are sufficiently represented in the training data. While OCR-related biases, such as errors arising from faded text or unusual typefaces, are different in nature from class imbalance in layout detection, both can contribute to errors in downstream tasks. Addressing these biases systematically through improved training and data selction strategies is an area for further study.

In our Active Learning loop, we will analyze the inferences made by our trained layout detection model and identify the elements whose class confidence scores were low and hovered around the threshold of 80%. Cases where the model exhibits uncertainty in classifying layout elements will be manually annotated and added to the training dataset. The goal of this strategy is to improve the model's performance by fine-tuning it on examples where its predictions were least confident, thereby ensuring that it learns from its own mistakes while also mitigating potential biases in class representation.

4.4.2 *Create a searchable database*

We aim to create a searchable platform to access the extracted metadata and images from the SweMPer project. This platform will be developed using GraphQL for efficient data querying and React for building the user interface. Users will be able to search, browse, and filter the dataset based on metadata, such as journal name, year, and layout elements. This database will serve as a valuable tool for researchers, allowing them to explore historical Swedish medical journals through an intuitive interface.

5 *Hands-on practical introduction*

5.1 *Converting Labelme rectangles to YOLO format*

Link to the code for converting Labelme Rectangles to YOLO Format: https://www.kaggle.com/code/daliaortizpablo/hands-on-labelme-to-yolo

5.2 *Code for selecting images in Section 4.3.2*

Link to the code used for selecting images: https://github.com/CDHUppsala/swemper_ML/blob/main/beast_code/utils/data_prep.py

5.3 *Code to create metadata for Section 4.3.5*

Link to the code: https://github.com/CDHUppsala/swemper_ML/blob/main/beast_code/create_metadata_v1.py

5.4 *Other code and instructions*

You can find more code related to the project and instructions at the GitHub repository here: https://github.com/CDHUppsala/swemper_ML/tree/main

References

Aangenendt, Gijs, Maria Skeppstedt & Ylva Söderfeldt. 2024. Curating a historical source corpus of 20th century patient organization periodicals. In *Huminfra conference*, 76–82.

AI, Deci. 2023. *Yolo-nas: a new state-of-the-art object detection model*. https://www.deci.ai/yolo-nas/. Accessed: September 23, 2024.

Arroyo-Ramirez, Elvia. 2016. Invisible defaults and perceived limitations: Processing the juan gelman files. *Medium*.

Bahrami, Mahdi & Amir Albadvi. 2024. Deep learning for identifying Iran's cultural heritage buildings in need of conservation using image classification and grad-cam. *ACM Journal on Computing and Cultural Heritage* 17(1). 1–20.

Bender, Edward A. 2000. *An introduction to mathematical modeling*. Courier Corporation.

Bolukbasi, Tolga, Kai-Wei Chang, James Zou, Venkatesh Saligrama & Adam Kalai. 2016. *Man is to computer programmer as woman is to homemaker? Debiasing word embeddings*. https://arxiv.org/abs/1607.06520.

Bunge, Mario & Mario Bunge. 1973. Mathematical modeling in social science. *Method, Model and Matter*. 131–142.

Colavizza, Giovanni, Tobias Blanke, Charles Jeurgens & Julia Noordegraaf. 2022. Archives and ai: An overview of current debates and future perspectives. *Journal on Computing and Cultural Heritage* 15(1). 1–15. DOI: 10.1145/3479010.

Ginzburg, Carlo. 1979. Clues: Roots of a scientific paradigm. *Theory and Society* 7(3). 273–288.

Goodfellow, Ian. 2016. *Deep learning*.

Google. N.d. *Machine learning glossary*. https://developers.google.com/machine-learning/glossary#t. Accessed: September 9, 2024. Updated: September 4, 2024.

Gruber, Ivan, Pavel Ircing, Petr Neduchal, Marek Hrúz, Miroslav Hlaváč, Zbyněk Zajíc, Jan Švec & Martin Bulín. 2020. An automated pipeline for robust image processing and optical character recognition of historical documents. In *International conference on speech and computer*, 166–175.

He, Kaiming, Georgia Gkioxari, Piotr Dollár & Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the ieee international conference on computer vision*, 2961–2969.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren & Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition*, 770–778.

Hill, Mark J & Simon Hengchen. 2019. Quantifying the impact of dirty ocr on historical text analysis: Eighteenth century collections online as a case study. *Digital Scholarship in the Humanities* 34(4). 825–843.

Jaillant, Lise & Annalina Caputo. 2022. Unlocking digital archives: Cross-disciplinary perspectives on ai and born-digital data. *AI & society* 37(3). 823–835.

Jarlbrink, Johan & Pelle Snickars. 2017. Cultural heritage as digital noise: Nineteenth century newspapers in the digital archive. *Journal of Documentation* 73(6). 1228–1243.

Jockers, Matthew L. 2013. *Macroanalysis: Digital methods and literary history*. University of Illinois Press.

Kaur, Ravpreet & Sarbjeet Singh. 2023. A comprehensive review of object detection with deep learning. *Digital Signal Processing* 132. 103812.

Krišto, Mate, Marina Ivasic-Kos & Miran Pobar. 2020. Thermal object detection in difficult weather conditions using yolo. *IEEE access* 8. 125459–125476.

Labellerr. 2023. *A detailed guide to best practices in image annotation*. https://www.labellerr.com/blog/best-practices-in-image-annotation/.

Written by Akshit Mehra. Accessed: August 29, 2024. Published: Nov 23, 2023.

Lee, Benjamin Charles Germain, Jaime Mears, Eileen Jakeway, Meghan Ferriter, Chris Adams, Nathan Yarasavage, Deborah Thomas, Kate Zwaard & Daniel S. Weld. 2020. The newspaper navigator dataset: Extracting and analyzing visual content from 16 million historic newspaper pages in chronicling america. *CoRR* abs/2005.01583. https://arxiv.org/abs/2005.01583.

Małecki, Piotr. 2006. *Road*. Wikimedis Commons. Licensed under the GNU Free Documentation License and Creative Commons Attribution-Share Alike 3.0 Unported License. Modified from the original. https://creativecommons.org/licenses/by-sa/3.0/ (25 June, 2024).

Marciano, Richard, Victoria Lemieux, Mark Hedges, Maria Esteva, William Underwood, Michael Kurtz & Mark Conrad. 2018. Archival records and training in the age of big data. In *Re-envisioning the mls: Perspectives on the future of library and information science education*, vol. 44, 179–199. Emerald Publishing Limited.

Matschak, Tizian, Florian Rampold, Malte Hellmeier, Christoph Prinz & Simon Trang. 2022. A digitization pipeline for mixed-typed documents using machine learning and optical character recognition. In *International conference on design science research in information systems and technology*, 195–207.

Medium. 2020. *Introducing label studio, a swiss army knife of data labeling*. https://towardsdatascience.com/introducing-label-studio-a-swiss-army-knife-of-data-labeling-140c1be92881. Written by Nikolai Liubimov. Accessed: September 6, 2024. Published: Januaty 8, 2020.

Medium. 2024. *A quick reference for bounding boxes in object detection*. https://medium.com/@rajdeepsingh/a-quick-reference-for-bounding-boxes-in-object-detection-f02119ddb76b/. Written by Rajdeep Singh. Accessed: June 25, 2024. Published: Jan 17, 2024.

Mikołajczyk, Agnieszka & Michał Grochowski. 2018. Data augmentation for improving deep learning in image classification problem. In *2018 international interdisciplinary phd workshop (iiphdw)*, 117–122.

Moretti, Franco. 2005. *Graphs, maps, trees: Abstract models for a literary history*. London/New York: Verso.

Naiman, Jill P, Peter KG Williams & Alyssa Goodman. 2023. The digitization of historical astrophysical literature with highly localized figures and figure captions. *International Journal on Digital Libraries*. 1–21.

Nanonets. N.d. *Annotation best practices for object detection*. https://nanonets.github.io/tutorials-page/docs/annotate. Accessed: August 29, 2024.

Neudecker, Clemens, Konstantin Baierer, Mike Gerber, Christian Clausner, Apostolos Antonacopoulos & Stefan Pletschacher. 2021. A survey of ocr evaluation tools and metrics. In *Proceedings of the 6th international workshop on historical document imaging and processing*, 13–18.

Nicoletti, Leonardo & Dina Bass. 2023. Humans are biased. Generative AI is even worse. *Bloomberg Technology + Equality*. https://www.bloomberg.com/graphics/2023-generative-ai-bias/.

Ntoutsi, E., P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl, M. Vidal, S. Ruggieri, F. Turini, S. Papadopoulos, E. Krasanakis, I. Kompatsiaris, K. Kinder-Kurlanda, C. Wagner, F. Karimi, M. Fernández, H. Alani, B. Berendt, T. Kruegel, C. Heinze, K. Broelemann, G. Kasneci, T. Tiropanis & S. Staab. 2020. Bias in data-driven artificial intelligence systems—an introductory survey. *WIREs Data Mining and Knowledge Discovery* 10. DOI: 10.1002/widm.1356.

Obeso, Abraham Montoya, Jenny Benois-Pineau, Alejandro Álvaro Ramirez Acosta & Mireya Saraí García Vázquez. 2017. Architectural style classification of mexican historical buildings using deep convolutional neural networks and sparse features. *Journal of Electronic Imaging* 26(1). 011016–011016.

OpenCV. 2024. *Data annotation – a beginner's guide*. https://opencv.org/blog/data-annotation/. Written by Alvi Farooq. Accessed: June 4, 2024. Published: Februry 21, 2024.

Pastaltzidis, Ioannis, Nikolaos Dimitriou, Katherine Quezada-Tavarez, Stergios Aidinlis, Thomas Marquenie, Agata Gurzawska & Dimitrios Tzovaras. 2022. Data augmentation for fairness-aware machine learning: Preventing algorithmic bias in law enforcement systems. In *Proceedings of the 2022 acm conference on fairness, accountability, and transparency*, 2302–2314.

Pavlopoulos, Georgios A, Maria Secrier, Charalampos N Moschopoulos, Theodoros G Soldatos, Sophia Kossida, Jan Aerts, Reinhard Schneider & Pantelis G Bagos. 2011. Using graph theory to analyze biological networks. *BioData mining* 4. 1–27.

Rafael Padilla, Amy Roberts & the Hugging Face Team. 2023. *Open object detection leaderboard*. https://huggingface.co/spaces/hf-vision/object_detection_leaderboard.

von Ranke, Leopold. [1824] 1885. *Geschichten der romanischen und germanischen völker von 1494 bis 1514*. Verlag von Dunder und Humblot.

Redmon, Joseph, Santosh Divvala, Ross Girshick & Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the ieee conference on computer vision and pattern recognition*, 779–788.

Reitermanova, Zuzana. 2010. Data splitting. In *Wds*, vol. 10, 31–36.

Roboflow. 2024a. *Best ocr models for text recognition in images*. https://blog.roboflow.com/best-ocr-models-text-recognition/. Written by Leo Ueno. Accessed: September 23, 2024. Published: March 16, 2024.

Roboflow. 2024b. *How to label image data for computer vision models*. https://blog.roboflow.com/tips-for-how-to-label-images/. Written by Joseph Nelson. Accessed: August 29, 2024. Published: January 5, 2024.

Russell, Bryan C, Antonio Torralba, Kevin P Murphy & William T Freeman. 2008. Labelme: A database and web-based tool for image annotation. *International journal of computer vision* 77. 157–173.

Settles, Burr. 2009. Active learning literature survey.

Shen, Zejiang, Ruochen Zhang, Melissa Dell, Benjamin Charles Germain Lee, Jacob Carlson & Weining Li. 2021. Layoutparser: A unified toolkit for deep learning based document image analysis. In *Document analysis and recognition–icdar 2021: 16th international conference, lausanne, switzerland, september 5–10, 2021, proceedings, part i 16*, 131–146.

Shorten, Connor & Taghi M Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. *Journal of big data* 6(1). 1–48.

Smith, Astrid J. & Bridget Whearty. 2023. All the work you do not see: Labor, digitizers, and the foundations of digital humanities. In Lauren F. Klein & Matthew K. Gold (eds.), *Debates in the digital humanities 2023*, 27–46. University of Minnesota Press.

Smith, Ray. 2007. An overview of the tesseract ocr engine. In *Icdar '07: Proceedings of the ninth international conference on document analysis and recognition*, 629–633. Washington, DC, USA: IEEE Computer Society. https://storage.googleapis.com/pub-tools-public-publication-data/pdf/33418.pdf.

Thomas, Julia & Irene Testini. 2024. Capturing captions: Using AI to identify and analyse image captions in a large dataset of historical book illustrations. *Digital Humanities Quarterly* 18(2).

Ultralytics. 2023. *Object detection datasets overview*. https://docs.ultralytics.com/datasets/detect/. Written by Glenn Jocher, Rizwan Munawar, Ivor Zhu, Laughing-Q. Accessed: June 26, 2024. Published: Nov 12, 2023. Updated: Jun 2, 2024.

van Strien, Daniel, Kaspar Beelen, Mariona Ardanuy, Kasra Hosseini, Barbara McGillivray & Giovanni Colavizza. 2020. Assessing the impact of OCR quality on downstream NLP tasks. In *Proceedings of the 12th international conference on agents and artificial intelligence*, 484–496. Valletta, Malta: SCITEPRESS. DOI: 10.5220/0009169004840496.

Villaespesa, Elena & Seth Crider. 2021. A critical comparison analysis between human and machine-generated tags for the metropolitan museum of art's collection. *Journal of Documentation* 77(4). 946–964.

Wang, Rui-Sheng, Assieh Saadatpour & Reka Albert. 2012. Boolean modeling in systems biology: An overview of methodology and applications. *Physical biology* 9(5). 055001.

Wevers, Melvin & Thomas Smits. 2020. The visual digital turn: Using neural networks to study historical images. *Digital Scholarship in the Humanities* 35(1). 194–207.

Xailient. 2024. *Exploring data labeling and the 6 different types of image annotation*. https://xailient.com/blog/exploring-data-labeling-and-the-6-different-types-of-image-annotation/. Written by Sabina Pokhrel. Accessed: June 25, 2024. Published: March 11, 2024.

Xie, Saining, Ross Girshick, Piotr Dollár, Zhuowen Tu & Kaiming He. 2017. Aggregated residual transformations for deep neural networks. In *Proceedings of the ieee conference on computer vision and pattern recognition*, 1492–1500.

Xie, Xiaoyuan, Joshua WK Ho, Christian Murphy, Gail Kaiser, Baowen Xu & Tsong Yueh Chen. 2011. Testing and validating machine learning classifiers by metamorphic testing. *Journal of Systems and Software* 84(4). 544–558.

Xu, Yun & Royston Goodacre. 2018. On splitting training and validation set: A comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. *Journal of analysis and testing* 2(3). 249–262.

Yosinski, Jason, Jeff Clune, Yoshua Bengio & Hod Lipson. 2014. How transferable are features in deep neural networks? *Advances in neural information processing systems* 27.

Zhang, Yuanyuan, Yixuan Zhang, Bence Mark Halpern, Tanvina Patel & Odette Scharenborg. 2022. Mitigating bias against non-native accents. In *Proceedings of the annual conference of the international speech communication association, interspeech*, vol. 2022, 3168–3172.

Zhong, Xu, Jianbin Tang & Antonio Jimeno Yepes. 2019. Publaynet: Largest dataset ever for document layout analysis. In *2019 international conference on document analysis and recognition (icdar)*, 1015–1022.

Zhu, Xizhou, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang & Jifeng Dai. 2020. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*.

Zou, Zhengxia, Keyan Chen, Zhenwei Shi, Yuhong Guo & Jieping Ye. 2023. Object detection in 20 years: A survey. *Proceedings of the IEEE* 111(3). 257–276.

List of abbreviations

AI	Artificial Intelligence
CDHU	Centre for Digital Humanities and Social Sciences (Uppsala University)
CNN	Convolutional Neural Network
mAP	mean Average Precision
ML	Machine Learning
NLP	Natural Language Processing
OCR	Optical Character Recognition
RCNN	Region Based Convolutional Neural Networks
SweMPer	Communicating Medicine: Digitalisation of Swedish Medical Periodicals, 1781–2011
ROI	Region of Interest

Corresponding authors

Dalia Ortiz Pablo
Department of ALM
Uppsala University
dalia.ortiz_pablo@abm.uu.se

Sushruth Badri
Department of ALM
Uppsala University
sushruth.badri@abm.uu.se

9 SweLL with pride: How to put a learner corpus to good use

Elena Volodina
University of Gothenburg

Arianna Masciolini
University of Gothenburg

Beáta Megyesi
Stockholm University

Julia Prentice
University of Gothenburg

Lisa Rudebeck
Stockholm University

Gunlög Sundberg
Stockholm University

Mats Wirén
Stockholm University

Second language (L2) learner corpora are collections of language samples that demonstrate learners' abilities to perform some learning tasks, e.g. an ability to write essays, answer to reading comprehension questions, or talk on a given topic. Such corpora are necessary for both empirical-based research within Second Language Acquisition (SLA), and for development of methods for automatic processing of such data. L2 corpora are notoriously difficult to collect, and their value depends to a greater degree on the representativeness and balance of the sampled data, type of associated metadata and reliability of manual annotations.

In this chapter we thoroughly describe the SweLL-gold corpus of L2 Swedish, its annotation, statistics and metadata, and showcase main types of its use, such as (1) in research on SLA through detailed instructions on how to perform corpus searches given SweLL-specific annotation, combined with guidelines for SVALA usage, a tool for correction annotation; and (2) in NLP research on problems such as grammatical error correction through guidelines on how to use the different available file formats that the SweLL-gold corpus is released in. Both cases are further supported by case studies and, where available, relevant scripts ready for reuse by researchers.

1 *Introduction*

Second Language Acquisition (SLA) is a broad research area that deals with theoretical, methodological and, by extension, practical aspects of how a language is acquired by non-native speakers. There exist multiple SLA theories and approaches (e.g. VanPatten et al. 2025), most of which need access to empirical data for generating hypotheses or for drawing conclusions. However,

HUM INFRA Elena Volodina, Arianna Masciolini, Beáta Megyesi, Julia Prentice, Lisa Rudebeck, Gunlög Sundberg & Mats Wirén. 2025. SweLL with pride: How to put a learner corpus to good use. In Gerlof Bouma, Dana Dannélls, Dimitrios Kokkinakis & Elena Volodina (eds.), *Huminfra handbook: Empowering digital and experimental humanities* (NEALT Proceedings Series 59), 251–306. University of Tartu Library. DOI: 10.58009/aere-perennius0178
© The authors, CC BY 4.0

learner-produced data is non-trivial to collect (e.g. Volodina et al. 2016), the real bottleneck being manual annotation necessary for effective SLA analysis. Relevant annotation includes, for example, mark-up of deviations, errors and their correction, identification of linguistic features that reflect maturity of language use, essay grading, to name just a few.[1]

The availability of manually annotated learner corpora is critical not only for traditional SLA research, but also for tasks at the intersection between Natural Language Processing (NLP) and SLA, such as training systems for automatic error correction and classification. So, what are learner corpora?

Learner corpora are collections of essays or transcripts of speech produced by learners of some language, where electronic data are used for empirical evidence in research on the development of learner language or in related fields. Some examples of written essay corpora are ASK for L2 Norwegian (Tenfjord et al. 2006), FALKO for L2 German (Lüdeling et al. 2005, Reznicek et al. 2012), MERLIN for L2 Czech, German and Italian (Boyd et al. 2014), COPLE2 for L2 Portuguese (Mendes et al. 2016), CzeSL for L2 Czech (Rosen et al. 2020), LAVA for L2 Latvian (Dargis et al. 2020), Icelandic L2 Error Corpus (Glisic & Ingason 2022) and multiple learner corpora for L2 English (e.g. Yannakoudakis et al. 2011, Paquot 2022, Vinogradova & Lyashevskaya 2022).

For L2 Swedish there exist several collections, such as CrossCheck with essays from different levels of schools/courses (Lindberg & Eriksson 2005), ASU with L2 essays and transcribed L2 speech (Hammarberg 2010), and Uppsala Corpus of Student Writings with an extensive collection of essays from Swedish national exams (Megyesi et al. 2016). These corpora are valuable, reflecting different aspects of L2 Swedish, but some are not easy to gain access to, and none has correction annotation, unlike the SweLL-gold corpus that we describe in this chapter.

This chapter presents SweLL-gold (Volodina et al. 2019, Volodina, Granstedt, et al. 2025), the first electronic corpus of learner Swedish that has been manually corrected, annotated with correction labels and pseudonymized to protect learner identities. SweLL-gold provides unique opportunities for SLA and NLP research. In this chapter we briefly introduce the SweLL infrastructure; describe the SweLL-gold corpus, summarizing corpus statistics, metadata, and annotations; describe research questions that can be answered with the help of such data, exemplifying spin-off projects and resources. We devote Sections 3 and 4 to two use cases, namely, use case 1 focusing on Second Language Acquisition research based on Sundberg &

1 Parts of this article have been previously published in https://ecp.ep.liu.se/index.php/hic/article/view/896, a conference paper that has been invited as an extended chapter for the current Huminfra Handbook

Prentice (2023) & step-by-step guidelines for SLA researchers; and use case 2 aimed at Natural Language Processing research & step-by-step guidelines for NLP researchers.

2 *SweLL-gold in a nutshell*

2.1 *The SweLL infrastructure project*

The purpose of the SweLL infrastructure project, funded by Riksbankens Jubileumsfond,[2] was to set up an infrastructure for collection, digitization, normalization and annotation of learner written production, as well as to make available a linguistically annotated corpus, where it would be possible to search for various types of linguistic structures, without the researcher having to guess what such a structure might look like, since there is a parallel normalized version available. For instance, by searching for the target form *mycket* 'much' it would be possible to see misspellings present in such a corpus; for example, in the SweLL-gold learner corpus we can find the following misspellings:[3] **mycke*, **mycker*, **myckt*, **mcyket*, **meka*, **myckena*, **mycki*, **myckit*.

The SweLL infrastructure released in its first version consists of:

1. a data collection portal, the SweLL portal (Mohammed et al. 2022), which is a user interface that regulates access to the SweLL database where all essays and their metadata are stored. The portal also facilitates administration of annotation steps. Essay upload and download in different formats is also handled through the SweLL portal.

2. the SVALA annotation tool for L2 analysis (Wirén et al. 2019), primarily, for manual normalisation of learner-written texts, manual addition of correction tags and manual pseudonymization of personal information. SVALA communicates directly with the SweLL portal database where versioning is automatically handled.

3. an annotated corpus of L2 written productions, SweLL-gold, consisting of 502 correction annotated essays collected from adult learners of Swedish (Volodina et al. 2019).[4]

2 https://spraakbanken.gu.se/en/projects/swell

3 see Korp search for "target form: mycket" in SweLL-gold original

4 A bonus corpus, SweLL-pilot (Volodina 2024), was released together with the SweLL-gold, although we will not describe it in this chapter. Apart from SweLL-gold and SweLL-pilot, a number of other essays are stored on the portal which are at different stages of annotation and release history.

4. specific search solutions for second language materials facilitating filtering for relevant variables, e.g. texts written by male writers or writers at a certain proficiency level, deployed on Korp corpus management system (Borin et al. 2012).
5. SweLL-gold metadata description and statistics.
6. multiple guidelines describing the process of annotation and collection:
   - transcription guidelines (Volodina & Megyesi 2021)
   - pseudonymization guidelines (Megyesi et al. 2021)
   - normalization[5] guidelines (Rudebeck et al. 2021)
   - correction annotation guidelines (Rudebeck & Sundberg 2021)
7. user manuals, among others:
   - SVALA manual for users,[6] i.e. for those who want to explore data in a demo mode (English only)
   - SVALA manual for annotators,[7] i.e. for those who annotate with the intention to save the annotated versions on the server (Swedish only)
   - SweLL portal manual[8] (English only)

Thorough work has been carried out to make sure that GDPR guidelines and ethical principles are followed. In consultation with university lawyers, the access principles have been defined and double-checked. Access is granted to researchers, developers and teachers following an application for use.[9] According to the GDPR, users outside Europe cannot get immediate access to the data in its entirety. Their applications need to be processed by the university lawyers on a case by case basis. Applicants inside European Union will get access to the full dataset provided their intended use targets L2-oriented research, development or pedagogical applications.

The data can be browsed through Korp (Borin et al. 2012) with specific search solutions for L2 material facilitating filtering for e.g. texts written by writers of a certain age, gender, mother tongue, or writers at a certain proficiency level with a possibility for full-text view.

5 We are aware of the other uses of the term *normalization*, e.g. with reference to standardization of spelling conventions in the field of social media data. We extend this term to standardization of grammar, syntax, punctuation and word choice as well. See our interpretation of the term *normalization* in Section 2.3.4.

6 https://docs.google.com/document/d/1YSpphG3tHe5UlkAjMZ2cfgsrtZHnO-05Uo-x0TKh4QY/

7 https://drive.google.com/file/d/18-VkZDecHXdq5DKaOYx5KsOP7BZTjsjW/

8 https://docs.google.com/document/d/11v6GGRcIkINrb0-HRD4Xrn7PFJQyjl-iUGy0VD-0rjI/

9 https://sunet.artologik.net/gu/swell

2.2 *SweLL-gold in numbers*

SweLL-gold is, as might be clear from the above, a corpus of essays written by adult learners of Swedish with non-Swedish backgrounds. The essays were collected during 2017–2021 from several schools in Sweden, in total 11 distinct schools from different geographic areas, who agreed to collaborate. The involved teachers assisted with consent forms, personal and task metadata forms, and essays. The type of the course was used as an indication of the approximate level of learners, e.g. *upper-secondary* and *university preparatory* courses being representative of *Advanced* levels (C); *SVA* (Swedish as a Second Language) courses for adults representing *Intermediate* levels (B); and *SFI* courses (Swedish For Immigrants) representing *Beginner* levels (A). Apart from the essays themselves, we collected rich metadata about schools, learners, tasks, and individual essay characteristics, which are well-documented on the SweLL-gold metadata webpage.[10]

2.2.1 *Metadata*

Metadata – i.e. data about data, such as information about authors or tasks – is extremely important for pursuing different types of research and for ensuring interoperability between corpora (Volodina et al. 2018, König et al. 2021). For example, age, gender and first languages are important for identifying learning problems specific to different demographic groups, while task metadata allows studying the impact of the task on the type of language produced by learners in the essays. However, there are many other metadata aspects that are easily overlooked by corpus compilers, although similarly important.

Work on metadata standardization in Learner Corpus Research (LCR) was initiated by Granger & Paquot (2017), was followed up by König et al. (2022) and is ongoing at the moment of writing (Paquot et al. 2023, 2024). Paquot et al. (2023) identify eight groups of metadata – administrative, corpus design, learner, text, task, annotation, annotator and transcriber[11] – with multiple subcategories divided into obligatory and optional. In the SweLL-gold corpus, most of the obligatory metadata is considered, however, they are grouped a bit differently into five categories: administrative, personal, task, essay, and school – shortly summarized below. Note that we did not collect metadata information on transcribers and annotators, which is difficult to rectify in retrospect.

10 https://spraakbanken.github.io/swell-release-v1/Metadata-SweLL
11 Work-in-progress document available here: https://tinyurl.com/L2metadataV2

Table 1: Overview of SweLL-gold statistics.

		Original essays	Normalized versions
Nr. tokens		147 842	151 851
Nr. sentences		7 807	8 137
Nr. essays	Beginner courses	289	289
	Intermediate courses	45	45
	Advanced courses	168	168
	Total	502	502

2.2.2 *Administrative metadata*

Administrative metadata covers, among other things, authorship, corpus design, period of collection, etc., corpus statistics being the most relevant for this article (Table 1). SweLL-gold contains 502 *original* (i.e. learner-written) essays and 502 *normalized* (i.e. minimally corrected, as explained in Section 2.3.5) versions of the original essays. The number of sentences and tokens in the original and normalized versions deviate slightly as a result of the corrections, as shown in Table 1. The same table also shows the distribution of essays from several course types, grouped under three headings: Beginner, Intermediate and Advanced.

All collected texts were *manually* processed as follows: transcribed from hand-written paper samples; pseudonymized; normalized (i.e. edited to create a separate, corrected, version); and labeled for the type of corrections that were made to the text. On top, *automatic* annotation (part of speech tagging, lemmatization, syntactic analysis, etc.) have been added to both versions of SweLL-gold essays (original and normalized).

2.2.3 *Personal metadata*

Personal metadata includes information about the data subjects, i.e. the learners. We collected information about learners' gender, age, educational background, time in Sweden, mother tongues, knowledge of other languages, etc. SweLL-gold essays represent 321 distinct learners, with some who wrote more than one essay (see Table 2), allowing insights into the individual development of their linguistic competence.

Learners represent a big age span (see Figure 1), which clearly shows that learners in the age span 20-30 years (born between 1985 and 1999) dominate courses of this type; the age range 50-65 (born between 1950 and 1979) is rather underrepresented among learners.

Table 2: Number of recurrent students in the SweLL-gold collection.

Nr essays per student	Nr students	Total nr essays
1	193	193
2	89	178
3	25	75
4	14	56
Total	321	502

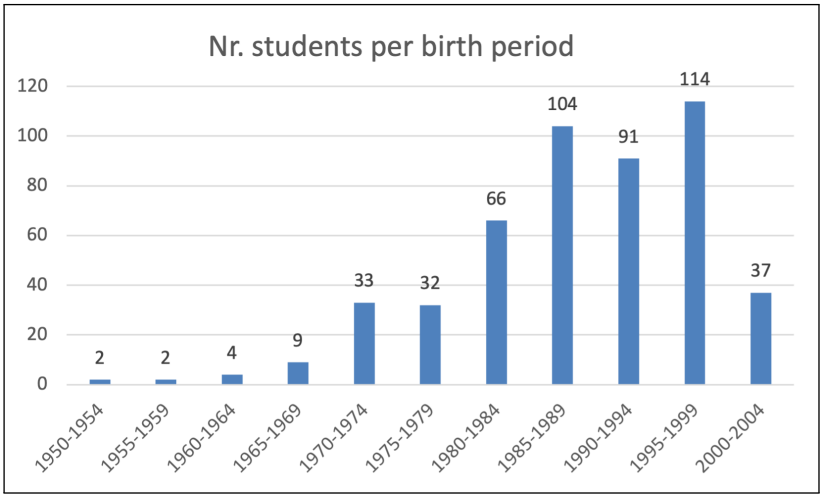


Figure 1: Distribution of age groups in the SweLL-gold data. All essays were collected during 2017–2021.

The distribution of mother tongues in the SweLL-gold corpus reflects the migrant situation of the time, namely, the big waves of immigration to Sweden, following several crises in the Middle East and Africa around 2015. Among the most frequent/well-represented mother tongues in SweLL-gold are Arabic, Dari, Kurdish, Persian, Tigrinya and Turkish, with an absolute dominance of Arabic. Simultaneously, we see English, Greek, Russian and Spanish among other frequent mother tongues, which hypothetically results from the fact that many learners report more than one first language. All in all, there are 81 unique mother tongues, in 117 unique language combinations,[12] 77 of which are represented by one or two speakers in the corpus (e.g. Badinani, Berber, Igbo, Tamazigh, Yoruba).

12 Learners were allowed to indicate more than one mother tongue.

Table 3: Number of essays written by learners with the different educational backgrounds.

Educational level	Learners' educational background	Nr. essays
1	0–6 years of schooling (incl. elementary school)	34
2	7–9 years (incl. high school)	56
3	10–13 years (incl. upper-secondary education)	155
4	14+ years (incl. university education)	257

Learners' educational background, i.e. length of previous schooling, can hypothetically influence the success of their learning. Therefore, we collected information about their previous education in number of months per school type, which we summarize in Table 3 in the form of a numeric representation of educational level. It is clear from the numbers that well-educated learners dominate in this essay collection.

Given the various metadata information about learners, it is possible to cross-reference, for example, educational backgrounds with learners' mother tongues, age groups and with the indicators of their linguistic performance, such as error types.

2.2.4 *Task metadata*

Task metadata includes formal characteristics (date, course level, handwritten vs digitally written, time allowed for the task, allowed help materials, etc.), and task characteristics (topic, genre, task instructions/task prompt, grading scale, etc.).

There are 44 distinct task IDs (task prompts), the number of essays per task prompt varies between 2 and 39. Task types are summarized in Table 4, text genres - in Table 5. A list of topics (in Swedish) is provided in Appendix 1.

2.2.5 *Essay metadata*

Essay metadata includes individual essay ID, grade (if any) and some optional metadata that are present in some other learner corpora, such as `cefr_level`, `final_grade`, etc. for homogeneous representation. In the essays where this information is not present, the value for these metadata attributes is empty.

Out of 502 essays, 108 essays received some form of grade, the breakdown of which is shown in Table 6.

Table 4: Number of essays per task type.

Task type	Nr. essays
Formative assessment task	226
Final test	149
Placement test	24
National exam	18
Entrance test	15
n/a	70
Total	502

Table 5: Approximation of the essay genres in the corpus. Terminology used by teachers varied, more than one genre could be assigned to the same essay task.

Genres	Nr. essays
Argumentative	214
Descriptive	133
Investigative	72
Narrative	58
Instruction	25
Total	502

It could be added here that since there are several learner corpora that are maintained through the SweLL portal, there is ongoing work on harmonizing metadata annotations between those. Among such initiatives, work is being performed on re-assessment of SweLL-gold essays with the CEFR levels (A1, A2, B1, B2, C1, C2) (Council of Europe 2001). This additional annotation will be added in the next SweLL release.

For people who want to have a look at the actual essays, we provide a link to the full essay (as part of the metadata file in the download package). Using the link, the user can view the chosen essay in the *demo version* of the SVALA annotation tool, which is disconnected from the SweLL portal database and, hence, presents no risk of corrupting the source files.

2.2.6 *School metadata*

School metadata, presented in Table 7, includes information about the schools the essays are sourced from. For each school, the metadata consists of a unique identifier, information about the type of education the

Table 6: Breakdown of graded essays in the SweLL-gold corpus per grading scale.

Grading scale		Nr. task IDs	Grades	Nr. essays
G/U	University admission, TISUS*	3	G (pass), U (fail)	41
			G	24
			U	17
1–7	Placement in university courses: preparatory courses (grades 2–4), qualifying courses (grades 5–7)	1	1 – 7	39
			2	4
			3	5
			4	20
			5	6
			6	4
A–F	Adult education	1	A (top) – E, F (fail)	4
			C	2
			E	1
			F	1
A–D	Placement in SFI A–D courses	3	A (lowest), B, C, D	24
			SFI B	3
			SFI C	12
			SFI D	9

*TISUS: Test in Swedish for university studies; SFI: Swedish for immigrants

school provides and a short description. Note that neither the name and address of the school nor the names of the involved teachers are included. This was done intentionally to lower the risk of re-identification of learners. In an additional step, each school/course type has been mapped to an approximate proficiency level: Beginner (A), Intermediate (B), and Advanced (C), which is also shown in the table.

2.3 *SweLL-gold annotation*

2.3.1 *Manual processing*

Annotation standards in Learner Corpus Research (LCR) cover both manual and automatic annotation, stratified further into linguistic annotation, anonymization (vs pseudonymization), normalization, error annotation (vs correction annotation), etc. (Stemle et al. 2019, Paquot et al. 2024). Included here are also tools for annotation and annotation management. Traditionally, XML has been the dominant format for learner corpora. In this case, corrections are directly applied to the original texts as markup (e.g. Tenfjord

Table 7: List of source schools and number of essays collected from those.

ID	School type	Description	Approx. level	Nr. essays
A	Center for adult education	SFI placement: A–D	A	24
B	Upper secondary school	SVA	C	48
C	Municipal adult education	SFI A–D	A	164
E	Preparatory courses / university	*equiv.* upper secondary school	C	46
F	Preparatory courses / university	*equiv.* upper secondary school	C	20
G	Test in SFI	SFI A–D	A	101
H	TISUS	*equiv.* upper secondary school	C	15
J	Basic adult education	SVA dk 1–4	B	4
K	Basic adult education	SVA dk 1–4	B	41
L	SVA at upper secondary level	*equiv.* Adult education	C	–
M	Placement test for preparatory and qualifying courses / university		C	39

School L provided a number of essays, which, however, are not included in the released version.
SFI: Swedish for immigrants; SVA: Swedish as a second language, a block of courses of the next level after SFI, consisting of basic course followed by courses 1-4; TISUS: Test in Swedish for university studies.
Approximate levels: A, beginner; B, intermediate; C, advanced.
Table 6 provides explanations to some of the grading systems referenced here.

et al. 2006, Mendes et al. 2016). However, it is increasingly more common to represent learner data as parallel corpora, where the original and corrected versions of the essays are stored in two separate files (e.g. Lee, Li, et al. 2017, Rosen et al. 2020, Dargis et al. 2020, Arhar Holdt et al. 2024). Unlike most predecessors, the SweLL-gold corpus has been *pseudonymized* (not anonymized) – i.e. personal information in texts has been substituted by alternative strings to preserve the integrity of learner texts and to conform to the requirements of the GDPR (EU Commission 2016); *normalized*, i.e. rewritten to an alternative independent corrected version, and corrections were *correction annotated*[13] for the nature of the difference between the original and normalized strings, i.e. the labels describe the difference between the two text versions rather than the learner's original language (Rudebeck and Sundberg, 2024).

13 In the majority of other learner corpora this is called "error annotation"

Below, we describe each of the four steps of manual processing of the SweLL-gold learner essays.

2.3.2 *Transcription*

The transcription of hand-written second language essays into a digital format followed a standardized, multi-phase workflow to ensure consistency, accuracy, and preservation of the original content. Two core principles have guided the transcription process:

- *Preservation of authenticity:* Annotators must not correct errors made by the original author. The transcription must faithfully reflect the student's writing, even when errors were present.
- *Avoidance of assumptions:* When transcription uncertainties arose (e.g., ambiguous spacing between words where the correct version would be writing two words separately), annotators were instructed to apply a "positive assumption." This means assuming the learner's intention in the most favorable light, such as in the case above – treating two closely spaced words as separate items, unless clearly written as one.

The transcription guidelines were designed to maintain fidelity to the original texts. Key rules include:

- *Retention of errors:* Spelling and grammatical errors are to be preserved without correction. Spell checkers should be disabled to prevent inadvertent corrections.
- *Illegible writing:* Unreadable characters should be marked with a "$" symbol, ensuring that transcription reflects the limitations of the original manuscript.
- *Non-standard letters:* In cases where learners invent letters or use unconventional symbols, annotators should use the closest recognizable letter or mark the symbol with a "$" if it is completely unidentifiable.
- *Graphical elements:* While certain graphical elements, such as capitalization and paragraph breaks, are preserved, elements like strikethroughs, indentations, and underlines are omitted. Marginal comments are only transcribed if they are clearly part of the running text.

To enhance accuracy, annotators are advised to re-read each essay after completing the initial transcription. This re-reading allows for a better understanding of the student's handwriting and ensures no unintentional corrections have been introduced.

The entire transcription workflow comprised five distinct stages:

1. *Acquaintance with guidelines:* Annotators familiarized themselves with the transcription guidelines through theoretical study and practical exercises using sample essays. This phase was essential for understanding how to handle various transcription challenges.

2. *Transcription workshop:* A one-day workshop where annotators collaborated with researchers to transcribe actual essays. The workshop facilitated the resolution of uncertainties and subjective judgments, aiming to standardize decision-making across the team. It also promoted the development of a support network for future cross-consultations.

3. *Individual transcription:* During this phase, each annotator worked independently on their assigned set of essays, adhering to the established guidelines to ensure uniformity.

4. *Cross-consultation:* In cases where uncertainties arose during individual transcription, annotators were encouraged to consult with peers or researchers. This collaborative approach minimized discrepancies and ensured alignment with the project's objectives.

5. *Transcription check:* A third-party annotator performed random checks on the transcribed essays, providing an additional layer of quality control. This step helped identify any deviations from the transcription rules and ensured compliance with the project's standards.

Time required for transcription varied depending on several factors, including the student's proficiency level, handwriting legibility, and the complexity of the text. Early in the training process, transcription tended to take longer as transcribers became accustomed to the guidelines.

2.3.3 *Pseudonymization*

To ensure that the data collected for the project can be openly used in research while protecting participants' privacy, we developed a comprehensive data handling process. Personal information handling included decisions on metadata on learners, pseudonymization principles for their texts, and tools to support the pseudonymization process.

Throughout data collection and storage, the data had to be securely managed, and all personal identifiers—such as names, age, locations, and dates—must be localized, masked, and eventually replaced in the corpus. These identifiers could appear both in the metadata and the learners' texts.

The SweLL project implemented a cautious approach to metadata, providing essential research information about each learner while maintaining anonymity. This metadata (see Section 2.2) included details like the learner's gender, age in 5-year intervals, time spent in Sweden, education level, mother tongue, and languages spoken in different contexts. However, we excluded specific details such as exact birthdate, arrival date in Sweden, country of origin, nationality, and information about the educational institution where the essays were collected.

While de-identification through metadata generally helps, it was not always sufficient, as learner texts often contained not only personal information but also sensitive personal information. Pseudonymization in SweLL-gold addressed this by identifying personal information (e.g., "My name is Ali") and classifying it into predefined categories (e.g., "My name is [first_name]"). The first step involved manually marking text segments containing personal details, categorizing them as names, institutions, geographic data, transportation information, dates, and other sensitive details like disabilities, political views, or unique family relationships.

Once marked, these text strings were systematically replaced to create as "natural" a text flow as privacy and semantics allow. We applied two strategies: for some catogories - hard replacement with a predefined token (e.g. emails with "email@domain"). For other categories - substitution with an equivalent from the same category, using general resource lists (such as names of people, places, and institutions) for random selection of alternatives. Two versions were tested for substitution of geographical entities - (1) using another geo-name and (2) using a surrogate substitute, such as "B-city". The currently distributed version contains "B-city" replacements to avoid semantic inconsistencies, such as "I live in *Barcelona* where I can ski every day", where *Barcelona* obviously is inappropriate for the semantic context. During the annotation process, unique ID numbers were assigned to recurring entities within categories so that repeated items in the text were replaced consistently. Additionally, morphological information was attached to ensure the masked text retains the correct grammatical form.

In cases where the annotator was unsure how to categorize a text string, the original text was kept but marked with a placeholder. A distinction was made between information that must be replaced due to sensitivity and information that might be handled later. Figure 2 provides an overview of pseudonymization tags in the Swell-gold data. More details on pseudonymization is available in Megyesi et al. (2021).

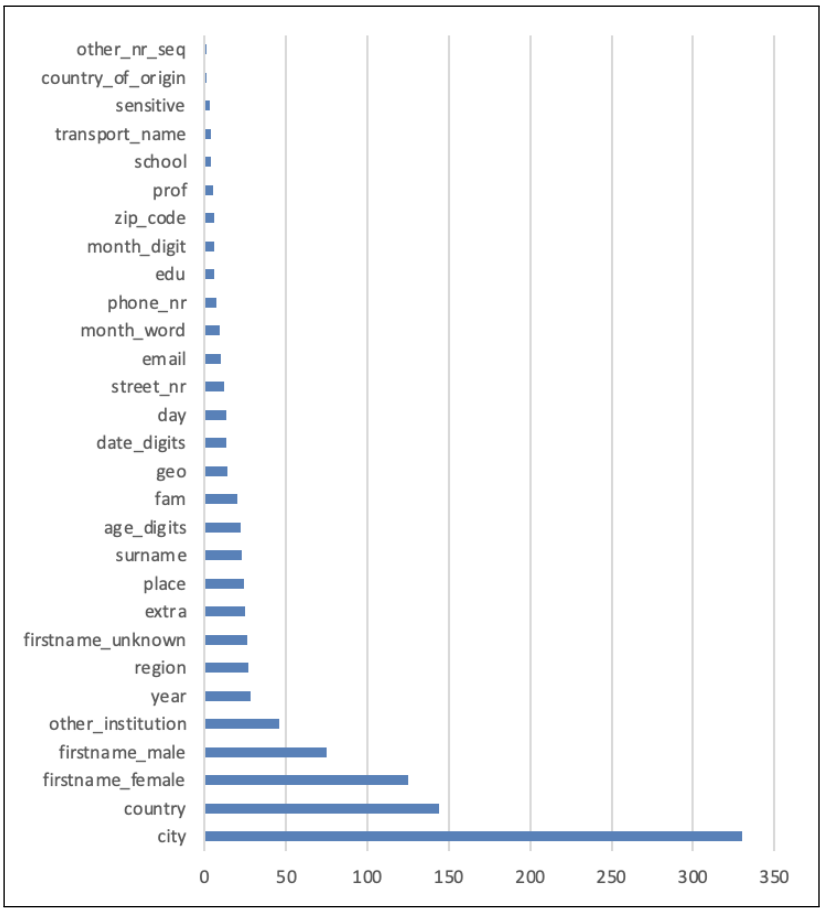


Figure 2: Overview of pseudonymization tags in the SweLL-gold data.

2.3.4 *Normalization*

The transcribed and pseudonomized learner texts were subsequently normalized. *Normalization* is the term which we have used in the SweLL project for the creation of a second, norm-adherent, version of the learner text.[14] The normalization had two primary purposes:

1. To provide a version of each learner text which is more amenable to automatic annotation using a standard linguistic analysis pipeline, optimized for "canonical" data in the form of texts adhering to standard norms (cf., Hirschmann et al. 2007, van Rooy 2015: 101).

14 The section on normalization is based on Rudebeck et al. (2021) and Rudebeck & Sundberg (2024). Some passages quoted from these sources are included.

2. To provide a basis for correction annotation through an explicit representation of the specific standard version of the text to which the original version of the text is related.

The normalization of each text was carried out through careful editing of the original (transcribed and pseudonomized) learner text. In practice, the normalization was made in the annotation tool SVALA (Wirén et al. 2019), allowing for it to be carried out as "normal text editing" similar to text editing in common word processing programs such as Word. This technical aspect of the normalization is closely intertwined with the general approach to normalization in SweLL-gold, which is based on a clear separation between normalization and correction annotation, as reflected by the separate sets of guidelines for normalization (Rudebeck et al. 2021) and correction annotation (Rudebeck & Sundberg 2021). While correction annotation means categorizing corrections according to a finite set of well-defined categories, the normalization process is open-ended and creative, guided by broad values and general methodological practices. It is a process which may be likened to translation. The theoretical foundation for and implications of this approach to normalization are throughly discussed in Rudebeck & Sundberg (2024).

The *broad values* guiding the normalization were that the resulting text version should both (1) adhere to the norms of written standard Swedish, and (2) be faithful to the original text. The second goal includes two different aspects: a normalized string should (a) be similar to the original text string, and (b) effectively communicate the content intended by the writer, according to the normalizer's interpretation (cf., Tenfjord et al. 2009: 60).

Concerning the first value (norm adherence), and the range of norms considered, the normalized text version should contain no obvious deviations from the norms of written standard Swedish at sentence level. However, the acceptance of unusual expressions was fairly high. The norm-related goal was not interpreted as to imply that the normalized texts should necessarily be "native-like". In fact, the normalized texts in SweLL-gold often give a learner-like impression, in spite of the lack of obvious norm deviations.

We dealt with all intra-sentence levels at once, and considered norms for spelling, morphology, punctuation, syntax, and the usage range of words and phrases. Norms concerning the composition of texts, beyond those which may be dealt with sentence-internally, were not considered. The normalization thus involved neither changes of the ordering of sentences or paragraphs, nor any deletions of whole sentences. However, the delimitation of sentences could be changed, for instance by exchanging a conjunction for a sentence-delimiting punctuation mark.

The fundamental values guiding the normalization are often conflicting. To start with, a conflict between norm adherence and fidelity to the original text is the very basis for editing the original text at all; only when the original text deviates from standard norms is a normalization called for. But the two aspects of fidelity to the original text (similarity to the original text string and effective communication of the writer's intended message) may also be conflicting, so that a greater similarity to the original text string may involve a less effective communication of the writer's intended message, as interpreted by the normalizer, and vice versa (see Rudebeck & Sundberg 2024 for a discussion of examples). The normalization process may thus be seen as an act of balancing these values.

To support the normalization process, a few *general methodological practices* were followed:

1. The normalization was carried out by a small number of *highly qualified annotators* (four in total), who all had expert knowledge of Swedish, but with complementary fields of expertise, including language structure, language norms, and Swedish as a second languge.

2. In an initial phase, problematic instances were *documented and tentatively categorized* as a basis for discussion in the project group.

3. Throughout the normalization process, particularly problematic instances were *discussed with a co-normalizer*.

4. The general approach to each learner text was *context sensitive* (a plausible interpretation relying on the context, e.g. Corder 1973: 247) in the sense that the text was interpreted and understood in relation to the task (cf., Callies 2015: 49) and to the learner's text as a whole.

These general practices were upheld in order to enhance the quality of the normalized texts as well-balanced "translations", which could both pass as samples of written standard Swedish, and as faithful representations of the original learner texts. Moreover, these practices served to improve the consistency of the normalized texts at the level of a general understanding of "norm-adherence," and in weighing the guiding values against each other.

The SweLL-gold approach to normalization, described above, differs from previously described approaches, according to which "target hypotheses" (normalized versions of expressions in the learner texts) should be chosen in a rule-based fashion, with the aim of achieving a replicable procedure where "analogous cases" of errors should be consistently categorized (Lüdeling & Hirschmann 2015: 144). In SweLL-gold, categorical consistency of this kind has been essential to correction annotation, but has not guided

the normalization process, since the validity of the correction annotation requires that the normalization step is based on an interpretation of the writer's intentions, and on careful consideration of contextual factors.

2.3.5 *Correction annotation*

Once a normalized version of a learner text has been finalized, it may be compared to the original version of the text. Any difference between the two versions is a *correction*. In SweLL-gold, these corrections have been *correction annotated*, that is, categorized according to a set of correction categories.

The purpose of the correction annotation is to make the texts searchable for specific kinds of "errors". But an "error" in a learner text can only be categorized on the basis of an assumed "correct", or "reconstructed" version of the text segment (Corder 1971: 155). By the choice of the term "correction annotation" (rather than the more common term "error annotation") the necessarily comparative nature of this kind of analysis is made explicit (Rudebeck & Sundberg 2024).

The correction annotation in SweLL-gold is thus a categorization of differences between the original learner texts and their corresponding normalized versions, and only indirectly indicates properties of the original learner texts. This means that the correction annotation is highly dependent on the preceding normalization, and that each correction label indicates a difference between the writer's choice of expression and an adjusted, norm-adhering, way of formulating the message intended by the writer, as interpreted by the normalizer.

The SweLL-gold correction categories, and principles for their application, are extensively described in Rudebeck & Sundberg (2021).[15] Here follows a summary of the major categories and principles.

A correction may consist of:

- an addition of a unit; the unit is only present in the normalized text,
- a deletion of a unit; the unit is only present in the original text,
- a movement of a unit; the unit is present both in the original text and the normalized text, but it is placed differently relative to the surrounding text,
- a change of a unit; a unit in the normalized text is a corrected version of a corresponding unit in the original text.

15 The section on correction annotation contains some passages quoted from this source.

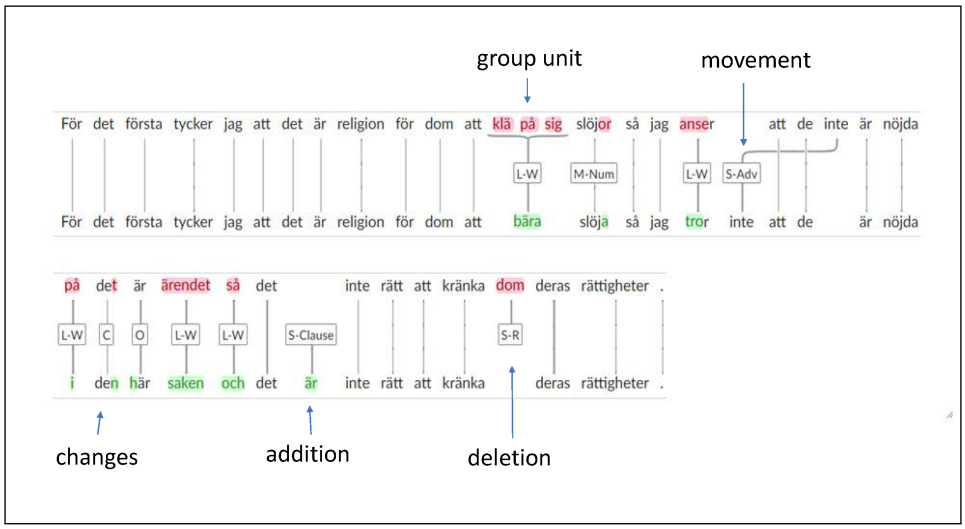


Figure 3: Types of correction and unit types in the SVALA annotation tool. The top level represents an original learner text, the lower level represents its normalized version.

A unit normally consists of one token (a token unit), but it may also consist of a group of tokens (a group unit). These correction types and unit types are illustrated in Figure 3. The correction of one single unit may reflect several correction categories, and the same unit may therefore be annotated with several correction tags.

The SweLL correction annotation taxonomy contains five main categories, with a number of sub-categories within each of them (33 in total):

- Orthographic corrections (O): regular spelling corrections, corrections between upper and lower case, and corrections of the use of spaces and hyphens between words.
- Lexical corrections (L): (1) corrections of the choice of word and (2) corrections of the internal morphological structure of word stems – i.e. the formation of words through compounding and derivation.
- Morphological corrections (M): corrections relating to inflectional morphology. This category also covers some extra-morphological corrections which are closely related to inflectional forms and grammatical categories involving inflections.
- Punctuation corrections (P): corrections of the choice of punctuation marks as well as the adding or removal of punctuation marks, and also instances of merging or splitting sentences.

- Syntactical corrections (S): corrections regarding the structure of multiword phrases and clauses, including corrections of missing or redundant words, word order, the choice between a compound and a multiword expression, and more complex syntactic corrections such as corrections between a finite clause structure and an infinitive phrase.

The correction annotation in SweLL-gold also contains a specific tag for corrections made as a consequence of other corrections (C), and a tag for unintelligible strings (X). The latter sometimes represents a correction, since a "guess" as to the intended meaning has been provided in the normalized version. In some cases unintelligible strings marked with the X tag have been left unchanged in the normalized version. In addition, there are three tags which are included in the correction annotation, although they are never associated with corrections: a tag for strings cited from a foreign language and thus not corrected (Cit-FL), as well as two tags for notes and comments (OBS! for internal work notes, and Com! for comments intended for corpus users).

A consequence of the fundamental principle of annotating corrections, understood as differences between the original text and one specific interpretation of this text (the normalized text version), is that certain clear deviations from the norms of written standard Swedish in the original texts are left without annotations – because they are not deviations in relation to the normalized text. This occurs for instance when a misspelled word in the original text has been corrected to another word altogether. Such a correction will be annotated as an instance of a corrected choice of word (L-W), and since the word in the original text cannot be analyzed as a misspelling of the word in the normalized text, the spelling error will be left without annotation. For instance, the string *monga nognting* may be interpreted as a misspelt version of the string *många någonting* 'many something,' but was normalized as *många saker* 'many things.' The link between the original token *nognting* and the normalized token *saker* was then annotated with the label L-W, reflecting the exchange of word made in the normalization, but not the faulty spelling of the original string. Table 8 shows the correction tags along with their meanings and frequencies in SweLL-gold.

The SweLL-gold correction taxonomy was developed through a series of pilots and workshops within the project group (Volodina et al. 2019). Initially, the project group looked at correction taxonomies developed for other learner corpora, and the taxonomies of ASK (Tenfjord et al. 2006) and Merlin (Boyd et al. 2014) were selected as a starting point. The final SweLL-gold taxonomy is clearly influenced by the ASK taxonomy, but includes a few more specific correction categories meant to capture common norm deviations

Table 8: The correction taxonomy: main types, specific correction tags, and frequencies in SweLL-gold. For further details, see Rudebeck & Sundberg (2021).

	Tag	Correction concerns	N
Ortography	O	Orthography (spelling)	3 269
4 381	O-Cap	Upper/lower case	566
	O-Comp	Spaces and hyphens within compounds and between words	546
Lexicon	L-Der	Word formation (compounding or derivation)	721
4 876	L-FL	Non-Swedish word corrected to Swedish word	89
	L-Ref	Choice of anaphoric expression	624
	L-W	Word choice (other)	3 442
Morphology	M-Adj/adv	Adjective form corrected to adverb fom	104
8 005	M-Case	Nominative vs genitive/accusative	257
	M-Def	Definiteness: articles; forms of nouns, adjectives	2 968
	M-F	Form corrected, same category (e.g. one plural suffix corrected to another plural suffix)	300
	M-Gend	Gender	953
	M-Num	Number	1 031
	M-Other	Other inflection-related correction	116
	M-Verb	Verb forms; auxiliaries	2 276
Syntax	S-Adv	Word order: adverbial placement	785
7 696	S-Clause	Basic clause structure	1 129
	S-Comp	Compound versus multi-word expression	136
	S-Ext	Extensive and complex syntactical correction	310
	S-FinV	Word order: placement of finite verb	701
	S-M	Word missing (added in target), other	1 904
	S-Msubj	Subject missing (added in target)	434
	S-Other	Other syntactical correction	80
	S-R	Word redundant (removed from target)	1 423
	S-Type	Change of phrase type/part of speech	595
	S-WO	Word order (other)	199
Punctuation	P-M	Punctuation missing (added in target)	1 834
2 754	P-R	Punctuation redundant (removed from target)	444
	P-Sent	Sentence segmentation	39
	P-W	Wrong (i.e. changed) punctuation	437
Other	C	Consistency correction, necessitated by other correction	1 208
1 399	X	Unintelligible string	137
	Cit-FL	Non-Swedish word kept (i.e. no correction)	54
Sum			29 111

in learner language, such as adverbial placement and omission of subjects. The current correction taxonomy has been evaluated through a comparison between annotations made independently by two different annotators, which resulted in Inter-Annotator Agreement of 88% by Fleiss's kappa and 76% by Krippendorff's alpha (Artstein & Poesio 2008, Krippendorff 2019) as measured on 10% of the essays (i.e. 50 essays).

2.3.6 *Automatic annotation*

The *automatic* linguistic annotation present in the SweLL-gold corpus comes from the Sparv annotation pipeline (Hammarstedt et al. 2022) and contains tokenization, lemmatization, word sense disambiguation, morpho-syntactic annotation, syntactic dependencies and markup of a few other linguistic features. The general reliability of the Sparv automatic annotation is high on standard texts, i.e. texts written by native speakers (Adesam & Berdicevskis 2021, Volodina et al. 2022). When it comes to language learners, Volodina et al. (2022) have shown that part of speech (POS) tagging, lemmatization and word sense disambiguation (WSD) stay similarly high. However, Volodina et al. (2022) have also shown that syntactic annotation and markup of multiword expressions drop significantly in performance when applied to learner language. These insights should be taken into account when using automatic linguistic annotation for theoretical generalizations and conclusions about language learning.

The added linguistic annotation makes it possible to search the corpus through the corpus management system Korp (Borin et al. 2012), building searches as a combination of metadata, manually added markup, automatic linguistic filters, and tokens, e.g. a search in Korp for the `target form` *mycket* 'a lot' followed by a `part of speech` *noun*, and sorted by `approximate level`.[16]

2.3.7 *SweLL contributions*

With regards to the manual processing and annotation practices in Learner Corpus Research, the SweLL project has contributed to:

1. increased attention to the need for structured pseudonymization of learner essays (Megyesi et al. 2018, 2021, Volodina et al. 2020, Volodina, Dobnik, et al. 2023);
2. an emerging new paradigm of learner corpora where the original and normalized versions are treated as parallel corpora (Wirén et al. 2019, Arhar Holdt et al. 2024, Rudebeck & Sundberg 2024);

16 To get access to the search,you have to be an approved SweLL user.

```
Essay ID: K44KT5 Metadata: [...] age="21-25" [...] l1="Tigrinska"
[...] edu_level="2" [...] school_type="Vuxenutbildningen" [...]
task_subject="Argumenterande text om slöjor eller krav för medborgarskap"
[...] Source: [...] För det första tycker jag att det är religion för dom att
klä på sig slöjor så jag anser att de inte är nöjda på det är ärendet så det
inte rätt att kränka dom deras rättigheter . [...]
```

Figure 4: Excerpt from an original essay in raw text format. Raw files include metadata and minimally preprocessed (tokenized and pseudonymized) text.

3. a greater focus on normalization, as a step of analysis which is separate from correction annotation, requiring a more context-sensitive and interpretation-based approach (Rudebeck & Sundberg 2024);

4. shifting the focus from "errors" in learner versions to their "corrections" since these corrections are only some of several possible hypothetical ways to interpret (errors in) learner writing (Rudebeck & Sundberg 2024).

2.4 *Access to the data and released formats*

The SweLL-gold corpus contains private information – both in the form of metadata and as private mentions in texts, and therefore falls under the GDPR (EU Commission 2016). This sets limitations to the openness of data, namely, that only individuals living and working in EU can have access to the data; with a further restriction that the area of application should be connected to education (teaching, learning, research or development). Due to that, SweLL-gold is distributed in pseudonymized form only and access is administered through an application form.[17,18] Recently, a new restriction has been added to the SweLL data usage, namely, that no API-based third-party Large Language Models (LLMs) are allowed for processing SweLL, with the aim of preventing these learner texts from being used to retrain LLMs.

To satisfy different user needs, the approved SweLL user gets access to the data in two ways: through a corpus search system Korp[19] (Borin et al. 2012), as well as through downloading the data in three file formats:

- raw text data (separately for original and normalized versions) (cf. Figure 4);

17 Application: https://sunet.artologik.net/gu/swell
18 DOI for SweLL-gold: https://doi.org/10.23695/2k47-y432
19 https://spraakbanken.gu.se/korp/

- JSON file format (linked original and normalized versions) where correction labels are included (cf. Figure 5); and
- TEI-compliant XML files containing correction labels and automatically added linguistic annotation (separately for original and normalized versions; cf. Figure 6).

Formats for distribution of learner corpora are largely influenced by the way annotation is conceptualized, such as whether to treat the corrected version as an independent text, or to attach a corrected string directly into the original sentence. However, even the search interfaces set limitations on formats, most prominently, corpus workbench depending heavily on TEI-XML. Most error-annotated corpora are, therefore, distributed in XML-based file formats with only a few distributed alternatively also in JSON format (e.g. Arhar Holdt & Kosem 2024).

SweLL-gold is the first and the only correction-annotated L2 learner corpus of Swedish.

3 *Use case 1: Corpus searches for second language acquisition*

> Given original learner essays and their corrections, we want to explore whether the use of verb constructions with the motion verbs *komma, gå, åka* 'come', 'go', 'move by vehicle' change over learners' proficiency levels in a way that might be indicative of development towards (a) a more accurate use and (b) a more abstract use of the investigated constructions that goes beyond the descriptions of concrete motion events (Based on a study by Sundberg & Prentice 2023).

The first case study illustrates how data from the learner corpus can be used for the study of Swedish as a second language, more specifically the development and use of *motion verbs* like 'go' (*gå*), 'come' (*komma*), 'walk' (*promenera*) and 'travel'/'go by vehicle' (*åka*) in texts by learners at different levels of acquisition (see Sundberg & Prentice 2023). There are several reasons why motion verbs are a relevant object of study.

Firstly, the acquisition of verb constructions constitutes a crucial aspect of language learning since they form the core of a sentence and thus are drivers in the learner's grammatical development.

Secondly, we also know that some verbs of motion, such as 'go' and 'come', are among the most common *nuclear lexical verbs* in a large number of languages; they are needed to express basic actions and thus appear and can be analyzed at all linguistic levels. For instance, in a previous corpus-based

```
{ [...],
  "K44KT5": {
    "Metadata": {[...], "age": "21-25", [...], "l1": "Tigrinska", [...],
        "edu_level": "2", [...], "school_type": "Vuxenutbildningen",
        "task_subject": "Argumenterande text om slöjor eller krav för
        medborgarskap", [...]},
    "Source": "[...] För det första tycker jag att det är religion för dom
        att klä på sig slöjor så jag anser att de inte är nöjda på det är
        ärendet så det inte rätt att kränka dom deras rättigheter . [...]",
    "Target": "[...] För det första tycker jag att det är religion för dom
        att bära slöja så jag tror inte att de är nöjda i den här saken och
        det är inte rätt att kränka deras rättigheter . [...]"
  }
}
```

Figure 5: Excerpt from SweLL-gold in JSON format. In this case, original (source) and normalized (target) essays are stored in a single file.

```
<text essay_id="K44KT5" [...] age="21-25" [...] l1="Tigrinska" [...]
    edu_level="2" [...] school_type="Vuxenutbildningen" [...]
    task_subject="Argumenterande text om slöjor eller krav för
    medborgarskap" [...]>
  [...]
  <link id="6398">
    <sentence>
      [...]
      <w ref="2">För</w> <w ref="3">det</w> <w ref="4">första</w>
      <w ref="5">tycker</w> <w ref="6">jag</w> <w ref="7">att</w>
      <w ref="8">det</w> <w ref="9">är</w> <w ref="10">religion</w>
      <w ref="11">för</w> <w ref="12">dom</w> <w ref="13">att</w>
      <w ref="14" target_form="bära" correction_label="L-W">klä</w>
      <w ref="15" target_form="bära" correction_label="L-W:2">på</w>
      <w ref="16" target_form="bära" correction_label="L-W:3">sig</w>
      <w ref="17" target_form="slöja" correction_label="M-Num">slöjor</w>
      <w ref="18">så</w>
      [...]
    </sentence>
  </link>
  [...]
</text>
```

Figure 6: Excerpt from an original essay in XML format. Each essay is segmented into <sentence> elements, which are in turn tokenized into <w> elements. Target forms and correction annotations are associated with individual tokens. For the sake of compactness, automatic linguistic annotation is omitted here.

study of verb learning in over 68 000 English learner texts at different levels, the verb 'go' is the fourth most common verb lemma, preceded by 'be', 'have' and 'do' (Römer 2019).

Thirdly, the number of correction tags used in the normalized version of the corpus might index fields of particular interest for second language researchers. To mark discrepancies and deviations between the learner texts and the reconstructed text versions, 31 different tags were used as subcategories of 6 general categories. The single tag most used belonged to deviations in vocabulary or lexicon, where the subcategory "Lexicon Wrong" (L-W), meaning an incorrect use of a word or a phrase, was used 3 442 times (see Table 8 on page 271). An in-depth analysis of a common lexical acquisition problem would therefore be relevant.

The development and acquisition of the Swedish verb lexicon, among them motion verbs, has previously been studied in learners' spoken language (Kotsinas 1982, Viberg 1992, 2004). They show a possible over- and underuse of more specific motion verbs in learners' Swedish, and a related question is whether the same pattern can be demonstrated in the written language at group level. Römer (2019) shows how verb-argument constructions (VAC) become increasingly varied, abstract and productive at higher learner levels in English learner texts.

An example from the SweLL-gold corpus may illustrate a typical deviation and overuse of *gå* in a text at beginner and advanced level respectively:

(a) *…ibland jag läsa en bok eller gå i skogen i närheten.*
'…sometimes I read a book or go to a forest nearby.' (beginner)

(b) *…storebor sager till lillebror att man går till Nagijala när man dör.*
'…big brother says to little brother that you go to Nagijala when you die.' (advanced)

These examples show that the motion verb *gå* 'go' in the Swedish language has a more limited semantic scope than in many other languages. In the normalized version, *gå* and *går* has been substituted by Swedish *promenera* 'walk' and *kommer* 'come', and thus has received the correction label L-W in the SweLL-gold corpus.

In a cross-sectional case study, Sundberg & Prentice (2023) used different approaches to the study of motion verb constructions at group level in SweLL-gold. The study was based on 502 written texts and 321 different learners categorized as beginners, intermediate and advanced. These categories are included in SweLL-gold metadata based on course level (see Section 2.2.6).

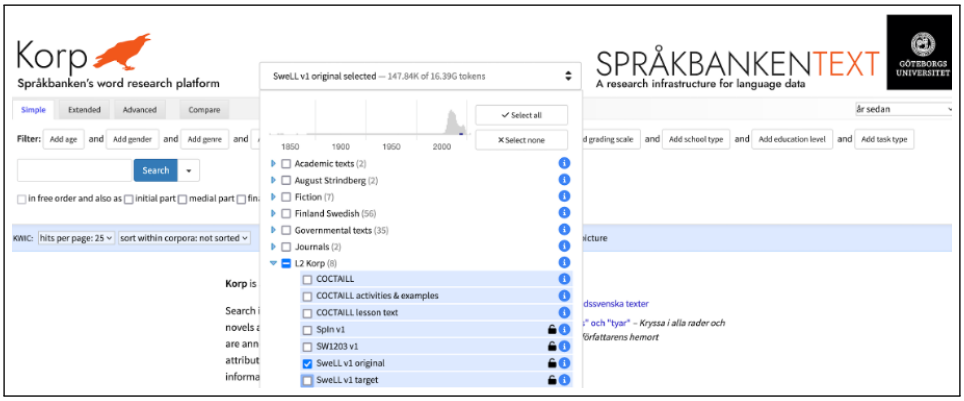


Figure 7: Corpus selection in Korp user interface.

3.1 *The SLA case study step-by-step*

An important starting point is to study both the learners' unproblematic use of motion verbs in the texts and to study deviations (Thewissen 2013). In this respect, both the original text and the reconstructed text versions play important roles for different types of analyses. Based on the data and metadata in the SweLL-gold corpus, one can, for example, describe the use of motion verb constructions with *gå*, *komma* and *åka*:

1. in general, at different learner levels,
2. in relation to metadata about the writing task and textual factors, such as text types, or
3. in relation to metadata about the learners, such as linguistic or educational background.

For the case study described here, we analyse data based on the first of these aspects, that is, we compare the use of the motion verb constructions at the ascribed language learner level, that is based on the course level of the learner. To do so, we take the following five steps:

Step 1: Choose "SweLL v1 original"[20] (under "L2 Korp")[21] among the corpora available through Korp[22] (Figure 7).

20 Note that you need to be an approved SweLL user to be able to search the SweLL corpora in Korp. Please, apply using https://sunet.artologik.net/gu/swell. More information on data access is provided in Section 2.4

21 Note also the use of the terms *SweLL original* and *SweLL target* with reference to the two versions of the SweLL-gold corpus. For sake of space, the *-gold* part has been left out in the corpus search platform Korp, which we try to reflect in the examples here.

22 https://spraakbanken.gu.se/korp/

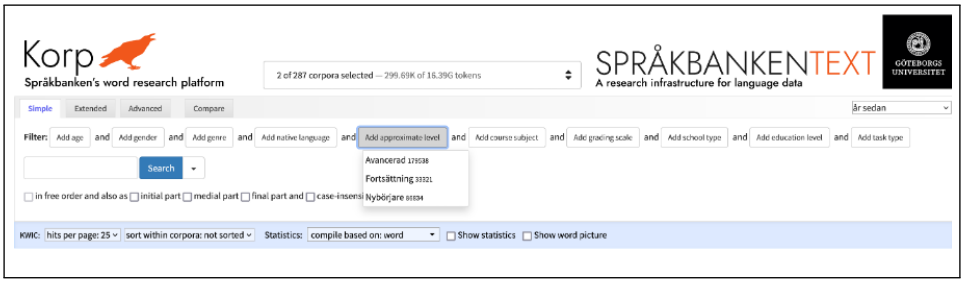


Figure 8: L2-specific filter in Korp user interface.

Table 9: Relative frequency for constructions with *komma, gå* and *åka* in SweLL-gold (Sundberg & Prentice 2023: 441). Raw frequency in brackets.

	Beginner		Intermediate		Advanced		Total	
komma 'come'	48.6	(207)	47.9	(78)	37.4	(333)	41.8	(618)
gå 'go'	63.2	(269)	64.5	(105)	17.3	(154)	35.7	(528)
åka 'move by vehicle'	19.7	(84)	14.1	(23)	1.2	(11)	8.0	(118)

Step 2: Conduct searches for lemmas *gå, komma, åka* in different sub corpora, i.e. Beginner, Intermediate and Advanced, by using the filter function in Korp. In this case we are interested in texts written by learners at three different "approximate levels" (Figure 8)

The results of the queries and the frequency of *komma*, *gå* and *åka* in the different subcorpora (Beginner, Intermediate and Advanced) are compared in Table 9. The relative frequency is reported per 10,000 tokens (raw frequency in brackets).

As Table 9 shows, the frequency level for *komma* is relatively similar in the different sub corpora, whereas both *gå* and *åka* are clearly more frequent at the beginners' level than in the texts written by advanced learners (Sundberg & Prentice 2023: 441). All differences between the beginners and advanced learners are statistically significant. These results are partly in line with previous research, indicating overuse of certain motion verbs at earlier stages of L2 development (e.g. Viberg 2004). However, we want to look more closely at the constructions behind these numbers, since especially *komma* and *gå* can be used in a variety of different constructions, not necessarily in the meaning of concrete motion events (Sundberg & Prentice 2023: 441).

Step 3: As mentioned initially, we are also interested in problematic versus unproblematic uses (see "Correct use" column in Table 10) of the verbs. Therefore, the next step is to compare the search results for the three verbs in

Table 10: Absolute frequency uses of *gå*, *komma* and *åka* in SweLL-gold original and SweLL-gold target, and % uncorrected uses at beginner and advanced level.

	Occurences		Correct use (%)	
	Original	Target	Beginner	Advanced
gå	414	374	68	84
komma	528	619	77	87
åka	107	144	77	100

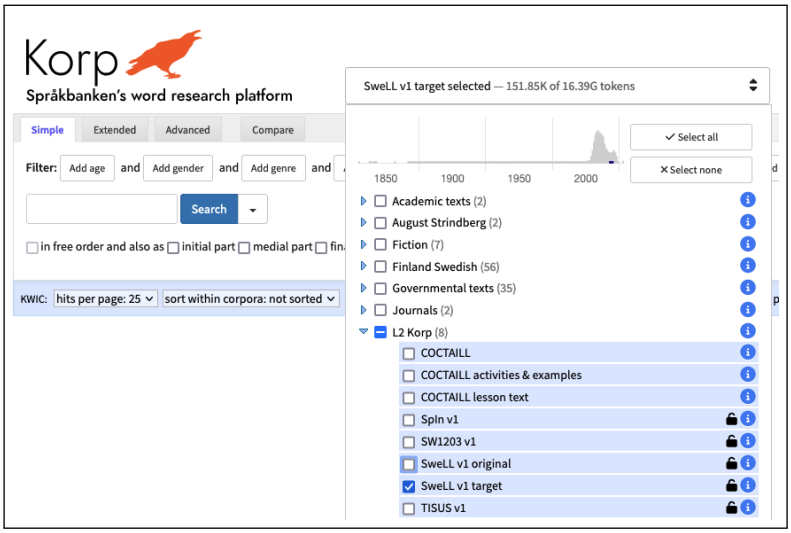


Figure 9: Select "SweLL v1 target" to compare frequency of use for the three verbs in normalized texts to usage in original texts.

"SweLL v1 original" with "SweLL v1 target", both with regard to corrected vs. uncorrected uses and different kinds of correction-labels. We, therefore, select "SweLL target" for this search (Figure 9).

Table 10 shows that the absolute frequency (i.e. raw count) of *gå* is slightly higher in the source texts in SweLL original than in the normalized versions in SweLL target, whereas the opposite is the case for *komma* and *åka*. In addition, we see that the percentage of uncorrected uses ("Correct use" column) in SweLL original increases between beginner and advanced levels for all three verbs (by 18% for *gå*, 10% for *komma* and 23% for *åka*).

Step 4: To learn more about non-standard usage we also look at the correction annotation of the three verbs. The first inquiry is about which type of corrections have been made regarding the learners' use of the verbs. To find

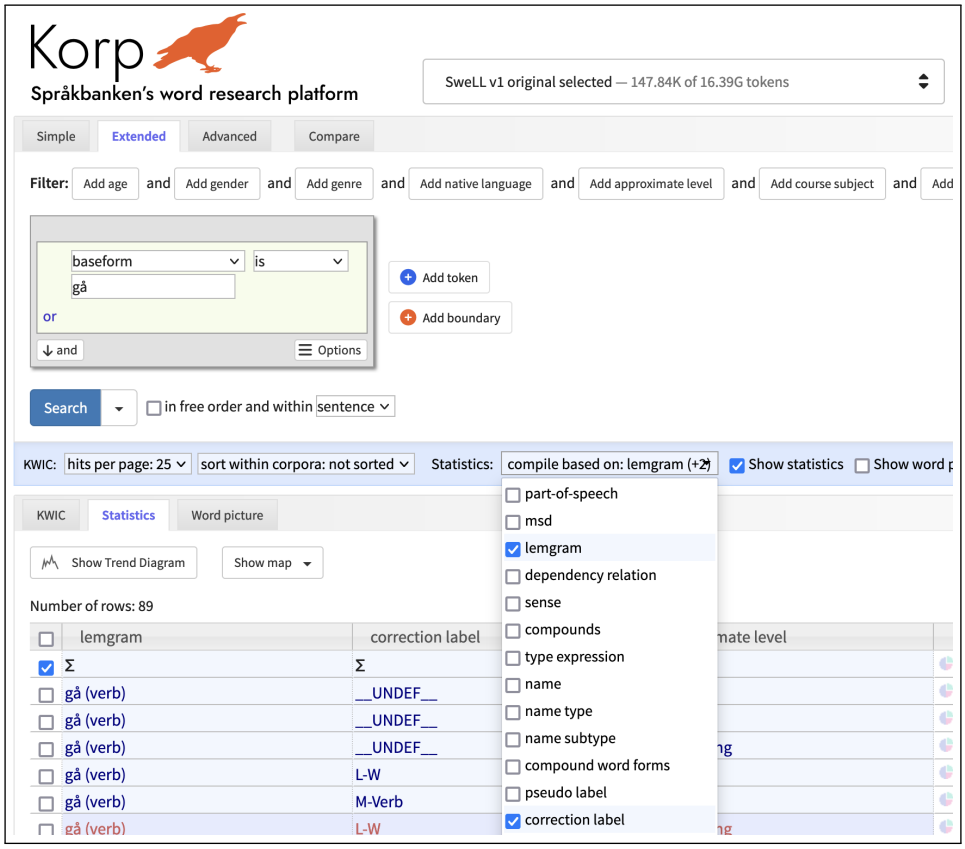


Figure 10: Define a search for base forms and compile statistics for *lemgram* (a combination of baseform and part of speech), correction label and approximate level.

that information in the corpus, we choose SweLL original (Figure 10), then search for the base form of the verbs in the extended search and compile the statistics based on lemgram, correction label and approximate level.

Figure 11 shows how frequently corrections made to constructions containing *gå*, *komma* and *åka* have been labeled as L-W ('wrong word'), M-Verb ('wrong verb form') or S-finV ('problem with finite verb placement'). These corrections would be relevant for a more in-depth analysis from both an SLA and a pedagogical perspective, since they tell us something about the use of the three verbs in the learner's texts in relation to common challenges for language learners (of Swedish), i.e. scope and depth in vocabulary, verb morphology and word order. An example of an L-W correction annotation in SVALA is shown in Figure 12.

The most obvious difference between the correction annotation of the

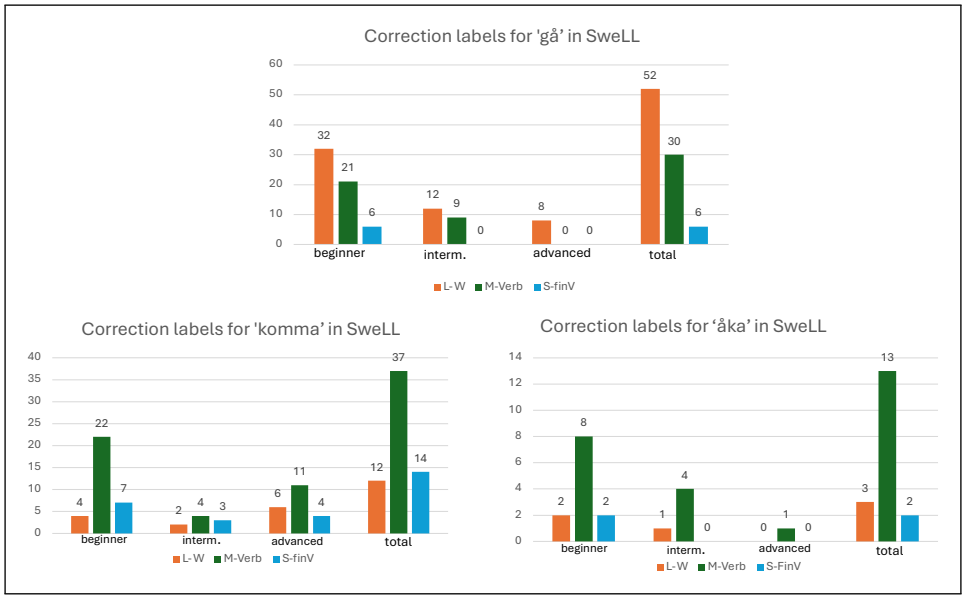


Figure 11: Frequency of correction labels for *gå*, *komma* and *åka* in SweLL-gold original at the three learner levels.

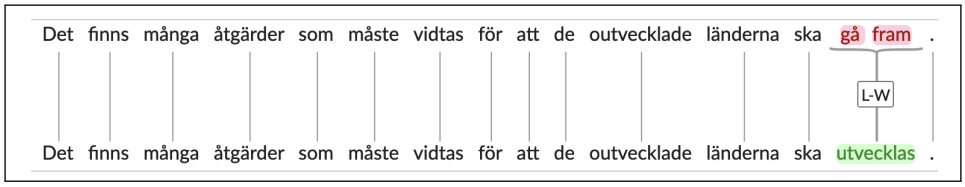


Figure 12: Target hypothesis *utvecklas* 'evolve' for *gå fram* 'go forward' in SweLL-gold (Advanced level).

three different verbs is, as shown in Figure 11, that the category L-W 'wrong word' is the most common one for *gå*, whereas M-Verb is the most frequently used correction label for both *komma* and *åka* (Sundberg & Prentice 2023: 442). This is not surprising, assuming a less developed vocabulary at beginner level. It is also in line with previous research, indicating overuse of verbs like *gå* in spoken learner language (Kotsinas 1982, Viberg 2004), even if we cannot draw generalizable conclusions from the limited data reported here.

Step 5: It is, nevertheless, worth taking a closer look at the target hypotheses, i.e. the words that *gå* have been replaced with in SweLL target, to learn more about which kind of verb constructions are affected by a possible overuse of the verb at different proficiency levels. This can be done by conducting a search of the base form *gå* in SweLL original (using the extended

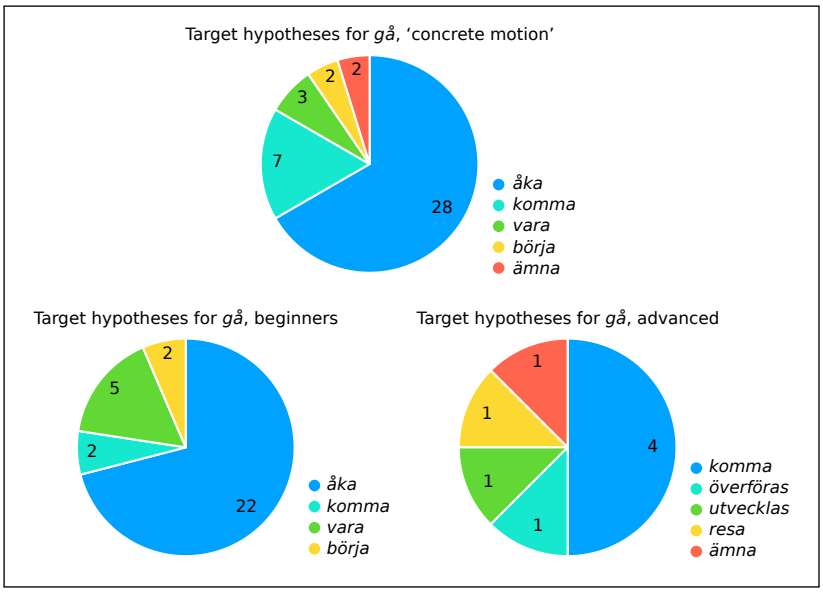


Figure 13: Target hypotheses for *gå* in SweLL target.

search mode) and compiling the statistics by lemgram, correction label and approximate level (cf. Figures 7 and 10). We have summarized the results of this search in Figure 13 (see also Sundberg & Prentice 2023: 443).

They show that *gå*, when used in verb constructions which express concrete motion is mostly replaced by *åka* and that most of these corrections have been made at beginner level (22 of 28 in total). Other target hypotheses are *komma* 'come', *vara* 'be', *börja* 'begin' and *lämna* 'leave'. If we look at the beginner level, *gå* has been replaced by *åka*, *komma*, *vara* and *börja*, which indicates (a) a possible overuse of *gå* related to a possible underuse of *åka* and (b) that the motion constructions in which *gå* occurs at beginner level are relatively concrete, event related constructions. To confirm this, we would need to do a qualitative analysis of the constructions and involve more metadata about the texts (e.g. task type) in the analysis. Regarding the target hypotheses for *gå* at advanced level, Figure 13 shows that few corrections have been made, which makes it difficult to draw any conclusions. The only replacement occurring more than once here (four times) is *komma*. Other verbs that occur are *överföra* 'transfer', *utvecklas* 'evolve', *resa* 'travel' and *lämna* 'leave'. Some of these verbs could be indicators of the emergence of more abstract constructions (cf. Römer 2019) with more figurative meanings of 'motion' as 'development' like in the example from the data illustrated in Figure 12.

The example translates into "There are many actions to be taken for the undeveloped countries to *go forward*". SweLL-gold can, however, only give some indications of what might be interesting to pursue. Due to the limited amount of data, we cannot actually draw any conclusions about development with regards to motion verb constructions between the levels at this point.

Summing up, the case study described here indicates certain tendencies which seem to be in line with previous research (e.g. Viberg 2004, Römer 2019). However, to confirm these indications one would have to look at larger amounts of data. In order to demonstrate individual variation and language development over time or seek explanations as to whether the use of motion verbs depends on, for example, the task, the type of text or the learner's first language, the description and analysis need to be refined with the help of metadata about both texts and learners (Sundberg & Prentice 2023: 443). Further, the case study shows the usefulness of a learner corpus like SweLL-gold within a well-developed infrastructure where a variety of different tools and interfaces for analysis can be cross-referenced.

4 *Use case 2: Source data for natural language processing*

Section 3 has demonstrated that Korp provides SLA researchers with a convenient interface to SweLL-gold. The corpus, however, is just as valuable for Natural Language Processing (NLP), where direct access to the source corpus for further processing is often crucial. Since its initial release, in fact, SweLL-gold has been used as training and/or evaluation data for a variety of machine learning (ML) tasks. This often involved some degree of task-oriented preprocessing, sometimes fairly sophisticated, which has led to the creation of a number of derivative corpora and conversion scripts. In this section, we give an overview of the NLP tasks SweLL-gold has proven to be valuable for and provide guidance on how to use the dataset for each of them.

4.1 *Linguistic acceptability judgment*

Linguistic acceptability judgement is a binary classification task used for evaluating the linguistic knowledge of LLMs. The input is a single sentence that the model needs to categorize as acceptable (i.e. grammatically and/or semantically correct) or not.

Linguistic acceptabily corpora exist for a variety of languages (Warstadt et al. 2019, 2020, Trotta et al. 2021, Mikhailov et al. 2022, Jentoft & Samuel 2023, Someya et al. 2024, Bel, Punsola & Ruíz-Fernández 2024, Bel, Punsola

& Ruiz-Fernández 2024). In their most basic form, they consist of sentences marked as either acceptable or unacceptable. The source of the sentences can vary: some corpora consist of synthetically generated linguistic examples, while others, like NoCoLA (Norwegian Corpus of Linguistic Acceptability, Jentoft & Samuel 2023), are collections of authentic sentences, sometimes extracted from learner corpora. Furthermore, some corpora, such as BLiMP (Benchmark of Linguistic Minimal Pairs, Warstadt et al. 2020), consist of *minimal confusion pairs* including a sentence with a single grammatical error and a corrected version of the same sentence. Often, these datasets are organized into subcorpora by target linguistic phenomenon. This allows for finer-grained analysis.

SweLL-gold is the main source for the Swedish Dataset for Linguistic Acceptability Judgments (DaLAJ)[23] (Volodina, Mohammed, et al. 2023, Volodina & Mohammed 2024), a linguistic acceptability corpus released as a part of the Superlim (Berdicevskis et al. 2023) natural language understanding benchmark. It consists of over 44 000 error-annotated sentences, most of which organized into minimal pairs. Minimally ungrammatical sentences are automatically obtained from original learner sentences with a process that involves using the correction annotations contained in the JSON version of the corpus (cf. Section 2.4) to isolate each individual grammatical error. Unlike the full SweLL-gold corpus, DaLAJ is released under the CC-BY 4.0 license and available for direct download as JSONL (JSON Lines; cf. Figure 14) and TSV (cf. Figure 15).[24]

4.2 *Grammatical error detection*

Another binary classification task SweLL-gold lends itself to is *Grammatical Error Detection* (GED). In its most basic variant, the task is similar to linguistic acceptability judgement, but performed at the level of individual tokens rather than sentences: given a sentence, the goal is to mark each of the words it consists of as either correct or incorrect. SweLL data was first used for this purpose in the context of the MultiGED-2023 Shared Task on Multilingual Grammatical Error Detection, which also provided data for Czech, English, German and Italian in a uniform TSV-based format (Volodina, Bryant, et al. 2023, Volodina, Bryant, et al. 2025)[25] (cf. Figure 16). The Swedish subset of MultiGED contains 8 553 sentences and 145 507 tokens.

23 DOI for DaLAJ: https://doi.org/10.23695/kxvz-tx42
24 DOI for DaLAJ-GED: https://doi.org/10.23695/kxvz-tx42
25 DOI for MultiGED: https://doi.org/10.23695/xe7r-k506

```
[...]
{"sentence": "För det första tycker jag att det är religion för dom att klä
    på sig slöja så jag tror inte att de är nöjda i den här saken och det är
    inte rätt att kränka deras rättigheter.", "label": "incorrect", "meta":
    {"error_span": {"start": 58, "stop": 68}, "confusion_pair":
    {"incorrect_span": "klä på sig", "correction": "bära"}, "error_label":
    "L", "education_level": "Fortsättning", "l1": "Tigrinska",
    "data_source": "Dalaj.v.2 -- SweLL gold"}}
{"sentence": "För det första tycker jag att det är religion för dom att bära
    slöjor så jag tror inte att de är nöjda i den här saken och det är inte
    rätt att kränka deras rättigheter.", "label": "incorrect", "meta":
    {"error_span": {"start": 63, "stop": 69}, "confusion_pair":
    {"incorrect_span": "slöjor", "correction": "slöja"}, "error_label": "M",
    "education_level": "Fortsättning", "l1": "Tigrinska", "data_source":
    "Dalaj.v.2 -- SweLL gold"}}
[...]
```

Figure 14: Excerpt from the training split of the DaLAJ dataset in JSONL format. The two training instances in this example are both obtained from the same original sentence.

```
sentence       label   error_span_start      error_span_stop
    confusion_pair_incorrect_spant confusion_pair_correction error_label
    education_level l1    data_source
[...]
För det första tycker jag att det är religion för dom att klä på sig slöja
    så jag tror inte att de är nöjda i den här saken och det är inte rätt
    att kränka deras rättigheter. incorrect 58 68 klä på sig bära   L
    Fortsättning   Tigrinska      Dalaj.v.2 -- SweLL gold
För det första tycker jag att det är religion för dom att bära slöjor så jag
    tror inte att de är nöjda i den här saken och det är inte rätt att
    kränka deras rättigheter. incorrect 63 69  slöjor slöja  M
    Fortsättning   Tigrinska      Dalaj.v.2 -- SweLL gold
[...]
```

Figure 15: Corresponding excerpt of the TSV version of the DaLAJ dataset.

All training and validation data for the task is publicly available on GitHub[26] and new systems can be automatically evaluated on CodaLab[27].

A derivative SweLL resource, MuClaGED (Multi-Class GED),[28] also exists for a variant of the task where tokens have to be assigned one of a set of pre-defined error labels (Casademont Moner & Volodina 2022b, Casademont Moner & Volodina 2025). In this case, the format is CoNLL-like, format:

26 https://github.com/spraakbanken/multiged-2023
27 https://codalab.lisn.upsaclay.fr/competitions/9784
28 DOI for MuClaGED: https://doi.org/10.23695/q9v4-vt57

```
[...]
det             c    'it'
är              c    'is'
religion        c    'religion'
för             c    'for'
dom             c    'them'
att             c    'to'
klä             i    'clothe'
på              i    'on'
sig             i    'themselves'
slöjor          i    'hijab'
[...]
```

Figure 16: Excerpt from the training split of the dataset used in the MultiGED-2023 shared task. Sentences are stored vertically and separated from each other by two newline characters. Each token is simply marked as correct (`c`) or incorrect (`i`). The data is organized in two columns. The translation column is added only for the sake of this article to help the reader undestand the example. For the sake of understanding – *klä på sig slöjor* is rewritten to *bära slöja* 'wear hijab' in the corrected version.

with five columns: 1. token ID (sequential); 2. word form as it occurs in the original learner text; 3. A (Addition); 4. D (Deletion); 5. R (Replacement). The last three columns are filled with one or more SweLL coarse-grained correction annotaton labels (cf. Section 2.3.5) based on the edit operation(s) performed to normalize the sentence (cf. Section 2.3.4). If, for instance, a noun is incorrectly inflected and misspelled, the R column of the corresponding noun contains the labels `M` (morphology) and `O` (orthography), while the A and D columns are left empty (see Figure 17).

4.3 *Grammatical error correction*

Beyond classification, SweLL-gold is suitable for *Grammatical Error Correction* (GEC), a sequence-to-sequence task whose goal is to rewrite ungrammatical text producing a correct version. What is considered "ungrammatical" can vary: corrections may include fluency edits or aim for mere formal correctness. Since SweLL-gold is minimally corrected, it can provide useful data for the latter case.

A SweLL-derived dataset, MultiGEC[29] (Masciolini, Caines, De Clercq, Kruijsbergen, Kurfalı, Muñoz Sánchez, Volodina, Östling, et al. 2025, Masciolini, Caines, De Clercq, Kruijsbergen, Kurfali, et al. 2025), has been used in the context of the MultiGEC-2025 Shared Task on Multilingual Gram-

29 DOI for MultiGEC: https://doi.org/10.23695/h9f5-8143

```
[...]
# text = Men det är så om man vill vinna då man måste jobba hårt för sitt
    mål . @
# sent id = 6
# metadata = Essay id = XXXX, Approximate level = Fortsättning, L1 = Tigrinska
1       Men     _       _       _       _       CCONJ
2       det     _       _       _       _       PRON
3       är      _       _       _       _       VERB
4       så      _       _       _       _       ADV
5       om      ['P']   _       _       ['P-M'] SCONJ
6       man     _       _       _       _       PRON
7       vill    _       _       _       _       AUX
8       vinna   _       _       _       _       VERB
9       då      ['P']   _       _       ['P-M'] SCONJ
10      man     _       _       _       _       PRON
11      måste   _       _       ['S']   ['S-FinV']     AUX
12      jobba   _       _       _       _       VERB
13      hårt    _       _       _       _       ADV
14      för     _       _       _       _       ADP
15      sitt    _       _       _       _       PRON
16      mål     _       _       _       _       NOUN
17      .       _       _       _       _       PUNCT
18      @       _       _       _       _       PUNCT

[...]
```

Figure 17: Excerpt from the training split of the MuClaGED dataset. Gloss: 'But it is so if you want to win then you must work hard for your goal.' Sentences are stored vertically and separated from each other by metadata information, sentence ID and the full sentence. The data is organized in five columns (as described in the text) plus two optional ones for detailed correction tags and part of speech tags.

matical Error Correction (Masciolini, Caines, De Clercq, Kruijsbergen, Kurfalı, Muñoz Sánchez, Volodina & Östling 2025),[30] which features Swedish as well as eleven other languages. Since the task consists of rewriting full essays – which, as mentioned in Section 2.4, cannot be freely redistributed – access to the full MultiGEC corpus requires a separate Terms of Use agreement.[31] SweLL-gold users interested in the Swedish subcorpus alone, however, can also convert the XML version of the source corpus into MultiGEC format through the `multigec.py` script, available as part of the SweLL-scripts repository, a growing collection of SweLL-related utilities.[32] `multigec.py` generates parallel original-normalized files in a simple

30 https://spraakbanken.gu.se/en/compsla/multigec-2025
31 https://lt3.ugent.be/resources/multigec-dataset
32 https://github.com/spraakbanken/SweLL-scripts

```
[...]
### essay_id = K44KT5
Förbjud inte slöjor på barn under 13 år!
[...]
För det första tycker jag att det är religion för dom att klä på sig slöjor
    så jag anser att de inte är nöjda på det är ärendet så det inte rätt att
    kränka dom deras rättigheter.
[...]
Tvinga inte dom! De får göra som de vill.


### essay_id = E6ET6
Familjen i samhället idag
[...]
```

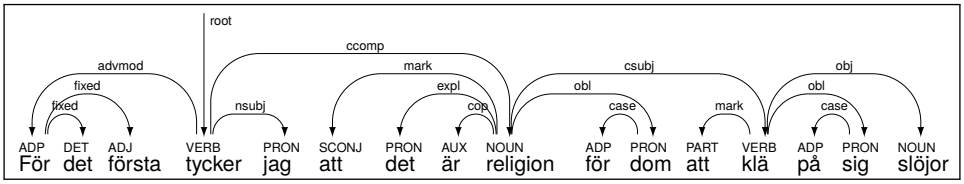
Figure 18: Excerpt from the test split of the dataset used in the MultiGEC-2025 shared task. Full essays are stripped from their metadata (except for the sentence ID) and stored in two separate Markdown files, one (partially reproduced here) containing original essays, and one containing their normalized versions.

Markdown-based format where texts are stripped from their metadata (cf. Figure 18).

It is also worth mentioning that, while the above mentioned shared task is at the essay level, GEC is traditionally performed on individual sentences. This poses a problem because in SweLL-gold learner sentences are not aligned with corrections and automatically extracting sentence-level correspondences is not always easy. This is due to the fact that normalization sometimes involves sentence re-segmentation, leading to one-to-many or many-to-one sentence correspondences in a number of essays. Several SweLL-derived datasets have been used in a variety of projects centered on developing and/or evaluating GEC systems for Swedish (Nyberg 2022, Ehnroth & Park 2023, Östling et al. 2024), but no comprehensive sentence-aligned dataset has been released at the time of writing. Over 5000 sentence pairs, however, can be obtained through the `extract_sentence_pairs.py` script – also available as part of the SweLL-scripts collection – which extracts them from trivially alignable essays. The latter can be used to output a TSV file where sentence-correction pairs are accompanied by their metadata and a list of labels relative to the errors encountered in the original sentence (cf. Figure 19). Alternatively, it is also possible to use the same script to generate two parallel CoNLL-U files where error labels are associated to individual tokens, just like in the source corpus.

```
essay_id      source student_id    nr_essay_student      age
    birthyear_interval    time_in_sweden gender l1     iso_l1 edu_level
    task_id task_date     datum  school_type    course_subject grading_scale
    task_subject  task_format    approximate_level     task_type
    text_types    svala_link     task_url       writing_language
    original sentence     corrected sentence    correction labels
[...]
K44KT5 _      _      _      21-25  _      _      _      Tigrinska
    _      _      _      _      Vuxenutbildningen     _      _
    Argumenterande text om slöjor eller krav för medborgarskap _     _
    _      _      _      För det första tycker jag att det är religion för
    dom att klä på sig slöjor så jag anser att de inte är nöjda på det är
    ärendet så det inte rätt att kränka dom deras rättigheter . För det
    första tycker jag att det är religion för dom att bära slöja så jag tror
    inte att de är nöjda i den här saken och det är inte rätt att kränka
    deras rättigheter . L-W,L-W:2,L-W:3,M-Num,L-W,S-Adv,L-W,C,O,L-W,L-W,S-R
[...]
```

Figure 19: Excerpt of a TSV file obtained with the `extract_sentence_pairs.py` script. For privacy reasons and for consistency with previous examples, some metadata have been replaced with underscores.



a. The dependency tree rendered as a graph.

```
1     För    för    ADP   PP    _     4     advmod _     _
2     det    en     DET   _     _     1     fixed  _     _
3     första en     ADJ   _     _     1     fixed  _     _
4     tycker tycka  VERB  _     _     0     root   _     _
5     jag    jag    PRON  _     _     4     nsubj  _     _
6     att    att    SCONJ _     _     9     mark   _     _
7     det    den    PRON  _     _     9     expl   _     _
8     är     vara   AUX   _     _     9     cop    _     _
9     religion      religion     NOUN  _     _     4     ccomp _     _
10    för    för    ADP   PP    _     11    case   _     _
11    dom    de     PRON  _     _     9     obl    _     _
12    att    att    PART  IE    _     13    mark   _     _
13    klä    klä    VERB  _     _     9     csubj  _     ErrorLabel=L-W
14    på     på     ADP   PP    _     15    case   _     ErrorLabel=L-W:2
15    sig    sig    PRON  _     _     13    obl    _     ErrorLabel=L-W:3
16    slöjor slöja  NOUN  _     _     13    obj    _     ErrorLabel=M-Num
```

b. The underlying CoNNL-U format. Fine-grained error labels are stored in the last column. For compactness, however, most other optional fields are left blank here.

Figure 20: Fragment of an original learner sentence annotated in UD.

4.4 *L2 parsing*

Syntactically annotated learner data can greatly benefit SLA research, as it makes it possible to look up and automatically extract syntactic patterns (Masciolini et al. 2023). Uses of such patterns include, but are not limited to, CEFR level identification (by matching them against L2 profiles), error retrieval and controlled feedback comment generation. For this reason, recent years have seen the development of numerous resources of this type, with Universal Dependencies (UD) (de Marneffe et al. 2021) becoming one of the most popular annotation standards (Lee, Leung, et al. 2017, Berzak et al. 2016, Kyle et al. 2022, Di Nuovo et al. 2022, Sung & Shin 2024, Rozovskaya 2024). UD annotation, exemplified on a SweLL sentence in Figure 20, is a time consuming task requiring extensive annotator training. In the case of learner data, this is brought to an extreme by the need to extend the general guidelines to cover and consistently deal with ungrammatical sentences (e.g. Berzak et al. 2016). This is, however, a necessary step towards reliable automatic annotation of learner language, something that multiple studies – one for all Volodina et al. (2022), conducted on SweLL data – have shown to be challenging for currently available off-the-shelf systems.

SweLL-gold has therefore become the object of a two-stage treebanking project, whose goals are: (1) to provide a high-quality, manually annotated test set for parser evaluation[33] and (2) to gradually extend the treebank up to training scale.[34] In line with previous work (Lee, Leung, et al. 2017, Berzak et al. 2016, Di Nuovo et al. 2022, Rozovskaya 2024), the treebank will include UD annotations not only of the learner productions, but also of their normalized versions, thus allowing comparisons between the two with tools such as STUnD, presented in Masciolini, Lange, et al. (2025; Chapter 14 in this handbook).

The full resource is planned to be distributed under the same terms of use as its parent SweLL-gold corpus. In addition, a freely downloadable version, consisting of sentence pairs isolated from their context and stripped from their metadata, will be part of future UD releases.

33 Masciolini et al. (2024) already used a small SweLL-derived test set for evaluating a parsing model trained on synthetically generated word order errors. The data, manually validated after automatic pre-annotation, is available at https://github.com/spraakbanken/seapass/tree/main/data/swell. This is, however, mostly for the sake of reproducibility: both the choice of sentences and the extent of the manual checks, limited to word order errors themselves, make the treebank hardly reusable.

34 It must be kept in mind that, in the era of LLMs, "training scale" does not necessarily imply tens of thousands of sentences: while training a parser from scratch still requires large amounts of annotated text, a diverse few hundred sentences may be sufficient for fine-tuning a pretrained model.

5 *Concluding notes: SweLL impact on Swedish L2 research*

It is a fact that languages and research domains, that can boast rich data collections, have more empirical and data-intensive research done on them (Perc 2014, Søgaard 2022). This makes us believe that now, with the SweLL data available for research and development, the field of Swedish as a second language and related research fields will receive more attention. Since the release of SweLL-gold in 2021, we can see a steady increase in interest to:

1. data-driven linguistic studies on vocabulary and grammar in second language learning, grammatical patterns, etc. (e.g. Sundberg & Prentice 2023, Liljegren 2023);

2. novel approaches to feedback generation (e.g. Masciolini 2023, Masciolini et al. 2023);

3. methodological studies, such as, pseudonymization of research data, fairness and bias in language assessment (e.g., Szawerna, Dobnik, Tiedemann, et al. 2024, Szawerna, Dobnik, Sánchez, et al. 2024, Muñoz Sánchez et al. 2024, Rudebeck & Sundberg 2024, Volodina et al. 2020, 2022, Volodina, Dobnik, et al. 2023)

4. development of approaches to automatic error detection and correction, etc. (e.g. Nyberg 2022, Östling et al. 2024, Volodina, Bryant, et al. 2023, Ehnroth & Park 2023).

A number of derivative resources have been developed since 2021 based on the SweLL-gold corpus, such as DaLAJ (Volodina, Mohammed, et al. 2023) for studies on linguistic acceptability (Klezl et al. 2022); MuClaGED (Casademont Moner & Volodina 2022b) for error classification; synthetic datasets imitating real-life errors (Casademont Moner & Volodina 2022a) and many others. The Swedish MultiGED dataset[35] based on SweLL-gold has been used for the MultiGED shared task (Volodina, Bryant, et al. 2023) and MultiGEC dataset (Masciolini, Caines, De Clercq, Kruijsbergen, Kurfalı, Muñoz Sánchez, Volodina, Östling, et al. 2025) was used for shared task on grammatical error correction (Masciolini, Caines, De Clercq, Kruijsbergen, Kurfalı, Muñoz Sánchez, Volodina & Östling 2025), where the Swedish part was based on the SweLL-gold data.[36]

In the future, we are expecting both short-term and long-term impact from the SweLL-gold corpus on the fields of Swedish as a Second Language,

35 https://github.com/spraakbanken/multiged-2023
36 https://spraakbanken.gu.se/en/compsla/multigec-2025

Learner Corpus Research (nationally and internationally), and NLP- and AI-based approaches to L2 Swedish.

First of all, we intend to *promote* the use of the datasets among SLA researchers through detailed guidelines and assistance, and among NLP researchers through organization of multilingual shared tasks.[37] The current chapter is a way to promote the use of the data among the two user groups.

Second, we will work towards *extending authentic* learner datasets through use of automatic methods for data annotation, among other things, through setting on-the-fly pseudonymization algorithms for *continuous collection of essays* directly from schools and for automatic error correction and labeling.

Finally, we will also work on generation of *synthetic* datasets with a basis in the current SweLL data. For example, experimenting with LLMs to generate mock learner essays at different levels of proficiency, using real-life essays as samples; or generating error datasets using linguistic patterns observed in the SweLL-gold data. The additional synthetic data could help train better models for performing automatic annotation on authentic data.

Acknowledgments

Work on the article has been supported by *Nationella språkbanken* (years 2018-2028, contracts 2017-00626 and 2023-00161) and *Huminfra* (years 2022-2028, contracts 2021-00176 and 2023-00171), both funded by the Swedish Research Council and their participating partner institutions. Work on SweLL infrastructure has been supported by the infrastructure grant from the Swedish Riksbankens Jubileumsfond *SweLL – research infrastructure for Swedish as a second language*, during years 2017-2020, grant IN16-0464:1.

References

Adesam, Yvonne & Aleksandrs Berdicevskis. 2021. Part-of-speech tagging of Swedish texts in the neural era. In Simon Dobnik & Lilja Øvrelid (eds.), *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, 200–209. Reykjavik, Iceland (Online): Linköping University Electronic Press, Sweden. https://aclanthology.org/2021.nodalida-main.20/.

37 https://spraakbanken.gu.se/en/compsla

Arhar Holdt, Špela, Tomaž Erjavec, Iztok Kosem & Elena Volodina. 2024. Towards an ideal tool for learner error annotation. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti & Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 16392–16398. Torino, Italia: ELRA & ICCL. https://aclanthology.org/2024.lrec-main.1424/.

Arhar Holdt, Špela & Iztok Kosem. 2024. Šolar, the developmental corpus of Slovene. *Language Resources and Evaluation* 59. 1151–1177. DOI: 10.1007/s10579-024-09758-4.

Artstein, Ron & Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics* 34(4). 555–596. DOI: 10.1162/coli.07-034-R2.

Bel, Nuria, Marta Punsola & Valle Ruíz-Fernández. 2024. EsCoLA: Spanish corpus of linguistic acceptability. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti & Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 6268–6277. Torino, Italia: ELRA & ICCL. https://aclanthology.org/2024.lrec-main.554/.

Bel, Núria, Marta Punsola & Valle Ruiz-Fernández. 2024. CatCoLA, Catalan corpus of linguistic acceptability. *Procesamiento del Lenguaje Natural* 73. 177–190. http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6609.

Berdicevskis, Aleksandrs, Gerlof Bouma, Robin Kurtz, Felix Morger, Joey Öhman, Yvonne Adesam, Lars Borin, Dana Dannélls, Markus Forsberg, Tim Isbister, Anna Lindahl, Martin Malmsten, Faton Rekathati, Magnus Sahlgren, Elena Volodina, Love Börjeson, Simon Hengchen & Nina Tahmasebi. 2023. Superlim: A Swedish language understanding evaluation benchmark. In Houda Bouamor, Juan Pino & Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 8137–8153. Singapore: Association for Computational Linguistics. DOI: 10.18653/v1/2023.emnlp-main.506.

Berzak, Yevgeni, Jessica Kenney, Carolyn Spadine, Jing Xian Wang, Lucia Lam, Keiko Sophie Mori, Sebastian Garza & Boris Katz. 2016. Universal Dependencies for learner English. In Katrin Erk & Noah A. Smith (eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 737–746. Berlin, Germany: Association for Computational Linguistics. DOI: 10.18653/v1/P16-1070.

Borin, Lars, Markus Forsberg & Johan Roxendal. 2012. Korp — the corpus infrastructure of Språkbanken. In Nicoletta Calzolari, Khalid Choukri,

Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 474–478. Istanbul, Turkey: European Language Resources Association (ELRA). https://aclanthology.org/L12-1098/.

Boyd, Adriane, Jirka Hana, Lionel Nicolas, Detmar Meurers, Katrin Wisniewski, Andrea Abel, Karin Schöne, Barbora Štindlová & Chiara Vettori. 2014. The MERLIN corpus: Learner language and the CEFR. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 1281–1288. Reykjavik, Iceland: European Language Resources Association (ELRA). https://aclanthology.org/L14-1488/.

Callies, Marcus. 2015. Learner corpus methodology. In Sylviane Granger, Fanny Meunier & Gaëtanelle Gilquin (eds.), *Cambridge Handbook of Learner Corpus Research*, 35–55. Cambridge University Press. DOI: 10.1017/CBO9781139649414.

Casademont Moner, Judit & Elena Volodina. 2022a. Generation of synthetic error data of verb order errors for Swedish. In Ekaterina Kochmar, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Nitin Madnani, Anaïs Tack, Victoria Yaneva, Zheng Yuan & Torsten Zesch (eds.), *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, 33–38. Seattle, Washington: Association for Computational Linguistics. DOI: 10.18653/v1/2022.bea-1.6.

Casademont Moner, Judit & Elena Volodina. 2025. *MuClaGED*. DOI: 10.23695/q9v4-vt57.

Casademont Moner, Judith & Elena Volodina. 2022b. Swedish MuClaGED: A new dataset for grammatical error detection in Swedish. In David Alfter, Elena Volodina, Thomas François, Piet Desmet, Frederik Cornillie, Arne Jönsson & Evelina Rennes (eds.), *Proceedings of the 11th Workshop on NLP for Computer Assisted Language Learning*, 36–45. Louvain-la-Neuve, Belgium: LiU Electronic Press. https://aclanthology.org/2022.nlp4call-1.4.

Corder, Stephen Pit. 1971. Idiosyncratic dialects and error analysis. *International Review of Applied Linguistics in Language Teaching* 9(2). DOI: 10.1515/iral.1971.9.2.147.

Corder, Stephen Pit. 1973. *Introducing Applied Linguistics*. Penguin.

Council of Europe. 2001. *Common European framework of reference for languages: Learning, teaching, assessment*. https://rm.coe.int/1680459f97.

Dargis, Roberts, Ilze Auzina, Kristne Levane-Petrova & Inga Kaija. 2020. Detailed error annotation for morphologically rich languages: Latvian use case. *Frontiers in Artificial Intelligence and Applications* 328. 241–244. DOI: 10.3233/FAIA200629.

de Marneffe, Marie-Catherine, Christopher D. Manning, Joakim Nivre & Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics* 47(2). 255–308. DOI: 10.1162/coli_a_00402.

Di Nuovo, Elisa, Manuela Sanguinetti, Alessandro Mazzei, Elisa Corino & Cristina Bosco. 2022. VALICO-UD: Treebanking an Italian learner corpus in Universal Dependencies. *Italian Journal of Computational Linguistics* 8(1). DOI: 10.4000/ijcol.1007.

Ehnroth, Joel & Yoonjoo Park. 2023. *Correction of grammatical errors in Swedish.* Master's thesis U-CS-EX 2023-29, Lund University, Lund, Sweden. https://lup.lub.lu.se/student-papers/search/publication/9130663.

EU Commission. 2016. *General data protection regulation*. Accessed 2019-11-19. Official Journal of the European Union, 59, 1–88. https://gdpr-info.eu/.

Glisic, Isidora & Anton Karl Ingason. 2022. The nature of Icelandic as a second language: An insight from the learner error corpus for Icelandic. In *CLARIN Annual Conference*, 23–33. DOI: 10.3384/ecp1893.

Granger, Sylviane & Magali Paquot. 2017. *Core metadata* [*schema*] *for learner corpora draft 1.0*. http://hdl.handle.net/20.500.12124/61.

Hammarberg, Björn. 2010. *Introduktion till ASU-korpusen: En longitudinell muntlig och skriftlig textkorpus av vuxna inlärares svenska med en motsvarande del från infödda svenskar*. Tech. rep. Stockholm University: Department of Linguistics. https://www.diva-portal.org/smash/get/diva2:778194/FULLTEXT01.pdf.

Hammarstedt, Martin, Anne Schumacher, Lars Borin & Markus Forsberg. 2022. *Sparv 5 user manual*. Tech. rep. Göteborg. https://hdl.handle.net/2077/73604.

Hirschmann, Hagen, Seanna Doolittle & Anke Lüdeling. 2007. Syntactic annotation of non-canonical linguistic structures. In Mark Davies, Paul Rayson, Susan Hunston & Pernilla Danielsson (eds.), *Proceedings of the Corpus Linguistics Conference (CL 2007)*. https://www.birmingham.ac.uk/documents/college-artslaw/corpus/conference-archives/2007/128Paper.pdf.

Jentoft, Matias & David Samuel. 2023. NoCoLA: The Norwegian corpus of linguistic acceptability. In Tanel Alumäe & Mark Fishel (eds.), *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, 610–617. Tórshavn, Faroe Islands: University of Tartu Library. https://aclanthology.org/2023.nodalida-1.60.

Klezl, Julia, Yousuf Ali Mohammed & Elena Volodina. 2022. Exploring linguistic acceptability in Swedish learners' language. In *Proceedings of the 11th Workshop on NLP for Computer Assisted Language Learning*, 84–94. DOI: 10.3384/ecp190009.

König, Alexander, Jennifer-Carmen Frey & Egon W Stemle. 2021. Exploring reusability and reproducibility for a research infrastructure for L1 and L2 learner corpora. *Information* 12(5). DOI: 10.3390/info12050199.

König, Alexander, Jennifer-Carmen Frey, Egon W Stemle, Aivars Glaznieks & Magali Paquot. 2022. Towards standardizing LCR metadata. In *6th International Conference for Learner Corpus Research* (*LCR 2022*). Padova, Italy. https://hdl.handle.net/10863/42994.

Kotsinas, Ulla-Britt. 1982. *Svenska svårt: Några invandrares svenska talspråk:*[*ordförrådet*]. Institutionen för nordiska språk, Stockholms universitet. (Doctoral dissertation).

Krippendorff, Klaus. 2019. *Content analysis: An introduction to its methodology*. Sage publications. DOI: 10.4135/9781071878781.

Kyle, Kristopher, Masaki Eguchi, Aaron Miller & Theodore Sither. 2022. A Dependency treebank of spoken second language English. In Ekaterina Kochmar, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Nitin Madnani, Anaïs Tack, Victoria Yaneva, Zheng Yuan & Torsten Zesch (eds.), *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications* (*BEA 2022*), 39–45. Seattle, Washington: Association for Computational Linguistics. DOI: 10.18653/v1/2022.bea-1.7.

Lee, John, Herman Leung & Keying Li. 2017. Towards Universal Dependencies for learner Chinese. In Marie-Catherine de Marneffe, Joakim Nivre & Sebastian Schuster (eds.), *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies* (*UDW 2017*), 67–71. Gothenburg, Sweden: Association for Computational Linguistics. https://aclanthology.org/W17-0408.

Lee, John, Keying Li & Herman Leung. 2017. L1-L2 parallel dependency treebank as learner corpus. In *Proceedings of the 15th International Conference on Parsing Technologies*, 44–49. Pisa, Italy: Association for Computational Linguistics. https://aclanthology.org/W17-6306.

Liljegren, Johan. 2023. *Komplexitetsdrag i andraspråkstexter. En korpusbaserad undersökning av syntaktisk komplexitet i andraspråksinlärares skriftliga svenska.* Master's thesis, diva2:1820372, Stockholms universitet, Lund, Sweden. https://su.diva-portal.org/smash/record.jsf?pid=diva2%3A1820372.

Lindberg, Janne & Gunnar Eriksson. 2005. CrossCheck-korpusen — en elektronisk L2-korpus för skriven svenska. In Boel De Geer & Anna Malmbjer (eds.), *Språk på tvärs. Rapport från ASLA:s höstsymposium, Södertörn*

11–12 november 2004, 89–98. Uppsala: Svenska föreningen för tillämpad språkvetenskap.

Lüdeling, Anke & Hagen Hirschmann. 2015. Error annotation systems. In *Cambridge Handbook of Learner Corpus Research*. Sylviane Granger, Fanny Meunier & Gaëtanelle Gilquin (eds.). Cambridge University Press. 135–157. DOI: 10.1017/CBO9781139649414.

Lüdeling, Anke, Maik Walter, Emil Kroymann & Peter Adolphs. 2005. Multilevel error annotation in learner corpora. In *Proceedings of Corpus Linguistics*, vol. 1, 14–17.

Masciolini, Arianna. 2023. A query engine for L1-L2 parallel dependency treebanks. In *Proceedings of the 24th Nordic Conference on Computational Linguistics* (*NoDaLiDa*), 574–587. Tórshavn, Faroe Islands: University of Tartu Library. https://aclanthology.org/2023.nodalida-1.57.

Masciolini, Arianna, Andrew Caines, Orphée De Clercq, Joni Kruijsbergen, Murathan Kurfali, Ricardo Muñoz Sánchez, Elena Volodina, Robert Östling, Kais Allkivi-Metsoja, Špela Arhar Holdt, Ilze Auzina, Roberts Darģis, Elena Drakonaki, Jennifer-Carmen Frey, Isidora Glišić, Pinelopi Kikilintza, Lionel Nicolas, Mariana Romanyshyn, Alexandr Rosen, Alla Rozovskaya, Kristjan Suluste, Oleksiy Syvokon, Alexandros Tantos, Despoina-Ourania Touriki, Konstantinos Tsiotskas, Eleni Tsourilla, Vassilis Varsamopoulos, Katrin Wisniewski, Aleš Žagar & Torsten Zesch. 2025. *MultiGEC*. DOI: 10.23695/h9f5-8143.

Masciolini, Arianna, Andrew Caines, Orphée De Clercq, Joni Kruijsbergen, Murathan Kurfalı, Ricardo Muñoz Sánchez, Elena Volodina & Robert Östling. 2025. The MultiGEC-2025 shared Ttask on multilingual grammatical Eerror correction at NLP4CALL. *Proceedings of the 14th Workshop on Natural Language Processing for Computer Assisted Language Learning* (*NLP4CALL 2025*). https://hdl.handle.net/10062/107166.

Masciolini, Arianna, Andrew Caines, Orphée De Clercq, Joni Kruijsbergen, Murathan Kurfalı, Ricardo Muñoz Sánchez, Elena Volodina, Robert Östling, Kais Allkivi, Špela Arhar Holdt, Ilze Auzina, Roberts Darģis, Elena Drakonaki, Jennifer-Carmen Frey, Isidora Glišić, Pinelopi Kikilintza, Lionel Nicolas, Mariana Romanyshyn, Alexandr Rosen, Alla Rozovskaya, Kristjan Suluste, Oleksiy Syvokon, Alexandros Tantos, Despoina-Ourania Touriki, Konstantinos Tsiotskas, Eleni Tsourilla, Vassilis Varsamopoulos, Katrin Wisniewski, Aleš Žagar & Torsten Zesch. 2025. Towards better language representation in Natural Language Processing – a multilingual dataset for text-level Grammatical Error Correction. *International Journal of Learner Corpus Research* vol. 11, issue 2. 309–335. DOI: 10.1075/ijlcr.24033.mas.

Masciolini, Arianna, Emilie Francis & Maria Irena Szawerna. 2024. Synthetic-error augmented parsing of Swedish as a second language: Experiments with word order. In Archna Bhatia, Gosse Bouma, A. Seza Dogruoz, Kilian Evang, Marcos Garcia, Voula Giouli, Lifeng Han, Joakim Nivre & Alexandre Rademaker (eds.), *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, 43–49. Torino, Italia: ELRA & ICCL. https://aclanthology.org/2024.mwe-1.7.

Masciolini, Arianna, Herbert Lange & Márton András Tóth. 2025. Exploring parallel corpora with STUnD: A search tool for Universal Dependencies. In Elena Volodina, Gerlof Bouma, Dana Dannélls & Dimitrios Kokkinakis (eds.), *Huminfra handbook: Empowering digital and experimental humanities* (NEALT Proceedings Series 59), 455–503. University of Tartu Library. DOI: 10.58009/aere-perennius0183.

Masciolini, Arianna, Elena Volodina & Dana Dannélls. 2023. Towards automatically extracting morphosyntactical error patterns from L1-L2 parallel dependency treebanks. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, 585–597. https://aclanthology.org/2023.bea-1.50/.

Megyesi, Beáta, Lena Granstedt, Sofia Johansson, Julia Prentice, Dan Rosén, Carl-Johan Schenström, Gunlög Sundberg, Mats Wirén & Elena Volodina. 2018. Learner corpus anonymization in the age of GDPR: Insights from the creation of a learner corpus of Swedish. In *Proceedings of the 7th workshop on NLP for Computer Assisted Language Learning*, 47–56. Stockholm, Sweden: LiU Electronic Press. https://aclanthology.org/W18-7106.

Megyesi, Beáta, Jesper Näsman & Anne Palmér. 2016. The Uppsala corpus of student writings: Corpus creation, annotation, and analysis. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 3192–3199.

Megyesi, Beáta, Lisa Rudebeck & Elena Volodina. 2021. *SweLL pseudonymization guidelines*. http://hdl.handle.net/2077/69431.

Mendes, Amália, Sandra Antunes, Maarten Janssen & Anabela Gonçalves. 2016. The COPLE2 corpus: A learner corpus for Portuguese. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the tenth international conference on language resources and evaluation (LREC'16)*, 3207–3214. Portorož, Slovenia: European Language Resources Association (ELRA). https://aclanthology.org/L16-1511/.

Mikhailov, Vladislav, Tatiana Shamardina, Max Ryabinin, Alena Pestova, Ivan Smurov & Ekaterina Artemova. 2022. RuCoLA: Russian corpus of lin-

guistic acceptability. In Yoav Goldberg, Zornitsa Kozareva & Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 5207–5227. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics. DOI: 10.18653/v1/2022.emnlp-main.348.

Mohammed, Yousuf Ali, Arild Matsson & Elena Volodina. 2022. Annotation management tool: A requirement for corpus construction. In *CLARIN Annual Conference*, 101–108. DOI: 10.3384/ecp18910.

Muñoz Sánchez, Ricardo, Simon Dobnik, Maria Irena Szawerna, Therese Lindström Tiedemann & Elena Volodina. 2024. Did the names I used within my essay affect my score? Diagnosing name biases in automated essay scoring. In *Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization (CALD-pseudo 2024)*, 81–91. https://aclanthology.org/2024.caldpseudo-1.10/.

Nyberg, Martina. 2022. *Grammatical error correction for learners of Swedish as a second language*. Master's Thesis, Uppsala university. http://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1666233&dswid=8031.

Östling, Robert, Katarina Gillholm, Murathan Kurfalı, Marie Mattson & Mats Wirén. 2024. Evaluation of really good grammatical error correction. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti & Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 6582–6593. Torino, Italia: ELRA & ICCL. https://aclanthology.org/2024.lrec-main.584/.

Paquot, Magali. 2022. Corpora and Second Language Acquisition. In *The Routledge Handbook of Corpora and English Language Teaching and Learning*, 26–40. Routledge.

Paquot, Magali, Alexander König, Egon W Stemle & Jennifer-Carmen Frey. 2023. A core metadata schema for L2 data. In *Book of Abstracts from the EuroSLA Conference 2023*. https://www.birmingham.ac.uk/research/lacab/events/eurosla-32.

Paquot, Magali, Alexander König, Egon W Stemle & Jennifer-Carmen Frey. 2024. The core metadata schema for learner corpora (LC-meta). Collaborative efforts to advance data discoverability, metadata quality and study comparability in L2 research. *International Journal of Learner Corpus Research* 10(2). 280–300. DOI: 10.1075/ijlcr.24010.paq.

Perc, Matjaž. 2014. The Matthew effect in empirical data. *Journal of The Royal Society Interface* 11(98). DOI: 10.1098/rsif.2014.0378.

Reznicek, Marc, Anke Lüdeling, Cedric Krummes, Franziska Schwantuschke, Maik Walter, Karin Schmidt, Hagen Hirschmann & Torsten Andreas. 2012.

Das Falko-Handbuch. Korpusaufbau und Annotationen. Version 2.01. Berlin, Germany: Humboldt-Universität zu Berlin.

Römer, Ute. 2019. A corpus perspective on the development of verb constructions in second language learners. *International Journal of Corpus Linguistics* 24(3). 268–290. DOI: 10.1075/ijcl.00013.roe.

Rosen, Alexandr, Jiří Hana, Barbora Vidová Hladká, Tomáš Jelínek, Svatava Škodová & Barbora Štindlová. 2020. *Compiling and annotating a learner corpus for a morphologically rich language: CzeSL, a corpus of non-native Czech*. Nakladatelství Karolinum.

Rozovskaya, Alla. 2024. Universal Dependencies for learner Russian. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti & Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 17112–17119. Torino, Italia: ELRA & ICCL. https://aclanthology.org/2024.lrec-main.1486.

Rudebeck, Lisa & Gunlög Sundberg. 2021. SweLL correction annotation guidelines. https://gupea.ub.gu.se/handle/2077/69434.

Rudebeck, Lisa & Gunlög Sundberg. 2024. On the other side of the error tag: Predifined normalized texts as a basis for correction annotation. In Katherine Ackerley & Erik Castello (eds.), *Continuing Learner Corpus Research: Challenges and opportunities* (Corpora and Language in Use Proceedings 7), 123–152. Presses universitaires de Louvain.

Rudebeck, Lisa, Gunlög Sundberg & Mats Wirén. 2021. SweLL normalization guidelines. https://gupea.ub.gu.se/handle/2077/69432.

Søgaard, Anders. 2022. Should we ban English NLP for a year? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 5254–5260. DOI: 10.18653/v1/2022.emnlp-main.351.

Someya, Taiga, Yushi Sugimoto & Yohei Oseki. 2024. JCoLA: Japanese corpus of linguistic acceptability. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti & Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 9477–9488. Torino, Italia: ELRA & ICCL. https://aclanthology.org/2024.lrec-main.828/.

Stemle, Egon W, Adriane Boyd, Maarten Jansen, Therese Lindström Tiedemann, Nives Mikelić Preradović, Alexandr Rosen, Dan Rosén & Elena Volodina. 2019. Working together towards an ideal infrastructure for language learner corpora. In Andrea Abel, Aivars Glaznieks, Verena Lyding & Lionel Nicolas (eds.), *Widening the Scope of Learner Corpus Research. Selected papers from the fourth Learner Corpus Research Conference. Corpora*

and Language in Use – Proceedings 5, 427–468. Louvain-la-Neuve: Presses universitaires de Louvain.

Sundberg, Gunlög & Julia Prentice. 2023. SweLL: En svensk inlärarkorpus. *ASLA:s skriftserie/ASLA Studies in Applied Linguistics* 30. 428–453.

Sung, Hakyung & Gyu-Ho Shin. 2024. Constructing a dependency treebank for second language learners of Korean. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti & Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 3747–3758. Torino, Italia: ELRA & ICCL. https://aclanthology.org/2024.lrec-main.332.

Szawerna, Maria Irena, Simon Dobnik, Ricardo Muñoz Sánchez, Therese Lindström Tiedemann & Elena Volodina. 2024. Detecting personal identifiable information in Swedish learner essays. In *Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization (CALD-pseudo 2024)*, 54–63. https://aclanthology.org/2024.caldpseudo-1.7/.

Szawerna, Maria Irena, Simon Dobnik, Therese Lindström Tiedemann, Ricardo Muñoz Sánchez, Xuan-Son Vu & Elena Volodina. 2024. Pseudonymization categories across domain boundaries. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 13303–13314. https://aclanthology.org/2024.lrec-main.1164/.

Tenfjord, Kari, Jon Erik Hagen & Hilde Johansen. 2009. Norsk andrespråkskorpus (ASK) — design og metodiske forutsetninger. *Norsk som Andrespråk 25(1)*. 52–81.

Tenfjord, Kari, Paul Meurer & Knut Hofland. 2006. The ASK corpus — a language learner corpus of Norwegian as a second language. In Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odijk & Daniel Tapias (eds.), *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. Genoa, Italy: European Language Resources Association (ELRA). https://aclanthology.org/L06-1345/.

Thewissen, Jennifer. 2013. Capturing L2 accuracy developmental patterns: Insights from an error-tagged EFL learner corpus. *The Modern Language Journal* 97(S1). 77–101.

Trotta, Daniela, Raffaele Guarasci, Elisa Leonardelli & Sara Tonelli. 2021. Monolingual and cross-Lingual acceptability judgments with the Italian CoLA corpus. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia & Scott Wen-tau Yih (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2929–2940. Punta Cana, Dominican Republic:

Association for Computational Linguistics. DOI: 10.18653/v1/2021.findings-emnlp.250.

VanPatten, Bill, Gregory D. Keating & Stefanie Wulff (eds.). 2025. *Theories in second language acquisition. An introduction*. 4th edition. New York: Routledge. DOI: 10.4324/9781003491118.

van Rooy, Bertus. 2015. Annotating learner corpora. In Sylviane Granger, Fanny Meunier & Gaëtanelle Gilquin (eds.), *Cambridge Handbook of Learner Corpus Research*, 79–106. Cambridge University Press. DOI: 10.1017/CBO9781139649414.

Viberg, Åke. 1992. Universellt och språkspecifikt i det svenska ordförrådets organisation. *Tijdschrift voor Skandinavistiek* 13(2).

Viberg, Åke. 2004. Lexikal utveckling i ett andraspråk. In *Svenska som andraspråk: i forskning, undervisning och samhälle*. Kenneth Hyltenstam & Inger Lindberg (eds.). Studentlitteratur, Lund. 197–200.

Vinogradova, Olga & Olga Lyashevskaya. 2022. Review of practices of collecting and annotating texts in the learner corpus REALEC. In *International Conference on Text, Speech, and Dialogue*, 77–88.

Volodina, Elena. 2024. On two SweLL learner corpora — SweLL-pilot and SweLL-gold. In *Proceedings of the Huminfra Conference* (*HiC 2024*), 83–94. Linköping Electronic Conference Proceedings 205. DOI: 10.3384/ecp205012.

Volodina, Elena, David Alfter, Therese Lindström Tiedemann, Maisa Susanna Lauriala & Daniela Helena Piipponen. 2022. Reliability of automatic linguistic annotation: Native vs non-native texts. In *Selected papers from the CLARIN Annual Conference 2021*. DOI: 10.3384/ecp18914.

Volodina, Elena, Chris Bryant, Andrew Caines, Orphée De Clercq, Jennifer-Carmen Frey, Elizaveta Ershova, Alexandr Rosen & Olga Vinogradova. 2025. *MultiGED*. DOI: 10.23695/xe7r-k506.

Volodina, Elena, Christopher Bryant, Andrew Caines, Orphée De Clercq, Jennifer-Carmen Frey, Elizaveta Ershova, Alexandr Rosen & Olga Vinogradova. 2023. MultiGED-2023 shared task at NLP4CALL: Multilingual Grammatical Error Detection. In *Proceedings of the 12th Workshop on NLP for Computer Assisted Language Learning*, 1–16. Linköping Electronic Conference Proceedings 197. DOI: 10.3384/ecp197001.

Volodina, Elena, Simon Dobnik, Therese Lindström Tiedemann & Xuan-Son Vu. 2023. Grandma Karl is 27 years old–research agenda for pseudonymization of research data. In *2023 IEEE Ninth International Conference on Big Data Computing Service and Applications* (*BigDataService*), 229–233.

Volodina, Elena, Lena Granstedt, Arild Matsson, Beáta Megyesi, Ildikó Pilán, Julia Prentice, Dan Rosén, Lisa Rudebeck, Carl-Johan Schenström, Gunlög

Sundberg, et al. 2019. The SweLL language learner corpus: From design to annotation. *Northern European Journal of Language Technology* (*NEJLT*) 6. 67–104. DOI: 10.3384/nejlt.2000-1533.19667.

Volodina, Elena, Lena Granstedt, Beáta Megyesi, Julia Prentice, Lisa Rudebeck, Gunlög Sundberg & Mats Wirén. 2025. *SweLL-gold*. DOI: 10.23695/2k47-y432.

Volodina, Elena, Maarten Janssen, Therese Lindström Tiedemann, Nives Mikelic Preradovic, Silje Karin Ragnhildstveit, Kari Tenfjord & Koenraad de Smedt. 2018. Interoperability of second language resources and tools. In *Proceedings of the CLARIN Annual Conference*, 90–94. https://ep.liu.se/en/conference-issue.aspx?series=ecp&issue=159.

Volodina, Elena & Beáta Megyesi. 2021. SweLL transcription guidelines, L2 essays. https://gupea.ub.gu.se/handle/2077/69429.

Volodina, Elena, Beáta Megyesi, Mats Wirén, Lena Granstedt, Julia Prentice, Monica Reichenberg & Gunlög Sundberg. 2016. A friend in need? Research agenda for electronic second language infrastructure. In *Swedish Language Technology Conference* (*SLTC*) *2016*. https://spraakbanken.gu.se/sites/default/files/d7/2016_SLTC_L2infra_v2.pdf.

Volodina, Elena & Yousuf Ali Mohammed. 2024. *DaLAJ-GED-Superlim 2.0*. DOI: 10.23695/KXVZ-TX42.

Volodina, Elena, Yousuf Ali Mohammed, Aleksandrs Berdicevskis, Gerlof Bouma & Joey Öhman. 2023. DaLAJ-GED — a dataset for grammatical error detection tasks on Swedish. In David Alfter, Elena Volodina, Thomas François, Arne Jönsson & Evelina Rennes (eds.), *Proceedings of the 12th Workshop on NLP for Computer Assisted Language Learning*, 94–101. Tórshavn, Faroe Islands: LiU Electronic Press. DOI: 10.3384/ecp197011.

Volodina, Elena, Yousuf Ali Mohammed, Sandra Derbring, Arild Matsson & Beata Megyesi. 2020. Towards privacy by design in Learner Corpora Research: A case of on-the-fly pseudonymization of Swedish learner essays. In *Proceedings of the 28th International Conference on Computational Linguistics*, 357–369. Barcelona, Spain (Online): International Committee on Computational Linguistics. https://aclanthology.org/2020.coling-main.32.

Warstadt, Alex, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang & Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics* 8. 377–392. DOI: 10.1162/tacl_a_00321.

Warstadt, Alex, Amanpreet Singh & Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics* 7. 625–641. DOI: 10.1162/tacl_a_00290.

Wirén, Mats, Arild Matsson, Dan Rosén & Elena Volodina. 2019. SVALA: annotation of second-language learner text based on mostly automatic alignment of parallel corpora. In *CLARIN Annual Conference, Pisa, Italy, 8–10 October 2018*, 222–234. https://ep.liu.se/ecp/159/023/ecp18159023.pdf.

Yannakoudakis, Helen, Ted Briscoe & Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In Dekang Lin, Yuji Matsumoto & Rada Mihalcea (eds.), *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, 180–189. Portland, Oregon, USA: Association for Computational Linguistics. https://aclanthology.org/P11-1019/.

List of abbreviations

API	Application Programming Interface
CEFR	Common European Fraework of Reference
DaLAJ	Dataset for Linguistic Acceptability Judgments
GEC	Grammatical Error Correction
GED	Grammatical Error Detection
GDPR	General Data Protection Regulation
JSON	JavaScript Object Notation
L2	Second Language
LCR	Learner Corpus Research
LLM	Large Language Model
ML	Machine Learning
MuClaGED	Multi-Class Grammatical Error Detection
MultiGEC	Multilingual Grammatical Error Correction
MultiGED	Multilingual Grammatical Error Detection
NLP	Natural Language Processing
SFI	Swedish For Immigrants
SLA	Second Language Acquisition
SVA	Swedish as a Second Language (abbreviation from the Swedish *SVenska som Andraspråk*)
SweLL	Swedish Learner Language (name of a corpus collection and of an infrastructure)
TEI	Text Encoding Initiative
TISUS	Test In Swedish for University Studies
TSV	Tab Separated Values

UD	Universal Dependencies
VAC	Verb-Argument Construction
XML	eXtensible Markup Language

Appendix 1 *Essay topics*

Topic	Nr. essays
Om din bostad och om att bo	16
Berätta hur du bor!	8
Utredande text (pm), övning inför NP	9
Referat av texten "Världens språk tynar bort"	4
Argumenterande text om språk	14
Enkel utredande text om litterära teman	17
Referat av texten "En modern folktro"	2
Referat av texten "Giftermål ett större steg än barn"	2
Skriv ett brev	16
Skriv ett mail	18
Argumenterande text/brev	13
Skriv en insändare	11
Ge tips och råd - en anställningsintervju	10
En kulturupplevelse	6
Ge tips och råd	15
Min första kärlek	27
Mina första intryck	13
Insändare	11
Beskriv En god relation	8
Behöver man en egen bil?	12
Stänga utomhusbadet	9
Objektivt utredande uppgift	6
Familjen	13
Kommunikation och sociala medier	14
Två sätt att uppfostra	5
Brott orsaker och konsekvenser	8
Argumenterande text om arbetsmoral	6
Världens lyckligaste länder	14
En viktig plats	29
Mejl till en vän	30
En plats du tycker om	42
Demokratiska val - hur gammal ska man behöva vara för att rösta och varför?	15

(*Continues on next page*)

(*Continued from previous page*)

Topic	Nr. essays
Diskuterande text om pengars betydelse	4
Om utsatta grupper	9
Den skrämmande resenären/Fängelset/Hundvalparna	10
Romanexamination	12
Argumenterande text om slöjor eller krav för medborgarskap	5
Hur viktiga är kläderna?	39
Total	502

Corresponding author

Elena Volodina
Språkbanken Text
Department of Swedish,
Multilingualism, Language
Technology
University of Gothenburg
elena.volodina@svenska.gu.se

10 From text to insight: Uncovering linguistic patterns with SWEGRAM

Beáta Megyesi
Stockholm University

Rex Ruan
Stockholm University

Empirical linguistic analysis provides valuable insights into textual data for researchers in the humanities and social sciences, enabling them to identify patterns and trends within large datasets. SWEGRAM is a freely available tool designed to annotate and analyze Swedish and English texts without requiring programming skills or a user account. Users can upload one or more texts for linguistic analysis, extracting morphological and syntactic features. The linguistically annotated texts can then be used for quantitative linguistic analysis, allowing researchers to systematically explore textual characteristics. Additionally, the tool visualizes syntactic relations between words in sentences and provides detailed insights into the distribution of syntactic functions and relations within the text. Users can also create their own linguistically annotated text collections and generate statistical summaries of the linguistic properties of their texts. The tool is available as both a web-based service, which requires no user login or account, and a downloadable version for local use when data privacy and security are a priority. This dual availability ensures accessibility and flexibility for diverse research needs.

1 *Introduction*

Linguistic annotation and text analysis play a crucial role in the humanities and social sciences, offering a systematic approach to capturing and examining language use across different contexts. By assigning linguistic features—such as morphological, syntactic, semantic, and discourse-related properties—to words, phrases, sentences, paragraphs, and texts, researchers can identify patterns and trends that deepen our understanding of human communication, cultural practices, and social interaction. This approach

HUM INFRA Beáta Megyesi & Rex Ruan. 2025. From text to insight: Uncovering linguistic patterns with SWEGRAM. In Gerlof Bouma, Dana Dannélls, Dimitrios Kokkinakis & Elena Volodina (eds.), *Huminfra handbook: Empowering digital and experimental humanities* (NEALT Proceedings Series 59), 307–353. University of Tartu Library. DOI: 10.58009/aere-perennius0179 © The authors, CC BY 4.0

allows for the creation of large, annotated corpora that can be analyzed quantitatively and qualitatively, offering insights into language use, social dynamics, and culture. The empirical foundation of these linguistic annotations ensures that analyses are grounded in actual language use, which is essential for developing robust theories and models in linguistics, history, history of ideas, sociology, psychology, anthropology, and related fields.

Moreover, empirical linguistic analysis facilitates interdisciplinary research, bridging gaps between traditional humanities scholarship and the social sciences. For instance, sociolinguistic studies that combine annotated linguistic data with demographic information can reveal how language reflects and reinforces social stratification or how linguistic practices vary across different communities. Similarly, in digital humanities, the analysis of annotated texts enables the exploration of literary styles, authorship attribution, and the evolution of genres over time. This empirical approach not only enriches our understanding of language as a social and cultural phenomenon but also enhances the methodological rigor of research in the humanities and social sciences, making it possible to address complex questions with greater precision and depth.

Several powerful tools are available for the automatic annotation and analysis of texts (e.g., Bird et al. 2009, Kilgarriff et al. 2014, Manning et al. 2014, Honnibal & Montani 2017), each offering unique features tailored to different research needs. In this chapter, we describe SWEGRAM,[1] a web-based tool designed for the automatic annotation and linguistic analysis of texts in both English and Swedish. It not only facilitates detailed linguistic analysis but also allows users to create and manage their own text corpora, thus serving as a valuable tool for researchers working with language data.

SWEGRAM was originally developed for Swedish (Näsman et al. 2017, Megyesi et al. 2019) and through a collaboration between the Department of Linguistics and Philology and the Department of Scandinavian Languages at Uppsala University, Sweden. The tool was adapted later to English. Since 2024, SWEGRAM has been hosted by the Department of Linguistics at Stockholm University, Sweden. SWEGRAM has been an integral part of the Swe-CLARIN[2] project, which serves as the Swedish node of the European CLARIN infrastructure, with financial support from the Swedish Research Council. The primary goal of SWEGRAM is to make language-based text material accessible as primary research data for the humanities and social sciences, utilizing advanced processing tools for both written and transcribed spoken language.

1 http://swegram.ling.su.se/
2 https://sweclarin.se

In the following sections, we will introduce the components of the tool and provide guidance on how it can be utilized for empirical linguistic analysis. We begin by outlining some related work to set the context.

2 *Related work*

Numerous robust tools have been developed to support the automatic annotation and analysis of texts, each equipped with unique features that cater to diverse research needs in the humanities and social sciences. Tools like Stanford CoreNLP (Manning et al. 2014) and spaCy (Honnibal & Montani 2017) offer comprehensive natural language processing (NLP) pipelines capable of performing a wide range of tasks, including tokenization, part-of-speech tagging, named entity recognition, and syntactic parsing. These tools are highly scalable and integrate seamlessly into larger data analysis workflows, making them indispensable for complex linguistic studies.

The Natural Language Toolkit (NLTK) (Bird et al. 2009) is another widely used resource, particularly valued in educational and research settings for its extensive documentation and user-friendly design. NLTK supports a broad spectrum of text processing tasks, while specialized tools like VADER (Hutto & Gilbert 2014) or MALLET (McCallum 2002) excel in areas such as sentiment analysis and topic modeling. These tools are especially well-suited for analyzing social media data and conducting large-scale text mining projects. In the realm of spoken language annotation, ELAN (Wittenburg et al. 2006) and EXMARaLDA (Schmidt & Wörner 2014) are popular for their ability to synchronize audio or video recordings with transcriptions, annotations, and metadata. Collectively, these tools significantly enhance the efficiency and precision of linguistic annotation and analysis, empowering researchers to handle large volumes of text data and uncover meaningful patterns.

Beyond these specific tools, a variety of others are available for different types of text analysis. Xaira (Burnard 2006) is an open-source software package that supports the indexing and analysis of corpora using Extensible Markup Language (XML); originally developed for the British National Corpus (BNC) (BNC Consortium 2007), it has become a standard tool for corpus-based research. Similarly, BNCWeb (Hoffmann et al. 2008) offers a web-based interface for the British National Corpus, enabling users to conduct in-depth searches and analyses.

Concordance programs such as AntConc (Laurence 2011), ProtAnt (Anthony & Baker 2015), The Sketch Engine and WebCorp (Kilgarriff et al. 2014) are also popular in the field of text analysis, valued for their ability to dis-

play various text-related features like word frequencies, collocations, and keywords. WordSmith Tools (Scott 2016) is another widely used suite of text analysis software, offering functionalities that include the creation of word lists, concordance lists, clusters, collocations, and keyword analysis. These tools collectively contribute to a comprehensive ecosystem for text analysis, supporting both qualitative and quantitative research across different languages and text types.

In the context of Swedish linguistic research, several specialized tools have been developed to facilitate the automatic annotation and analysis of Swedish texts. SPARV, developed by Språkbanken at the University of Gothenburg (Borin et al. 2016, Hammarstedt et al. 2022), is a powerful tool offering a comprehensive suite of annotation services tailored for Swedish. It provides functionalities such as tokenization, sentence segmentation, morphological analysis, named entity recognition, and dependency parsing. SPARV's high degree of customizability allows users to fine-tune the annotation pipeline according to specific research requirements.

Another notable tool from Språkbanken is Korp (Borin et al. 2012), a corpus search engine that grants access to extensive annotated corpora of Swedish texts. Korp enables users to conduct complex searches across various linguistic features, including word forms, lemmas, and syntactic structures, facilitating in-depth analysis and exploration of language usage in corpora that are available through Språkbanken's web portal.

In parallel with the development of Korp which later also featured an annotation pipeline called Annotation Lab (Borin et al. 2012), the first version of SWEGRAM (Näsman et al. 2017, Megyesi et al. 2019) was released in 2015. At that time, there was no tool available for the automatic annotation of Swedish that was accessible to researchers without programming skills. SWEGRAM aimed to enrich the landscape of Swedish linguistic tools by offering comprehensive analysis capabilities for Swedish and later for English texts, providing services such as tokenization, part-of-speech tagging, lemmatization, and syntactic parsing. Another equally important feature of SWEGRAM is its ability to conduct linguistic analysis based on annotations within the same tool, streamlining the research process. Additionally, SWEGRAM allows users to create and manage their own corpora, making it a versatile resource for both corpus creation and analysis. Over the past few years, SWEGRAM has been updated and recently released as SWEGRAM 2.0 with new linguistic features, enhanced visualization options, and a redesigned user interface, further solidifying its utility in linguistic research, which we will describe in the subsequent sections.

In summary, linguistic annotation tools are crucial for advancing research, offering robust resources for both small-scale and large-scale studies. These

tools streamline text processing and enable detailed, scalable linguistic investigations, benefiting the broader humanities and social sciences. The array of tools for automatic annotation and analysis enhances researchers' ability to conduct large-scale empirical studies.

3 *SWEGRAM*

SWEGRAM (Näsman et al. 2017, Megyesi et al. 2019) was developed to provide automatic annotation and analysis of texts, enabling empirical linguistic analysis of large text corpora within a single tool. The primary goal was to create a web-based platform that allows users to perform annotation and statistical linguistic analysis without requiring any programming skills. Emphasis was placed on making the tool user-friendly, offering a web service that allows users to upload texts without needing to log in or provide personal information. The tool is designed to be easily accessible for researchers and others interested in the humanities and social sciences and requires no knowledge of automated text processing, nor any programming skills.

In SWEGRAM, the user can upload one or more texts and have them linguistically analyzed for morphological and syntactic features. These linguistically annotated texts can then be used to perform quantitative linguistic text analysis or distant reading with the tool providing statistics on sentence length, total word count, readability metrics, part-of-speech (PoS) distribution, and the frequency of lemmas, PoS categories, or misspelled words. Additionally, SWEGRAM supports qualitative text analysis in a close reading scenario, allowing users to closely examine linguistic patterns and structures within individual texts or across a corpus. The tool visualizes the syntactic relationships between words in sentences and provides detailed information about the distribution of various syntactic functions and relations in the text. Moreover, users can easily compile their own linguistically annotated text collection (corpus) and generate comprehensive statistics on its linguistic characteristics.

SWEGRAM 1.0, initially introduced in 2015, has since been employed in numerous research projects (Megyesi et al. 2016, 2019). Recently, SWEGRAM has been updated to support the processing of English texts. In addition, we have expanded the range of linguistic features for both English and Swedish, incorporating readability metrics for English texts and second language acquisition analysis, inspired by Ildikó Pilán's PhD thesis (Pilán 2018). The visualization of morphosyntactic analysis has also been enhanced, now offering graphical representations of syntax trees for individual sentences. SWEGRAM now features an upgraded back-end, enabling faster and more

efficient text processing. In the latest version, even the analysis of relatively large texts—though still requiring considerable processing time—can be displayed incrementally. In the online version, this is facilitated by temporarily storing annotated texts in a database, reducing the burden of data loading during processing. These temporary records are automatically removed after a few days. Users are fully responsible for ensuring compliance with copyright and any other applicable restrictions when uploading data to the system. For non-public or sensitive data, SWEGRAM can be run locally using the downloadable version (see Section 6).

Next, we give an overview of the various functionalities of SWEGRAM 2.0, from text upload and linguistic annotation to text analysis, visualization of features, and exporting annotated text and statistics for further processing.

4 *Using SWEGRAM*

SWEGRAM comprises two major components, the annotation taking care of the labeling on various linguistic levels and the analysis part, deriving statistics of a wide range of linguistic features. Figure 1 illustrates the two major components of the tool. SWEGRAM offers functionalities for language selection, text upload, linguistic annotation pipeline, linguistic analysis, and the export of text and data files. We provide an overview of each. A more detailed description of these functionalities can be found in the SWEGRAM 2.0 Guidelines (Megyesi & Ruan 2024).

4.1 *Preparing for processing*

Before processing begins, the user must select the language to be processed: English or Swedish. Depending on the chosen language, the system applies the appropriate language-specific pipeline. Then, one or more files can be uploaded for annotation (up to 10 MB raw text), preferably in the Universal Character Encoding Standard Unicode UTF-8 format. Alternatively, the user can choose to paste text directly into the designated field, as illustrated in Figure 2. The uploaded files are displayed in the text selection area under the *Statistics* or *Visualize* modules.

To enable text comparison, the uploaded texts can also contain optional metadata of any type. Each line of metadata must follow a certain format to be processed correctly by the tool. The format is illustrated in Figure 3a. Each line of the metadata must start with < and end with >. Inside the metadata line, the various types of information are structured as `FEATURE:VALUE` pairs, with a colon (`:`) as a separator between the feature and its value. Several

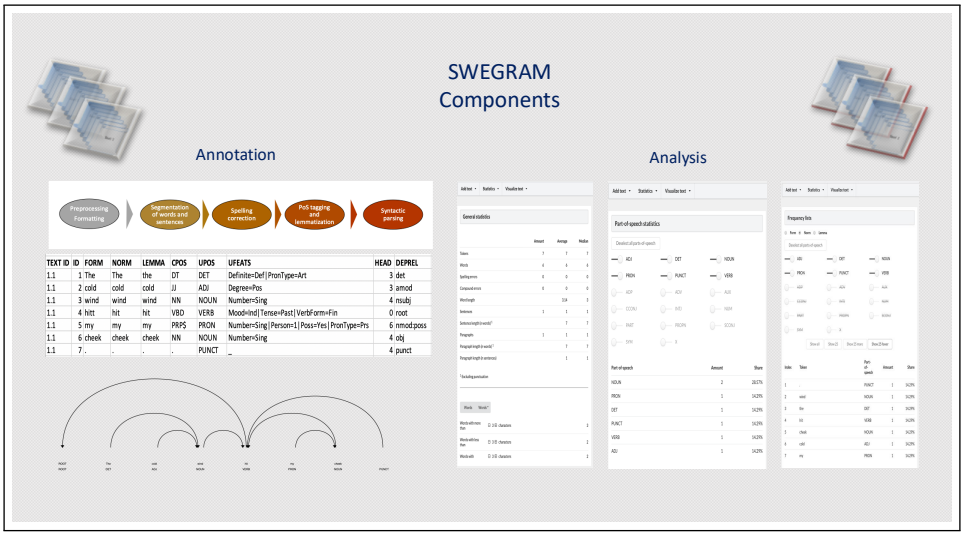


Figure 1: The SWEGRAM components.

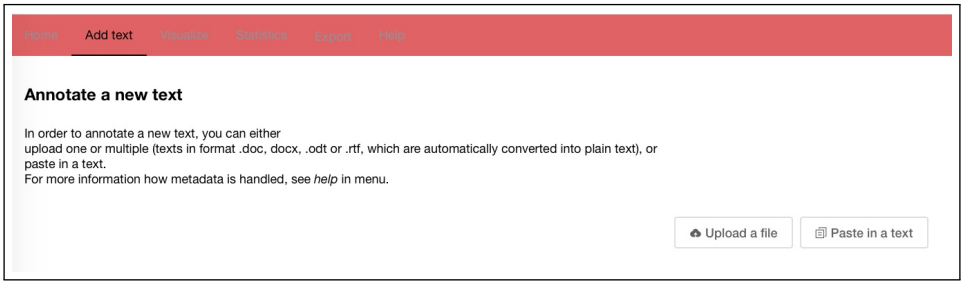


Figure 2: Upload texts.

pairs are separated by a semicolon (`;`). For example, a metadata line might appear as follows:

```
<AUTHOR:John Joe; GENRE:Narrative; YEAR:2023>
```

In this example, `AUTHOR`, `GENRE`, and `YEAR` are features, while `John Joe`, `Narrative`, and `2023` are their corresponding values.

If the uploaded file contains several texts, the metadata is displayed at the beginning of each text, as shown in Figure 3b. The metadata line can be also used for text selection, as illustrated in Figure 4. The system filters texts that match the specified feature-value pairs and proceeds with them for further analysis.

Lines can be added as additional information without being treated as part of the text to be analyzed by marking them with a hashtag (#). SWEGRAM excludes all lines starting with a hashtag during processing.

```
1 <FEATURE1:VALUE1;FEATURE2:VALUE2>
2
3 This is a text to be annotated and analyzed.
```

a. *Plain text*



b. *Visualized text with metadata*

Figure 3: Metadata label format in text files and in analysis.

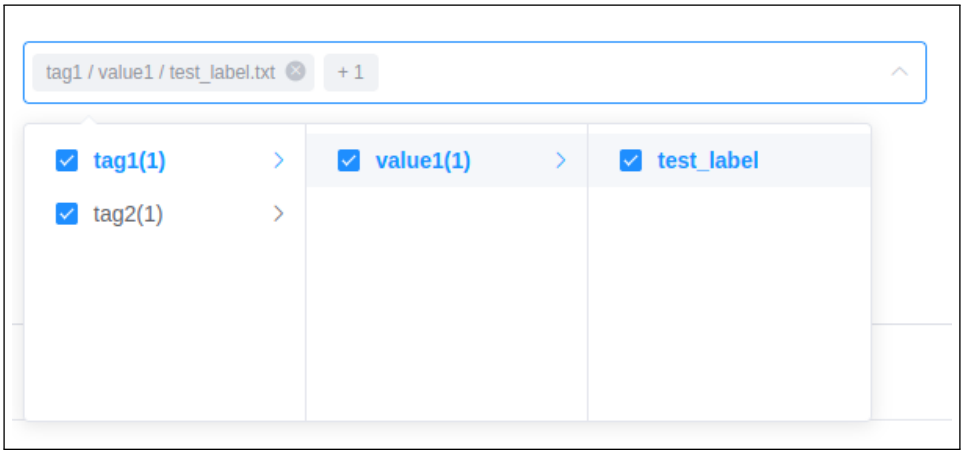


Figure 4: Metadata for selection.

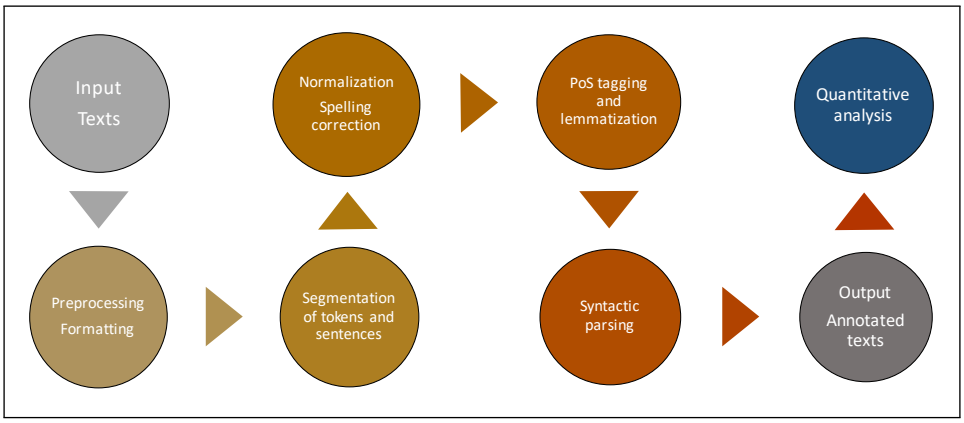


Figure 5: The components for linguistic annotation in SWEGRAM.

SWEGRAM can also be used to modify or add annotations to an already annotated text by selecting the "This is an annotated text" option. For this purpose, the CSV file format is accepted. When enabling this option, the uploaded file will not go through the standard annotation process; instead, the system will call for the annotation checker to ensure that the file is in the required format. All errors detected by the annotation checker are displayed with an index reference, indicating the type of errors that the file might have.

4.2 *Linguistic annotation*

Once the files are uploaded, linguistic annotation can be performed. SWEGRAM provides a comprehensive chain of tools to annotate a text on various linguistic levels. These include tokenization for the segmentation of words and sentences, normalization to correct misspelled words, and PoS tagging/-parsing for the morpho-syntactic analysis of sentences.

The components of the annotation pipeline are illustrated in Figure 5. This pipeline is built on a chain of automated tools for the analysis of the English and Swedish languages. We use state-of-the-art tools developed in the fields of natural language processing to automatically process and annotate texts with proven accuracy.

The system processes all running texts in the file except the lines marked with a hashtag (#).

Texts annotated using the tool are represented in a standardized format, with annotations at various linguistic levels adhering to specific guidelines. We use Unicode (UTF-8) for character encoding. To represent annotations at various linguistic levels and comply with international standards, we use

Table 1: Annotation representation.

TEXT-ID	Paragraph and sentence index, integer starting at 1 for each new paragraph and sentence.
ID	Token index, integer starting at 1 for each new sentence; may be a range for originally multiword tokens that have been split due to misspelling.
FORM	Word form or punctuation mark, so called token.
NORM	Corrected/normalized token for misspelled words.
LEMMA	Lemma of word form.
UPOS	Part-of-speech based on **Universal PoS tagset**.
XPOS	Part-of-speech based on the **Stockholm-Umeå Corpus** PoS tagset for Swedish and **Penn Treebank** PoS tagset for English.
CFEATS	List of morphological features from the **Stockholm-Umeå Corpus**; "_" if feature is missing. This is only valid for Swedish text analysis.
UFEATS	List of morphological features from the **Universal PoS tagset**; "_" if feature is missing.
HEAD	Head of the current word, which is either a value of ID or zero (0) if the word is the ROOT of the sentence.
DEPREL	Dependency relation to the HEAD based on the **Universal dependency relations**.
MISC	Any other annotation.

the CoNLL-U tab-separated format. Each token (i.e., each separated word and punctuation mark) occupies its own line along with its analysis, with each new sentence beginning on a new line shift.

Every token is analyzed at various levels, which are represented on the same line as the token itself in dedicated columns, separated by tabs. These columns contain both the ID number for the paragraph, sentence, and token, and the linguistic analysis. The linguistic analysis is represented at both the word level, through lemma, PoS annotation and morphological information, and the sentence level, through syntactic analysis. Table 1 provides a summary of annotation representation. On the left-hand side, the names of the features listed for each token in the columns are described, followed by the description of each feature.

Output from the linguistic annotation is illustrated for the sentence *Lixards like to eat bugs.*, which includes the misspelled token *Lixards* that SWEGRAM's normalization module corrected to *Lizards*. The linguistic annotation labels in the CoNLL-U format are shown in Figure 6 and the syntax tree with PoS tags and syntactic functions in Figure 7.

Next, we describe each component involved in the linguistic analysis in detail.

TEXTID	ID	FORM	NORM	LEMMA	UPOS	XPOS	U-FEATS	HEAD	DEPREL	MISC
1.1	1	Lixards	Lizards	Lizard	NOUN	NNS	Number=Plur	2	nsubj	_
1.1	2	like	like	like	VERB	VBP	Mood=Ind|Tense=Pres|VerbForm=Fin	0	root	_
1.1	3	to	to	to	PART	TO	_	4	mark	_
1.1	4	eat	eat	eat	VERB	VB	VerbForm=Inf	2	xcomp	_
1.1	5	bugs	bugs	bug	NOUN	NNS	Number=Plur	4	obj	_
1.1	6	.	_	.	PUNCT	.	_	2	punct	_

Figure 6: Output – linguistic annotation labels in CoNLL-U format.

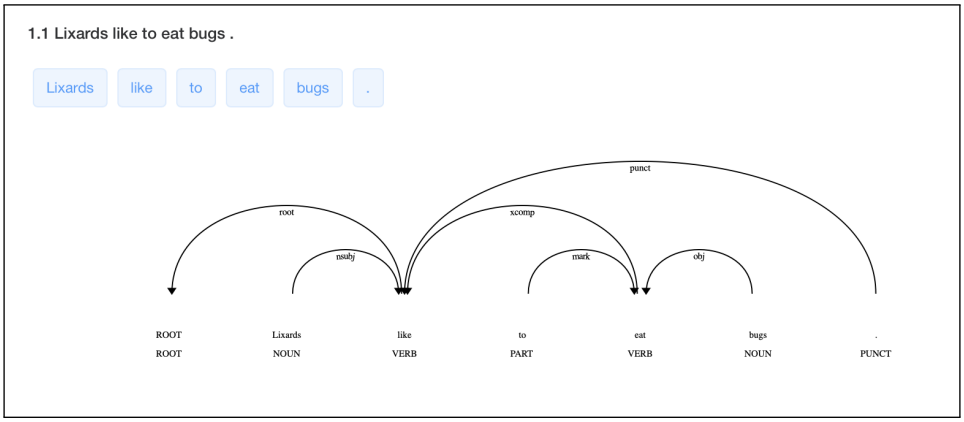


Figure 7: Output – annotated syntax tree.

4.2.1 *Tokenization and sentence segmentation*

In the formatted and uploaded text, words are separated from punctuation and segmented into linguistic units using tokenization, leaving each word and punctuation mark on its own line and each sentence followed by an empty line, see Figure 8. Each token in the sentence is then numbered in numerical order to receive an ID-number. Tokenization and sentence segmentation are based on the tokenization module used for the Swedish Treebank (Cap et al. 2016), which is also a component of EFSELAB (Östling 2016) applied for linguistic analysis.

4.2.2 *Normalization, spelling and separated compound words*

Misspelled words are identified and corrected by the normalization component. The normalized word is added to the analysis and displayed in a separate field to ensure that the original word form is kept and displayed. In Figure 9, the original word form is shown in the `FORM` column, while the corrected, normalized version appears in the `NORM` column. This shows the original and corrected text side by side, token by token.

```
This
is
e.g.
tokenized
.

The
sentences
are
segmented
.
```

Figure 8: An example of a tokenized and sentence segmented text.

English translation	TEXT ID	ID	FORM	NORM
My	1.1	1	Min	Min
	1.1	2–3		inspirationskälla
inspiration	1.1	2	inspirations	
source	1.1	3	källa	
comes	1.1	4	kommer	kommer
from	1.1	5	från	från
text	1.1	6	texten	texten
.	1.1	7	.	.

Figure 9: Normalized words (NORM).

The module also corrects incorrectly separated compound words in Swedish, see Figure 9. In such cases, the corrected form is indexed with a hyphen (2–3) and shown in the NORM column in Figure 9. The corrected version is used for further annotation, and the original form is used for linguistic analysis.

Spelling corrections are handled by a modified version of Hist-Norm (Pettersson et al. 2013), a tool designed for normalizing historical text spellings. While effective, the tool may not correct every error, so manual review is highly recommended for reliable analysis.

4.2.3 *Part-of-speech analysis*

The corrected, normalized text is analyzed using a part-of-speech (PoS) tagger, which annotates each token—both words and punctuation—with its appropriate PoS tag and morphological information. PoS annotation includes

Table 2: Universal PoS tags (UPOS).

Tag	Explanation	Example
ADJ	Adjective	*fine*
ADP	Adposition (preposition)	*on*
ADV	Adverb	*quick*, *very*
AUX	Auxiliary verb	*have*
CCONJ	Conjunction	*and*
DET	Article/Determiner	*a*
INTJ	Interjection	*yeah*
NOUN	Noun	*car*
NUM	Numeral	*two*
PART	Particle	*out*
PRON	Pronoun	*he*
PROPN	Proper noun	*Jenny*
PUNCT	Punctuation mark	, .
SCONJ	Subordinating conjunction	*that*
SYM	Symbol	☺
VERB	Verb	*in*
X	Other	*xbbe*

two types: universal tagsets, as defined by the Universal Dependency (UD) framework (Nivre et al. 2016) and listed in Table 2 as well as language-specific tagsets. The Stockholm-Umeå Corpus (SUC) tagset (Ejerhed et al. 1992, Gustafson-Capková & Hartmann 2006) is utilized as the language-specific tagset for Swedish as specified in Table 3, and the Penn Treebank Tagset (Mitchell et al. 1993) for English as illustrated in Table 4.

The PoS analysis with UD tags (see Table 2) is shown in column UPOS, and the language specific PoS tags (see Table 3 and 4) are shown in the column XPOS.

Morphological analysis reproduces a set of features that describe the word's lexical and grammatical characteristics. Lexical features include subcategories of PoS, such as nouns/proper nouns/types of proper noun. Grammatical features describe the categorization of PoS for a given word form; for example, gender, numeral expression, case and species for nominal word classes (noun and pronoun) or case, species, comparative degree and numeral expression for adjectives. We use the UD formalism for both English and Swedish, with an additional morphological feature set using SUC for Swedish.

Universal morphological features are stated in the form of Feature = Value pairs, where the feature indicates the morphological category (which may

Table 3: PoS tags in SUC 2.0 (XPOS).

Tag	Explanation	Example
AB	Adverb	*inte* 'not'
DT	Determiner	*ett* 'an'
HA	Interrogative/relative adverb	*när* 'when'
HD	Interrogative/relative determiner	*vilken* 'which'
HP	Interrogative/relative pronoun	*som* 'as'
HS	Interrogative/relative possessive pronoun	*vars* 'whose'
IE	Infinitive marker	*att* 'to'
IN	Interjection	*ja* 'yes'
JJ	Adjective	*fin* 'pretty'
KN	Conjunction	*och* 'and'
MAD	Major delimiter	. ? ! :
MID	Minor delimiter	, - ; / *
NN	Noun	*bil* 'car'
PAD	Pairwise delimiter	( )
PC	Participle	*dansande* 'dancing'
PL	Particle	*in* 'in'
PM	Proper noun	*Jenny*
PN	Pronoun	*han* 'he'
PP	Preposition	*på* 'on'
PS	Possessive pronoun	*hennes* 'her'
RG	Cardinal number	*två* 'two'
RO	Ordinal number	*andra* 'second'
SN	Subjunction	*att* 'to'
UO	Foreign word	*nota bene*
VB	Verb	*fira* 'celebrate'

be in abbreviated form) and the value describes the actual features of the word. Some of the most important morphological features in UD for English and Swedish are listed in Table 5.

Unlike UD, SUC does not explicitly state the name of the morphological feature but only provides its corresponding value. In contrast, UD spells out both the feature name and its value. SUC's morphological features are described in Table 6.

The interested reader is referred to UD's description (Nivre et al. 2016), the SUC 2.0 Manual (Gustafson-Capková & Hartmann 2006) and the Penn Treebank documentation (Mitchell et al. 1993) for further details. Annotation of PoS with morphological features is performed with the aid of EFSELAB (Östling 2016) and represented in the column UFEATS for the UD annotation and CFEATS for the SUC morphological annotation, as illustrated in Figure 10.

Table 4: Penn Treebank PoS tagset (`XPOS`.)

PoS Tag	Description
CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PRP	Personal pronoun
PP$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	infinitival to
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Wh-determiner
WP	Wh-pronoun
WP$	Possessive wh-pronoun
WRB	Wh-adverb

Table 5: Examples of morphological features in UD (UFEATS).

Feature	Value	Explanation	PoS
Case	Nom	Nominative	ADJ, NOUN, PRON, PROPN
	Acc	Accusativ	
	Gen	Genitive	
Definiteness	Ind	Indefinite	ADJ, DET, NOUN, PRON, PROPN
	Def	Definite	
Gender	Com	Utrum	ADJ, DET, NOUN, PRON, PROPN
	Neut	Neutrum	
Number	Sing	Singular	ADJ, DET, NOUN, PRON, PROPN
	Plur	Plural	
Possessive	Yes	Possessive	DET
Degree	Pos	Positive	ADJ, ADV
	Cmp	Comparative	
	Sup	Superlative	
Mood	Ind	Indicative	AUX, VERB
Tense	Pres	Present	AUX, VERB
	Past	Past	
Verb form	Fin	Finite	AUX, VERB
	Inf	Infinite	
Voice	Act	Active	AUX, VERB
	Pass	Passive	
Abbreviation	Yes	Abbreviation	ADV

4.2.4 *Lemmatization*

The lemma of each word is identified, reducing each token to its base form, known as its "lemma." Unlike stemming, which simply removes word endings, lemmatization takes into account the context and meaning of the word to return a valid dictionary form, as shown in the column labeled LEMMA in Figure 10. Lemmatization uses vocabulary and morphological analysis to accurately transform words based on their part of speech and is performed in conjunction with PoS tagging, utilizing EFSELAB (Östling 2016).

Table 6: Morphological features in SUC 2.0 (CFEATS).

Feature	Value	Explanation	PoS
Gender	UTR	Common/Utrum	DT, HD, HP, JJ, NN, PC, PN, PS, (RG, RO)
	NEU	Neuter	
	MAS	Masculine	
Number	SIN	Singular	DT, HD, HP, JJ, NN, PC, PN, PS, (RG, RO)
	PLU	Plural	
Definiteness	IND	Indefinite	DT, (HD, HP, HS), JJ, NN, PC, PN, (PS, RG, RO)
	DEF	Definite	
Case	NOM	Nominative	JJ, NN, PC, PM, (RG, RO)
	GEN	Genitive	
Tense	PRS	Present	VB
	PRT	Past	
	SUP	Supinum	
	INF	Infinite	
Voice	AKT	Active	VB
	SFO	Passive	
Mood	KON	Conjunctive	VB
Participle	PRS	Present	PC
	PRF	Perfect	
Degree	POS	Positive	(AB), JJ
	KOM	Comparative	
	SUV	Superlative	
Pronoun	SUB	Subject form	PN
	OBJ	Object form	
	SMS	Compound	All PoS

```
TEXTID  ID FORM        NORM     LEMMA   UPOS    XPOS   U-FEATS
1.1     1  Lixards     Lizards  Lizard  NOUN    NNS    Number=Plur
1.1     2  like        like     like    VERB    VBP    Mood=Ind|Tense=Pres|VerbForm=Fin
1.1     3  to          to       to      PART    TO     _
1.1     4  eat         eat      eat     VERB    VB     VerbForm=Inf
1.1     5  bugs        bugs     bug     NOUN    NNS    Number=Plur
1.1     6  .           _        .       PUNCT   .      _
```

Figure 10: Linguistic annotation labels for lemma (LEMMA), PoS (UPOS, XPOS) and morphological features (UFEATS).

4.2.5 *Syntactic analysis*

Finally, sentences are syntactically analyzed by a parsing module that identifies the relationships between words and their grammatical functions within the sentence. This analysis follows the Universal Dependency formalism, version 2 (UD2)[3]. In the UD framework (Nivre et al. 2016), words are linked in pairs, forming binary relationships where one word serves as the head and the other as its dependent. The verb typically occupies the central role in the sentence and acts as the head. Unlike other dependency formalisms, where the finite verb (including auxiliary verbs) is the head, in UD Version 2 (UD2), and as used by SWEGRAM, the verb carrying the semantic content is considered the head. All other tokens (including words and punctuation marks) are connected either directly or indirectly to the semantic verb through directed arcs. In UD2, these arcs point from the dependent words to their respective heads, showing the relationship between the head and its dependents through linked arcs, with words acting as nodes. The syntactic or grammatical functions are typically specified on these links between each head-dependent pair. The syntactic dependency relationships are listed in Table 7.

We illustrate the syntactic analysis for the sentence *The cold wind hit my cheek* both in CoNLL-U format (shown in Figure 11) with the syntactic analysis in the columns `HEAD` and `DEPREL` and as a dependency graph in Figure 12. Each word in the sentence forms a pair, with one word serving as the head and the other as its dependent, linked by a specific syntactic function. Dependents point toward their head words, with the syntactic relationship clearly stated. Every sentence is organized around an intended root node, labeled as 0, which serves to connect the words in the sentence. Each word is sequentially numbered based on its ID (in our example, 1–7). The head word in the sentence is the semantic verb *hit* (word number 4), which has several direct dependents: the subject of the clause *wind* (`NSUBJ`), the direct object *cheek* (`OBJ`), and the punctuation mark "." (`PUNCT`). In turn, these words also have their own dependents. For instance, *wind* (word number 3) has *The* (word 1) and *cold* (word 2) as its dependents, functioning as determiner (`DET`) and adjectival modifier (`AMOD`), respectively. Similarly, *my* (word number 6) has *cheek* (word number 7) as its head.

4.3 *Correction of automatic analysis*

Modularity has been a key factor in the development of the tool, allowing users to customize the annotation process by enabling or disabling specific

3 https://universaldependencies.org/v2/

Table 7: Syntactic relations in UD2 (DEPREL).

Name	Description	Name	Description
acl	clausal modifier of noun (adjectival clause)	fixed	fixed multiword expression
advcl	adverbial clause modifier	flat	flat multiword expression
advmod	adverbial modifier	goeswith	goes with
amod	adjective modifier	iobj	indirect object
appos	appositional modifier	list	list
aux	auxiliary verb	mark	infinitive marker
case	case marking	nmod	nominal modifier
cc	conjunction	nsubj	nominal subject
ccomp	clausal complement	nummod	numerical modifier
clf	classifier	obj	object
compound	compound	obl	oblique nominal
conj	conjunction	orphan	orphan
cop	copula	parataxis	parataxis
csubj	clausal subject	punct	punctuation
dep	unspecified	reparandum	reparation
det	determiner	root	root
discourse	discourse element	vocative	vocative
dislocated	dislocated element	xcomp	open clausal complement
expl	expletive		

ID	FORM	NORM	LEMMA	U-POS	X-POS	U-FEATS	HEAD	DEPREL
1	The	The	the	DET	DT	Definite=DEF|PronType=Art	3	det
2	cold	cold	cold	ADJ	JJ	Degree=Pos	3	amod
3	wind	wind	wind	NOUN	NN	Number=Sing	4	nsubj
4	hitt	hit	hit	VERB	VBD	Mood=Ind|Tense=Past|VerbForm=Fin	0	root
5	my	my	my	PRON	PRPS	Number=Sing|Person=1|Poss=Yes|PronType=Prs	6	nmod:poss
6	face	face	face	NOUN	NN	Number=sing	4	obj
7	.	.	.	PUNCT	.	_	4	punct

Figure 11: Annotation in CoNLL-U format.

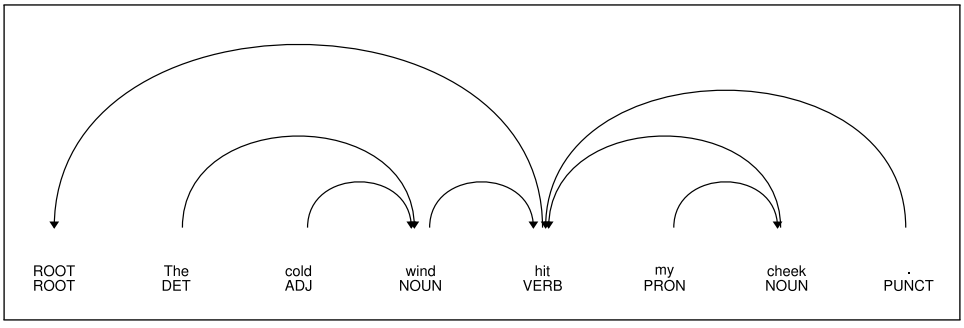


Figure 12: Example of syntactic analysis with dependency relations.

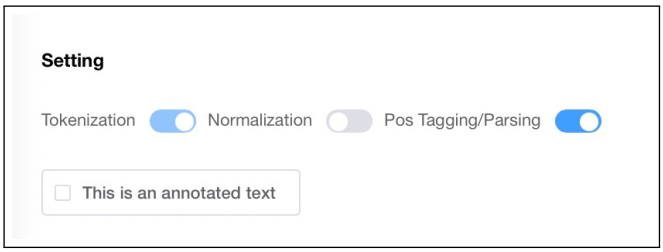


Figure 13: Choosing annotation components.

modules. This flexibility supports both tailored text analysis and manual correction of automatically assigned labels for subsequent processing.

The tool performs automatic linguistic analysis through several core components: tokenization, normalization, PoS tagging, and syntactic analysis (parsing), as shown in Figure 13. By default, all components except normalization are activated from the start.

To facilitate manual correction, users can download the automatically annotated file, review incorrectly assigned labels, and replace them with the correct ones, ensuring consistency with the tagset defined in the corresponding column. This step enhances the accuracy of annotations before further processing, making the tool adaptable to specific research needs.

Once corrections have been made, users can upload the revised annotations in the CoNLL-U format, as previously defined, for further automatic analysis by downstream components. This approach enables a refined workflow, where automatic linguistic analysis can be manually corrected and seamlessly reintegrated into the processing pipeline, ensuring both accuracy and flexibility.

4.4 *Storage of annotated texts*

The annotated texts are stored in a database on the server for a limited period. The system automatically deletes them after a few days. Since user registration is not required, the system does not track either the annotated texts or their statistics. Therefore, users should download and save their annotated texts and analyses for future processing.

4.5 *Linguistic analysis*

Once the uploaded texts are analyzed, SWEGRAM can perform quantitative analysis of various linguistic features based on the annotations. Statistics are calculated at multiple levels, including individual texts, several texts within

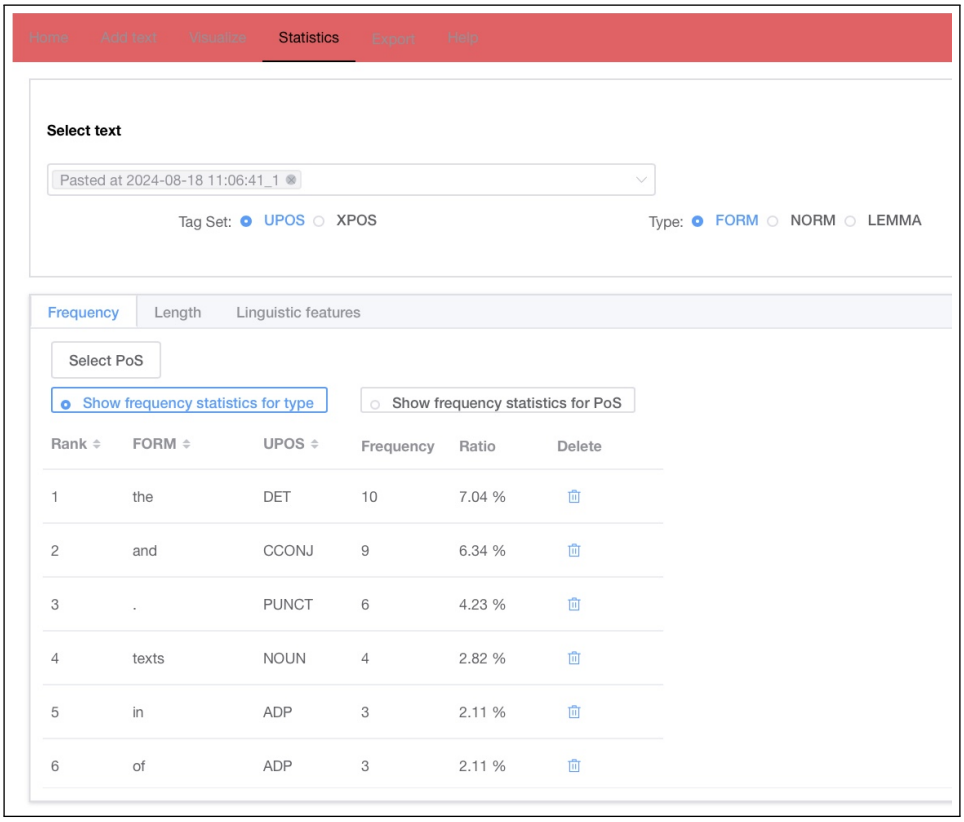


Figure 14: Frequency analysis.

a single file, or multiple uploaded files. General statistics can be obtained on word, sentence, and paragraph length, misspellings, PoS distribution, as well as readability metrics. Individual texts can also be studied using the visualization tool. Additionally, search filters can be applied to display statistics for selected texts among those that have been uploaded. Further details on statistics, filtering and visualization of various features are provided below.

4.5.1 *Frequencies*

SWEGRAM provides a display of tokens along with their PoS tags, which includes both the universal dependency tagset and a chosen language-specific tagset. The display also includes the total number of occurrences and the ratio (percentage) relative to the selected text(s). Figure 14 illustrates an example of the start of the frequency list for the abstract of this chapter, which will also serve as a demo text for the subsequent sections to illustrate the various features offered by SWEGRAM.

By default, the frequency analysis is based on the normalized word form (NORM). However, if this form is not available in the annotation, the analysis defaults to the surface form (FORM). Alternatively, the lemma can be selected in the top row of the frequency window. The frequency list can be sorted either alphabetically by token or by part-of-speech category.

Additionally, the frequency can be displayed separately for each PoS type, as shown in Figure 15. The search can be refined by deselecting unwanted PoS categories using the toggle buttons.

4.5.2 *Word length*

The statistics of word length excluding punctuation marks are shown in the *Length* feature set. SWEGRAM visualizes the number of tokens associated with specific PoS categories (displayed in columns) according to their character length, from 1 up to the length of the longest token in the text (displayed in rows). In the table, the rows represent word lengths, while the columns correspond to the parts of speech. The cells show the absolute frequency of tokens based on their PoS and length. The cells also show the specific words contributing to the counts and how often they appear in the selected texts. The list can be sorted by increasing or decreasing frequency.

Figure 16 presents the statistics for the length features based on the abstract of this chapter. In this specific example, there are eight nouns with a character length of four, six nouns with a length of five, two nouns with a length of six, and one noun with a length of three. The last row shows that there are 38 nouns in total. The last column lists the total number of words for each corresponding length.

4.5.3 *Syllables*

Some linguistic features and readability measures consider the number of syllables a word contains, and SWEGRAM estimates syllable counts using a rule-based heuristic approach for both English and Swedish. In English, the function `_syllable_count_en()` first converts the word to lowercase and determines the initial syllable count based on whether the first letter is a vowel. It then scans the word, increasing the count when a vowel is followed by a consonant, as this transition often marks the beginning of a new syllable. Additionally, it accounts for silent final *e*, subtracting one syllable if the word ends in *e*, and ensures that every word has at least one syllable by returning the higher value between the computed count and one. In Swedish, the function `_syllable_count_sv()` follows a similar approach but primarily counts the number of vowels in a word, as Swedish syllable structure typically

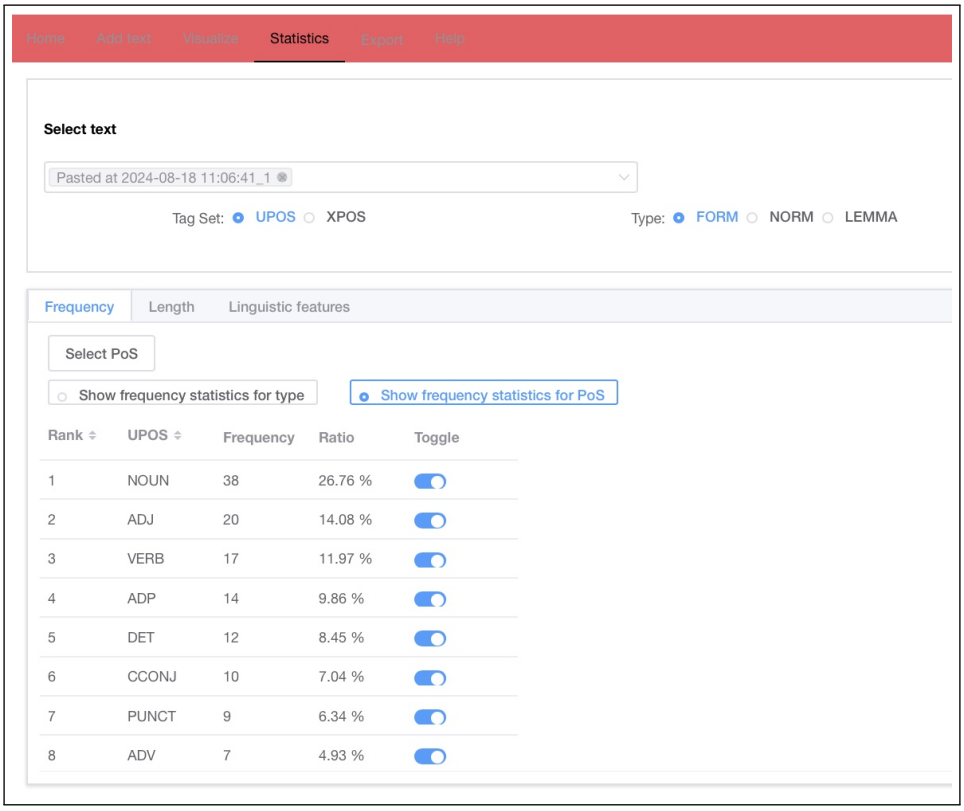


Figure 15: Frequency analysis of PoS categories.

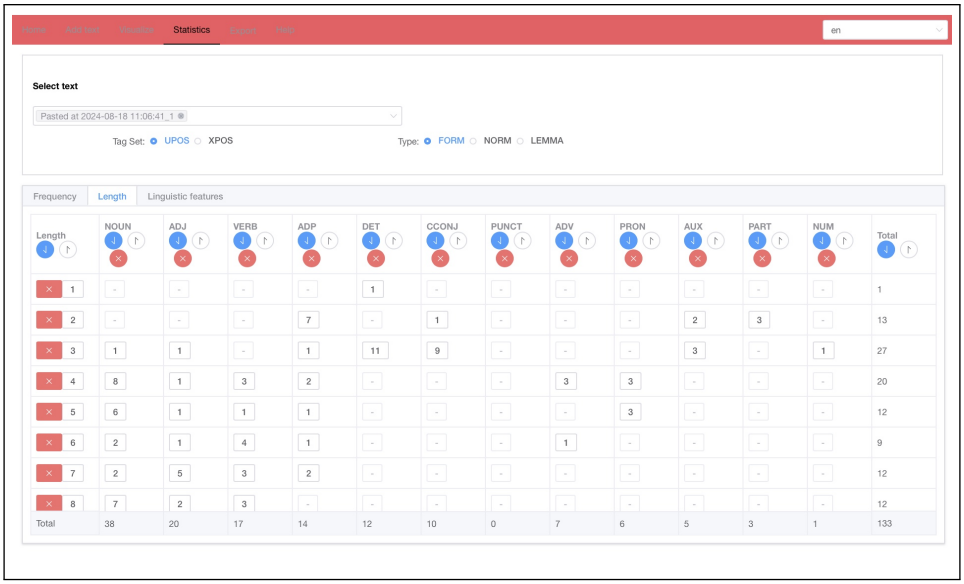


Figure 16: The analysis of character length sorted by PoS.

aligns with vowel occurrences. A notable exception is the word *journalist*, where the *ou* sequence is treated as a single syllable, reducing the overall count by one. If no vowels are found, the function guarantees a minimum syllable count of one. While these methods efficiently approximate syllable counts, they do not capture irregular pronunciations or complex phonetic patterns, making them useful but approximate tools for syllable extraction.

4.5.4 *Linguistic features*

In addition to frequency list of tokens and PoS tags, and the analysis of word length distribution of the uploaded text(s), SWEGRAM provides quantitative measurements on linguistic features generated from the annotated text, paragraph and sentence levels.

The content of the text under analysis—whether it is a full text, paragraph, or sentence—is displayed in the *Content* section, as illustrated in Figure 17 for the abstract of the article.

General features

Clicking on *Detail* in the top right corner displays the linguistic features in a popup window. The features for the specific content are computed immediately after the text is uploaded to the system. Once the level of analysis is selected, the statistics in *Overview* are displayed as a list consisting of *General*, *Readability*, *Lexical*, *Morphological*, and *Syntactic features*.

The general features of the text under analysis include details such as the total number of tokens, types, sentences, paragraphs, misspellings, separated compound words, as well as word, sentence, and paragraph lengths. Additionally, the median and mean values are provided, calculated based on the specific level at which the statistics are generated. Figure 18 illustrates the presentation of these general statistics for the abstract of this chapter.

Readability

SWEGRAM also calculates various readability, word variation, and nominality metrics. These are gathered under *Readability features*, which cover some common readability measures. Since readability measures are language dependent, two different sets have been developed, one for Swedish and another for English.

For Swedish, the Readability index (Lix) (Björnsson 1968), Word variation index (Ovix) (Hultman & Westman 1977), Type-token ratio (TTR) as discussed by (Johansson 2009), Simple nominal ratio (Af Geijerstam 2006: 108), and Full nominal ratio (Hultman & Westman 1977) are calculated.

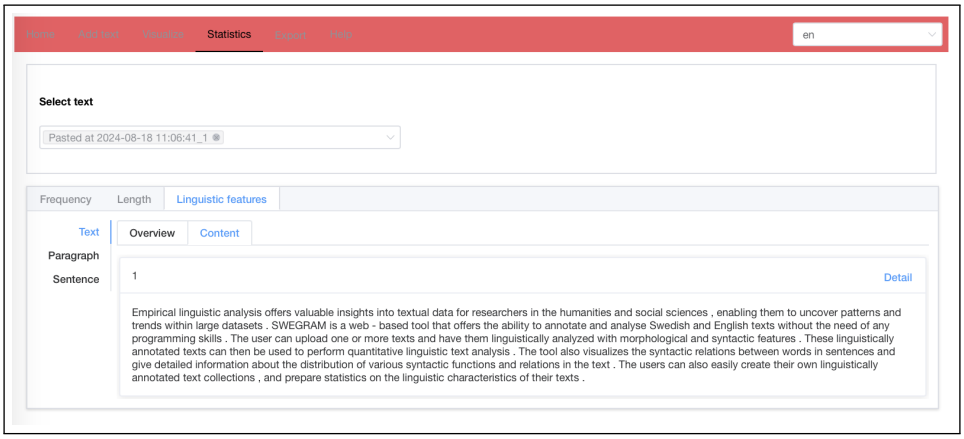


Figure 17: The analyzed text shown in *Content*.

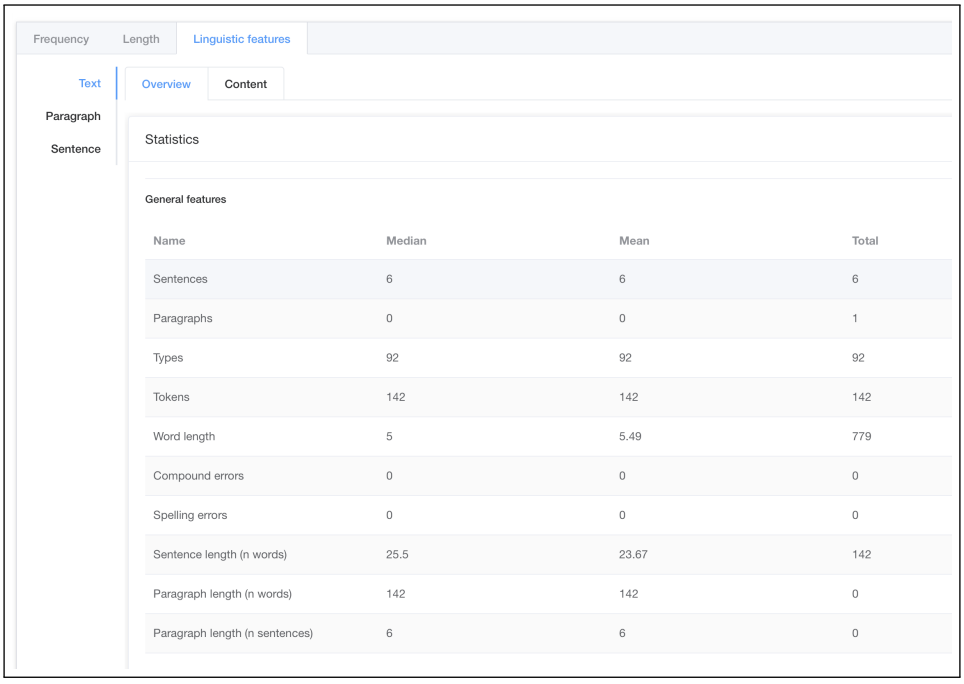


Figure 18: General features.

Table 8: Lix values, different genres.

Type of text	Lix	Words/ sentences	Long words	Interpretation
Books for children and young people	27	12	15	Very easy
Fiction	33	15	18	Easy
Daily and weekly press	39	14	25	Intermediate
Factual literature	47	18	29	Difficult
Nonfiction	56	20	35	Very difficult

Lix LIX, short for *läsbarhetsindex* in Swedish (readability index in English), is a formula used to evaluate the readability of a text, specifically how easy or difficult it is to comprehend (Björnsson 1968). It is calculated as shown in Equation 1, where a long word is defined as any word with more than six characters. The readability scores can be interpreted according to Table 8 (Melin & Lange 2000).

$$\text{LIX} = \frac{\text{total number of tokens}}{\text{total number of sentences}} + \frac{\text{total number of long tokens} \times 100}{\text{total number of tokens}} \quad (1)$$

Ovix Ovix stands for *ordvariationsindex* (English, word variation index) (Hultman & Westman 1977) and measures the total number of lexemes (minimal meaningful unit of language) in relation to the total number of tokens in a text. The formula used is defined in Equation 2. A higher Ovix value implies a greater word variation in relation to the length of the text (Hultman & Westman 1977: 56).

$$\text{OVIX} = \frac{\ln(\text{total number of tokens})}{\ln\left(2 - \frac{\ln(\text{total number of types})}{\ln(\text{total number of tokens})}\right)} \quad (2)$$

Type-token ratio The type-token ratio, abbreviated TTR, and discussed by Johansson (2009) is a metric used to measure the lexical diversity of a text. It is calculated by comparing the number of unique words (types) to the total number of tokens in a given text, as defined in Equation 3. A higher TTR indicates greater lexical diversity, meaning the text uses a wider variety of words while a lower TTR suggests that the text repeats the same words more often, indicating less lexical diversity. TTR can be affected by text length, as

shorter texts tend to have a higher TTR due to fewer opportunities for word repetition, while longer texts naturally tend to have a lower TTR.

$$\text{TTR} = \frac{\text{total number of types}}{\text{total number of tokens}} \tag{3}$$

Instead of using a simple ratio of the total number of types and tokens to mitigate the risk of the results affected by text length, bi-logarithm and square root TTR (see Equations 4) are also implemented. Here, logarithmic and square root transformations are applied to the number of types and the number of tokens, making the metrics more stable across texts of varying lengths.

$$f(\text{Bi-logarithm TTR}) = \frac{\ln(\text{total number of types})}{\ln(\text{total number of tokens})} \tag{4a}$$

$$f(\text{Square Root TTR}) = \frac{\text{total number of types}}{\sqrt{\text{total number of tokens}}} \tag{4b}$$

Simple nominal ratio The simple nominal ratio (Af Geijerstam 2006: 108) gives an indication of the text's information density by indicating the relationship between nouns (`NN`) and verbs (`VB`). The calculation of simple nominal ratio is defined in Equation 5. The higher percentage of nouns compared to verbs, the more information dense the text is.

$$\text{Simple nominal ratio} = \frac{\text{total number of nouns (NN)}}{\text{total number of verbs (VB)}} \tag{5}$$

Full nominal ratio The full nominal ratio offers an advanced measure of nominality. It highlights the relationship between nouns and noun-related parts of speech on one side, and verbs, adverbs, and pronouns on the other. Elements such as nouns (`NN`), prepositions (`PP`), and participles (`PC`) (both present and past) are associated with higher nominality and greater information density. In contrast, verbs (`VB`), adverbs (`AB`), and pronouns (`PN`) tend to dilute the informational content of a text. The formula for the full nominal ratio is presented in Equation 6.

$$\text{Full nominal ratio} = \frac{\text{total number of NN+PP+PC}}{\text{total number of VB+AB+PN}} \tag{6}$$

To measure readability for English texts, two types of type-token ratio, Bilogarithm and Root (Guiraud 1954), as well as the Coleman Liau Index (Coleman

& Liau 1975), Flesch Reading Ease (Flesch 1948), Flesch Kincaid Grade level (Kincaid et al. 1975), Automated Readability Index (Kincaid & Delionbach 1973), and Simple Measure of Gobbledygook (SMOG) index (McLaughlin 1969) are implemented.

Coleman Liau Index Coleman–Liau readability test (CLI) (Coleman & Liau 1975) is a formula used to evaluate the readability level of a text, specifically estimating the U.S. grade level required to understand the text. It is based on characters per tokens and sentence length indices and is calculated as shown in Equation 7. The formula first computes the average number of characters per 100 tokens to estimate word complexity. Then, it calculates the average number of sentences per 100 tokens to estimate sentence length and structure complexity.

$$\text{CLI} = 5.88 \times \frac{\text{total number of characters}}{\text{total number of tokens}} - 29.6 \times \frac{\text{total number of sentences}}{\text{total number of tokens}} - 15.8 \quad (7)$$

The resulting number is an estimate of the U.S. school grade level necessary to understand the text. The interpretation of the scores is shown in Table 9.

Flesch Reading Ease The Flesch Reading Ease (FRE) (Flesch 1948) analyzes the sentence lengths and word complexity by calculating the average length of the sentences measured by the number of words and the average number of syllables per word in the text.

$$\text{FRE} = 206.835 - 1.015 \times \frac{\text{total number of words}}{\text{total number of sentences}} - 84.6 \times \frac{\text{total number of syllables}}{\text{total number of words}} \quad (8)$$

It provides a score between 0 and 100; the higher the reading score, the easier the text is to read. The interpretation of the scores is described in Table 10.

Flesch Kincaid Grade level Flesch-Kincaid Grade Level (FKG) (Kincaid et al. 1975) is a variation of the Reading Ease formula with readjusted weights. The measure assesses the approximate reading grade level of a text, based on average sentence length and word complexity, see Equation 9.

Table 9: Coleman Liau Index levels.

Score	School level	Comprehension
5	5th grade and below	Very easy to read
6	6th grade	Easy to read
7	7th grade	Quite easy to read
7–10	8th, 9th, and 10th grade	Conversational English
11–12	11th and 12th grade	Quite hard to read
13–16	College	Difficult to read
17+	Professional	Very hard to read

Table 10: Flesch Reading Ease levels.

Score	Reading level
90–100	very easy to read, easily understood by an average 11-year-old
80–90	easy to read
70–80	fairly easy to read
60–70	easily understood by 13- to 15-year-old
50–60	fairly difficult to read
30–50	difficult to read, best understood by college graduates
0–30	very difficult to read, best understood by university graduates

Table 11: Flesch Kincaid scores and US grade levels.

Score	Reading level	School level	Age
0–3	Basic	Kindergarden	5–8
3–6	Basic	Elementary	8–11
6–9	Average	Middle school	11–14
9–12	Average	High school	14–17
12–15	Advanced	College	17–20
15–18	Advanced	Post-grad	20+

$$\text{FKG} = 0.39 \times \frac{\text{total number of words}}{\text{total number of sentences}} + 11.8 \times \frac{\text{total number of syllables}}{\text{total number of words}} - 15.59 \quad (9)$$

The scores correspond to US grade levels and age as listed in Table 11.

Automated Readability Index The Automated Readability Index (ARI) (Kincaid & Delionbach 1973) considers the average number of characters per

Table 12: ARI Scores and US grade levels.

ARI Score	Grade level	Reading level	Ages
1–5	Kindergarten	Extremely easy	5–6 y
1–5	1st grade	Extremely easy	6–7 y
6–7	2nd grade	Very easy	7–8 y
8–9	3rd grade	Very easy	8–9 y
10–11	4th grade	Easy	9–10 y
12–13	5th grade	Fairly easy	10–11 y
14–15	6th grade	Fairly aasy	11–12 y
16–17	7th grade	Average	12–13 y
18–19	8th grade	Average	13–14 y
20–21	9th grade	Slightly difficult	14–15 y
22–23	10th grade	Somewhat difficult	15–16 y
24–25	11th grade	Fairly difficult	16–17 y
26–27	12th grade	Difficult	17–18 y
28+	College	Very difficult	18–22 y

token and the average number of tokens per sentence. Characters include any letters, numbers, symbols, etc. with the exception for white space between characters.

$$\text{ARI} = 4.71 \times \frac{\text{total number of characters}}{\text{total number of tokens}} + 0.5 \times \frac{\text{total number of tokens}}{\text{total number of sentences}} - 21.43 \quad (10)$$

Table 12 summarizes the ARI scores and their corresponding interpretations.

SMOG The Simple Measure of Gobbledygook (McLaughlin 1969) is a measure of readability that estimates the years of education needed to understand a piece of writing. The SMOG formula scores a text based on the complexity of the sentences and words by considering the number of polysyllabic words, i.e. words with three or more syllables and the number of sentences in a text. The SMOG formula is defined In Equation 11.

$$\text{SMOG} = 1.043 \times \left(\sqrt{\frac{\text{total number of polysyllabic words} \times 30}{\text{total number of sentences}}} + 3.1291 \right) \quad (11)$$

The interpretation of the SMOG scores is detailed in Table 13.

Table 13: SMOG Scores and US grade levels.

SMOG Score	Approx. Grade level (+1.5 grades)
1–6	5
7–12	6
13–20	7
21–30	8
31–42	9
43–56	10
57–72	11
73–90	12
91–110	13
111–132	14
133–156	15
157–182	16
183–210	17
211–240	18

Table 14: Readability features for English texts. Scores are calculated on the chapter's abstract.

Name	Median	Mean	Total
SMOG	17.26	17.26	17.26
Root TTR	10.83	10.83	7.72
Bilogarithm TTR	0.98	0.98	0.91
Coleman Liau Index	17.30	17.30	17.30
Flesch Reading Ease	22.13	22.13	22.13
Flesch Kincaid Grade Level	15.68	15.68	15.68
Automated Readability Index	17.24	17.24	17.24

Each of the readability measures is calculated for the abstract of this chapter and shown in Table 14. All metrics indicate a professional text which is difficult to read with the exception of the ARI and SMOG scores, resulting in a grade level of 7.

Lexical features

To measure lexical diversity and proficiency, the Swedish KELLY-list[4] (Volodina & Kokkinakis 2012) is consulted. This list is based on the project *KEywords for Language Learning for Young and Adults Alike* (KELLY). The KELLY-list is

4 https://spraakbanken.gu.se/en/projects/kelly. Available on Sep 14, 2020.

Table 15: Example entries from the Swedish KELLY-list.

CEFR	Lemma	PoS	WPM
A1	*all* 'all'	pronoun	2975.47
	för 'for'	adverb	421.08
A2	*påstående* 'statement'	noun-ett	63.24
	för 'therefore'	conjunction	44.36
B1	*avgå* 'resign'	verb	23.46
	stabilitet 'stability'	noun-en	23.32
B2	*ödmjuk* 'humble'	adjective	11.86
	i natt 'tonight'	adverb	11.85
C1	*foster* 'fetus'	noun-ett	7.06
	enastående 'outstanding'	adjective	7.05
C2	*allergisk* 'allergic'	adjective	3.40
	eskalera 'escalate'	verb	3.40

a vocabulary sample that represents vocabulary usage, with each selected word or phrase classified into one of the six levels of the *Common European Framework of Reference for Languages* (CEFR) (Pilán & Volodina 2016) in ascending order of difficulty: A1, A2, B1, B2, C1, and C2. The Swedish KELLY-list contains 8,409 items generated from a corpus of web texts, as illustrated in Table 15 with examples of two entries for each CEFR level. Each word entry is represented by its lemma, part of speech, relative word frequency measured in words per million (WPM), followed by its English translation.

The distribution of lemmas according to their difficulty levels (A1, A2, B1, B2, C1, C2) found in the text is presented first, along with their median, mean, and total values. The number of "Difficult Words," referring to any token with a CEFR level of B1 or higher, and "Difficult NOUN & VERB," referring to nouns or verbs with a CEFR level of B1 or higher are provided. Finally, the number of tokens in the text that are not part of the KELLY list is listed. Table 16 illustrates the lexical features of the abstract of this chapter, indicating that the text contains many difficult words, both nouns and verbs.

Morphological features

The frequency of a part of speech or a morphological feature relative to other parts of speech or morphological features is measured using relative

Table 16: Lexical features. Scores are calculated on the chapter's abstract.

Name	Median	Mean	Total
A1 Lemma	415.49	415.49	415.49
A2 Lemma	133.80	133.80	133.80
B1 Lemma	98.59	98.59	98.59
B2 Lemma	14.08	14.08	14.08
C1 Lemma	14.08	14.08	14.08
C2 Lemma	49.30	49.30	49.30
Difficult Word	176.06	176.06	176.06
Out of Kelly list	274.65	274.65	274.65
Difficult Noun or Verb	126.76	126.76	126.76

frequencies, expressed through the incidence score (INCSC) (Graesser et al. 2004). The INCSC computation is defined as:

$$\text{INCSC} = 1000 \times \frac{N_t}{N_c} \tag{12}$$

where N_c is the number of tokens that belong to a certain category we are interested in and N_t is the number of tokens used for comparison. In most cases, N_c is a subset of N_t.

SWEGRAM covers 30 morphological features grouped into six subcategories based on part-of-speech (PoS) usage in the INCSC calculation. These linguistic features have been found to correlate with language development in both first and second language research (Pilán 2018). The features are defined using Universal Dependency (UD) tags.

SWEGRAM calculates the proportion of the following parts of speech relative to all tokens: adjectives (`ADJ`), adverbs (`ADV`), nouns (`NOUN`), particles (`PART`), verbs (`VERB`), punctuation (`PUNCT`), and subordinating conjunctions (`SCONJ`).

In addition, SWEGRAM analyzes the distribution of one part of speech in relation to others (PoS-PoS), namely: noun-to-verb (`NOUN - VERB`), pronoun-to-noun (`PRON - NOUN`), and pronoun-to-preposition (`PRON - ADP`).

SWEGRAM also computes the proportion of three specific subcategories of parts of speech (SUBPoS) relative to all tokens in the text: passive verb forms (s-verbs), third-person singular personal pronouns (`PRON 3SG`), and neuter nouns (`NOUN NEU`).

Additionally, SWEGRAM calculates the proportion of adjectives (`ADJ`), adverbs (`ADV`), nouns (`NOUN`), and verbs (`VERB`) relative to the sum of these four parts of speech (PoS-Multiple PoS).

Table 17: Specification of categories of N_c and N_t given the specific morphological features. *ALL* refers to all parts of speech categories.

Feature	N_c	N_t
Functional token	Functional PoS	ALL PoS
Lexical token	Lexical PoS	ALL PoS
Conjunction och subjunction	PoS = CCONJ+SCONJ	ALL PoS
Interrogative and relative	PoS = Int+Rel	ALL PoS
Lexical - functional token	Lexical PoS	Functional PoS
Nominal - Verbal	PoS = NOUN+ADP+PC	PoS = PRON+ADV+VERB

Different types of verb forms (N_c) are analyzed with respect to modality, tense, and aspect in relation to the total number of verbs.

Finally, SWEGRAM also calculates the ratio of groups of parts of speech relative to other word groups. For instance, it can compute the ratio between lexical and functional words. Lexical parts of speech in the UD framework include adjectives (ADJ), adverbs (ADV), interjections (INTJ), nouns (NOUN), proper nouns (PROPN), and verbs (VERB). Functional parts of speech include prepositions (ADP), auxiliaries (AUX), coordinating conjunctions (CCONJ), determiners (DET), numbers (NUM), particles (PART), pronouns (PRON), subordinating conjunctions (SCONJ), punctuation (PUNCT), symbols (SYM), and other (X). The distribution of these parts of speech groups, based on N_c and N_t, is detailed in Table 17.

Table 18 summarizes the morphological features used in the abstract of this chapter.

Syntactic features

The syntactic analysis, as provided by the UD2 annotation, is based on the number of dependency arcs in the sentence, along with the type and frequency of the dependency relations. Five syntactic features are extracted from statistics on the dependency arcs:

- Dependence length: The total number of dependency arcs from each token in the sentence to the ROOT.
- Longest dependency length: The longest dependency arc from any token to the ROOT.
- Dependency arcs longer than 5: The total number of tokens with dependency arcs that exceed a length of 5.

Table 18: Lexical features. Scores are calculated on the chapter's abstract.

Name	Median	Mean	Total
POS-ALL ADJ INCSC	140.85	140.85	140.85
POS-ALL ADV INCSC	49.30	49.30	49.30
POS-ALL NOUN INCSC	267.61	267.61	267.61
POS-ALL PART INCSC	21.13	21.13	21.13
POS-ALL VERB INCSC	119.72	119.72	119.72
POS-ALL PUNCT INCSC	63.38	63.38	63.38
POS-ALL SCONJ INCSC	0	0	0
POS-POS NOUN to VERB	2235.39	2235.39	2235.39
POS-POS PRON to NOUN	157.89	157.89	157.89
POS-POS PRON to PREP	6000.00	6000.00	6000.00
SUBPOS-ALL S-VERB INCSC	0	0	0
POS-MultiPOS ADJ Variation	243.90	243.90	243.90
POS-MultiPOS ADV Variation	85.37	85.37	85.37
VERBFORM Past VERB to VERB	45.45	45.45	45.45
MultiPOS-MultiPOS Rel INCSC	7.04	7.04	7.04
POS-MultiPOS NOUN Variation	463.41	463.41	463.41
POS-MultiPOS VERB Variation	207.32	207.32	207.32
VERBFORM Modal VERB to VERB	227.27	227.27	227.27
VERBFORM Supine VERB to VERB	0	0	0
VERBFORM Present VERB to VERB	227.27	227.27	227.27
MultiPOS-MultiPOS Lex to Token	577.46	577.46	577.46
MultiPOS-MultiPOS Nominal Ratio	2000.00	2000.00	2000.00
MultiPOS-MultiPOS Lex to Non-lex	1366.67	1366.67	1366.67
VERBFORM Past Participle to VERB	181.82	181.82	181.82
VERBFORM Present Participle to VERB	0	0	0
MultiPOS-MultiPOS CCONJ SCONJ INCSC	70.42	70.42	70.42
MultiPOS-MultiPOS Functional Token INCSC	422.54	422.54	422.54

- Ratio of right dependency arcs: The proportion of right-directed arcs relative to the total number of arcs in the sentence.
- Ratio of left dependency arcs: The proportion of left-directed arcs relative to the total number of arcs in the sentence.

Syntactic complexity is also measured through six syntactic functions, quantified using N_c incidence scores. The syntactic relations analyzed include variation in modifiers (both pre-modifiers and post-modifiers), subordinate clauses, relative clauses, and prepositional complements.

- A modifier refers to any dependent whose dependency relation to its head is one of the following: adjectival modifier (`amod`), nominal modi-

fier (nmod), appositional modifier (appos), numeric modifier (nummod), adverbial modifier (advmod), or discourse element (discourse). A modifier occurring before its head is classified as a pre-modifier, while a post-modifier appears after its head.

Modifier Variation refers to the total number of modifiers, measured in relation to all syntactic functions.

- A subordinate refers to a dependency relation (DEPREL) annotated with one of the following labels: clausal subject (csubj), clausal complement (ccomp), open clausal complement (xcomp), adverbial clause modifier (advcl), clausal modifier of noun (acl), or relative adnominal clauses (acl:relcl) between the current token and any of its direct dependents.

- A relative clause refers to any dependent whose relation to its head includes the feature PronType=Rel, indicating that the Pronominal Type is a relative pronoun (PRON), determiner (DET), numeral, (NUM) or adverb (ADV). The N_c for the feature Rel Clause INCSC represents the number of tokens in relative clauses.

- A prepositional complement refers to a dependent whose dependency relation with its head is annotated as case. The N_c for the feature PREP Comp INCSC corresponds to the number of prepositional complements and other dependents subordinated under the heads of prepositional complements.

Table 19 provides a summary of the syntactic features of the abstract of this chapter.

4.6 *Viewing and exporting data*

The results from the analyzed texts are displayed directly under the *Statistics* and *Visualize* modules, providing insights into various linguistic characteristics of the selected texts. The *Statistics* module presents numerical data, such as word frequency, sentence length, and readability scores, offering a quantitative overview of the corpus. In contrast, the *Visualize* module transforms these data points into graphical representations, such as bar charts or graphs, making patterns and relationships more accessible and interpretable.

All data, including the annotated texts, statistical summaries, and visualizations, can be downloaded as a text file (.txt) or CSV file (.csv) for further analysis.

Table 19: Syntactic features. Scores are calculated on the chapter's abstract.

Name	Median	Mean	Total
Dependency length	8.00	8.00	8.00
Subordinate INCSC	225.35	225.35	225.35
Modifier variation	295.77	295.77	295.77
Pre-modifier INCSC	218.31	218.31	218.31
Post-modifier INCSC	77.46	77.46	77.46
Relative clause INCSC	133.80	133.80	133.80
Longest dependency length	8.00	8.00	8.00
Dependency arcs longer than 5	14.00	1400	35.00
Ratio of left dependency arcs	36.62	36.62	36.62
Prepositional complement INCSC	98.59	98.59	98.59
Ratio of right dependency arcs	63.38	63.38	63.38

5 *SWEGRAM in use*

SWEGRAM is a powerful tool for annotating and analyzing texts in both English and Swedish. It provides clear, easily interpretable linguistic annotations alongside statistical analyses derived from these annotations. As demonstrated in this chapter, SWEGRAM has broad applications across various fields, including corpus linguistics, text analysis, language education, and computational linguistics.

Researchers utilize SWEGRAM for quantitative linguistic analysis, extracting statistics on text length, word frequency, readability metrics, and syntactic complexity. It also supports text comparison and corpus creation, allowing users to analyze multiple texts with metadata to study genre variation, writing development, or historical language change. In educational settings, SWEGRAM is valuable for assessing students' writing skills, tracking syntactic development, and examining linguistic patterns in national exam essays. Additionally, the tool facilitates qualitative text analysis, enabling detailed investigations of specific grammatical structures and stylistic choices.

One of SWEGRAM's key functionalities is its ability to create corpora by allowing users to upload multiple text files, which are then processed through the SWEGRAM annotation pipeline. Users can attach metadata–such as author details (e.g., identity, gender, age, residence) and text attributes (e.g., text ID, year, genre, place of publication)—to enhance filtering and statistical analysis. A prominent example of SWEGRAM's application is its use in creating the Uppsala Corpus of Student Writings (Megyesi et al.

2016). This corpus comprises approximately 2,000 texts (about one million words) written by students for the national test in Swedish, spanning grades 3 to high school. The corpus includes both native Swedish speakers (L1) and learners of Swedish as a second language (L2). Each text was enriched with metadata, including the year and semester of production, text type (e.g., narrative, descriptive, investigative, argumentative), grade level, gender, and Swedish proficiency (L1 or L2). The full SWEGRAM annotation pipeline—covering tokenization, sentence segmentation, lemmatization, morphological analysis, part-of-speech tagging, and syntactic annotation following the Universal Dependencies standard—was applied. The Uppsala Corpus of Student Writings has been instrumental in quantitative empirical studies of student writing, examining linguistic variation across different groups based on gender, geographic location, age, and grade level in synchronic and diachronic perspectives. Designed as a monitor corpus, it supports longitudinal analyses of writing development over time.

SWEGRAM has been widely used in research on language and writing development, particularly in studies focusing on syntactic complexity across different age groups, educational levels, and text types. In a study by Anne Palmér, two essays from a national test in Swedish were compared in terms of lexical features, part-of-speech distribution, and syntax to demonstrate the potential applications of SWEGRAM (Näsman et al. 2017). Other studies have analyzed student texts from national exams in larger scale.

A study by Palmér & Hussenius (2022) examines the writing development of sixth-grade students, focusing on discursive writing and the linguistic tools they use. Using SWEGRAM for automated quantitative analysis, supplemented by minor qualitative assessments, the study analyzes 240 argumentative and narrative texts from Swedish national exams. Findings show that argumentative texts are shorter and have a more limited vocabulary than narrative ones, with a strong reliance on verbal structures. However, some students have started incorporating nominalization and denser informational content. Adverbs play a key role in structuring arguments, expressing evaluations, logical connections, and writer stance, while participles contribute to expressive and varied language, particularly in high-scoring texts. The study concludes that mastering verbal lexicogrammar is essential for writing development, with nominalization as the next step. The applied text measures effectively capture both genre and developmental aspects, providing insights for writing instruction and assessment.

A study by Bendegard et al. (2023) investigated productivity, syntactic complexity, and variation in third-grade student texts. The findings showed that while narrative texts were significantly longer than explanatory ones, they did not necessarily exhibit greater syntactic complexity. Instead, syntac-

tic variation differed by genre, suggesting that different text types activate distinct linguistic structures, which has implications for writing instruction and development. SWEGRAM was instrumental in computing key linguistic metrics such as total word count and average words per macro-syntagm.

Several master's and doctoral theses have also employed SWEGRAM to analyze linguistic patterns and writing development in Swedish. For example, Josefsson (2017) examined how SWEGRAM assigns part-of-speech tags to Swedish and Swedish as a second language student texts, particularly focusing on the classification of the word *så* as an adverb. In another study, Jönsson-Ahola (2024) conducted a comparative investigation of high school students' writing development between the fall and spring semesters of Swedish II. The analysis, based on 124 student texts, used SWEGRAM to compute LIX, OVIX, and full nominal ratios to measure linguistic complexity and syntactic variation. The study found that all quantitative measures increased between the two semesters, reflecting writing progression. A qualitative analysis further revealed distinct patterns of individual development, leading to discussions on how to assess and promote writing development effectively.

Beyond educational and research settings, SWEGRAM has been applied in workplace language development initiatives. For instance, the ESF-project *Kompetensutveckling i äldreomsorgen* (Competence Development in Elderly Care) led by Arbetsmarknadsförvaltningen (The Labour Market Administration), Stockholms stad (Stockholm City), aims to develop a model for language and professional competence development among elderly care staff. In this project, SWEGRAM has been used to assess participants' linguistic development. Due to time constraints, only text data were collected for later analysis. However, SWEGRAM's ability to measure language development through lexical variation and grammatical complexity has been recognized as a valuable tool for future evaluations. SWEGRAM has been particularly useful in this project because it enables teachers to analyze lexical variation automatically, eliminating the need for manual categorization. By calculating the distribution of different parts of speech, the tool facilitates rapid assessments of vocabulary changes over time and differences between participant texts within the same genre (personal communication with Ulf Sparredal, 2025).

SWEGRAM is expected to remain a helpful tool in linguistic research, writing development studies, and educational assessment. Its ability to process both small and large-scale text collections, generate comprehensive linguistic annotations, and support both quantitative and qualitative analyses makes it invaluable for researchers, educators, and language professionals. Whether used for corpus creation, writing assessment, workplace language development, or a combination of these, SWEGRAM provides powerful analytical capabilities that enhance a wide range of linguistic investigations.

6 *Released versions of SWEGRAM*

SWEGRAM is freely available in two versions: a web-based service accessible at http://swegram.ling.su.se and downloadable versions for local installation hosted on GitHub at https://github.com/bmegyesi/swegram-v2.

The web-based version operates without requiring an account or login, and no personal user information is stored. Uploaded texts remain on the server for a maximum of one week before being automatically deleted. As a result, users must re-upload their files if they wish to continue working on them at a later time.

In contrast, the downloadable versions allow users to save files directly to their own computers and are powered by Docker containers. One version features a command-line interface, while the other is a web-based version that mirrors the functionality of the online version. Both options have been thoroughly tested in Linux and macOS environments.

To run SWEGRAM locally, users need to install three additional packages: (i) Pandoc,[5] a universal document converter, (ii) EFSELAB (Östling 2018)[6] for morpho-syntactic annotation, and (iii) udpipe (Nivre et al. 2016)[7] for linguistic analysis. We recommend setting up a virtual environment to ensure smooth processing. Detailed installation and setup instructions are provided on the GitHub page.

The current online production environment for SWEGRAM is containerized and comprises three main components: the frontend user interfaces are built using a VUE project with an Nginx proxy; the backend API, responsible for text annotation and data processing, is powered by the FastAPI framework; and the latest MySQL image is used to store data and facilitates interactions with the backend API. Each container operates independently, allowing for separate maintenance and updates.

SWEGRAM is licensed under the Creative Commons CC BY-SA license, enabling users to freely use, share, copy, distribute, modify, and adapt the tool in various forms and for any purpose, including commercial use. Proper attribution to SWEGRAM is required for any use, distribution, or modification. Additionally, we kindly request that users reference this chapter and our previous publication (Megyesi et al. 2019) where applicable.

5 https://pandoc.org
6 https://github.com/robertostling/efselab
7 https://ufal.mff.cuni.cz/udpipe/1/install

7 Conclusion

This chapter provides a comprehensive overview of SWEGRAM, a tool designed for the automatic annotation and analysis of Swedish and English texts. SWEGRAM allows researchers in the humanities and social sciences to perform detailed linguistic analyses without requiring programming skills. The tool enables the upload of texts, which are then annotated for morphological and syntactic features, providing quantitative insights such as word frequencies, sentence structures, and readability metrics. Users can also visualize syntactic relations and create their own annotated corpora for further study. This chapter discusses the components, functionalities, and case studies involving SWEGRAM, highlighting its role in facilitating large-scale empirical linguistic research. The tool is available as both a web-based service and a downloadable version, making it accessible and flexible for various research needs.

Acknowledgments

We would like to extend our heartfelt gratitude to everyone who has contributed to the development of this tool. First and foremost, we are deeply thankful to Anne Palmér from the Department of Scandinavian Languages at Uppsala University for her insightful ideas and contributions to the initial version of SWEGRAM, as well as to Jesper Näsman for implementing the first online version of the tool. Our sincere thanks also go to Shifei Chen for upgrading the back-end, significantly improving the tool's processing speed. Additionally, we are grateful to the Master's program students in language technology at Uppsala University, who, over the years, have helped evaluate, debug, and provide invaluable user feedback. Lastly, our big thanks go to the reviewers and editors of the book for their thorough review and valuable feedback.

This work is part of Swe-CLARIN,[8] the Swedish branch of the European CLARIN infrastructure, and has received long-term financial support from the Swedish Research Council.

References

Af Geijerstam, Åsa. 2006. *Att skriva i naturorienterande ämnen i skolan* [Writing in science subjects in school]. Department of Linguistics & Philology,

8 https://sweclarin.se

Uppsala University. (Doctoral dissertation). https://uu.diva-portal.org/smash/record.jsf?pid=diva2%3A169321&dswid=4726.

Anthony, Laurence & Paul Baker. 2015. ProtAnt: A tool for analysing the prototypicality of texts. *International Journal of Corpus Linguistics* 20(3). 273–292. https://doi.org/10.1075/ijcl.20.3.01ant.

Bendegard, Saga, Josefin Lindgren & Anne Palmér. 2023. Produktivitet och syntaktisk komplexitet och variation i 9-åringars berättande och beskrivande texter [Productivity, syntactic complexity, and variation in 9-year-olds' narrative and descriptive texts]. In *Språk i praktiken: I en föränderlig värld* [Language in practice: In a changing world] (ASLA:s skriftserie 30), 39–59. ASLA, Svenska föreningen för tillämpad språkvetenskap. https://doi.org/10.17045/sthlmuni.24321526.

Bird, Steven, Ewan Klein & Edward Loper. 2009. *Natural language processing with Python: Analyzing text with the Natural Language Toolkit*. O'Reilly.

Björnsson, Carl Hugo. 1968. *Läsbarhet* [Readability]. Liber.

BNC Consortium. 2007. *British national corpus, XML edition*. Oxford Text Archive. http://www.natcorp.ox.ac.uk/corpus/index.xml?ID=consortium.

Borin, Lars, Markus Forsberg, Martin Hammarstedt, Dan Rosén, Roland Schäfer & Anne Schumacher. 2016. Sparv: Språkbanken's corpus annotation pipeline infrastructure. In *SLTC 2016. The sixth Swedish language technology conference, Umeå university, 17-18 november, 2016*. http://www8.cs.umu.se/~johanna/sltc2016/abstracts/SLTC_2016_paper_31.pdf.

Borin, Lars, Markus Forsberg & Johan Roxendal. 2012. Korp – the corpus infrastructure of Språkbanken. In *Proceedings of the 8th international conference on language resources and evaluation* (*lrec 2012*), 474–478. http://www.lrec-conf.org/proceedings/lrec2012/pdf/248_Paper.pdf.

Burnard, Lou. 2006. *Xaira: software for language analysis*. https://doi.org/10.4230/DagSemProc.06491.17.

Cap, Fabienne, Yvonne Adesam, Lars Ahrenberg, Lars Borin, Gerlof Bouma, Markus Forsberg, Viggo Kann, Robert Östling, Aaron Smith, Mats Wirén, et al. 2016. Sword: Towards cutting-edge Swedish word processing. In *SLTC 2016. The sixth Swedish language technology conference, Umeå, Sweden, 17-18 November, 2016*.

Coleman, Meri & Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology* 60. 283–284. https://api.semanticscholar.org/CorpusID:144250124.

Ejerhed, Eva, Gunnel Källgren, Ola Wennstedt & Magnus Åström. 1992. *The linguistic annotation system of the Stockholm-Umeå corpus project*. University of Umeå, Department of General Linguistics.

Flesch, Rudolph. 1948. A new readability yardstick. *Journal of Applied Psychology* 32(3)(221–233). https://doi.org/10.1037/h0057532.

Graesser, Arthur C, Danielle S McNamara, Max M Louwerse & Zhiqiang Cai. 2004. Coh-metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers* 36(2). 193–202. DOI: 10.3758/BF03195564.

Guiraud, Pierre. 1954. *Les caractères statistiques du vocabulaire: Essai de méthodologie* [The statistical characteristics of vocabulary: An essay on methodology]. Presses universitaires de France. https://books.google.se/books?id=XSUeAAAAIAAJ.

Gustafson-Capková, Sofia & Britt Hartmann. 2006. Manual of the Stockholm Umeå corpus version 2.0. *Unpublished Work*.

Hammarstedt, Martin, Anne Schumacher, Lars Borin & Markus Forsberg. 2022. *Sparv 5 user manual*. Tech. rep. Göteborg: University of Gothenburg, Sweden. https://hdl.handle.net/2077/73604.

Hoffmann, Sebastian, Stefan Evert, Nicholas Smith, David Lee & Ylva Berglund Prytz. 2008. *Corpus linguistics with bncweb – A practical guide*. Frankfurt am Main: Peter Lang. https://www.peterlang.com/document/1104732.

Honnibal, Matthew & Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. https://sentometrics-research.com/publication/72/.

Hultman, Tor G & Margareta Westman. 1977. *Gymnasistsvenska* [Upper secondary school Swedish]. Institutionen för nordiska språk, Lund.

Hutto, C.J. & Eric Gilbert. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media* 8(1). 216–225. DOI: 10.1609/icwsm.v8i1.14550.

Johansson, Victoria. 2009. *Developmental aspects of text production in writing and speech*, vol. 48. Department of Linguistics, Phonetics, Centre for Languages & Literature, Lund University. https://lucris.lub.lu.se/ws/files/5221582/1487260.pdf.

Jönsson-Ahola, Holger. 2024. *Den oundvikliga skriftspråksutvecklingen: en komparativ undersökning av gymnasieelevers skriftspråksutveckling mellan hösttermin och vårtermin under kursen svenska II* [The inevitable development of written language: A comparative study of upper secondary students' writing development between the fall and spring semester in the course Swedish II]. Bachelor's thesis, Stockholm University. https://su.diva-portal.org/smash/record.jsf?pid=diva2%3A1873436&dswid=-6759.

Josefsson, Eva. 2017. *"Så", ska det taggas som adverb? En granskning av hur annoteringsverktyget swegram ordklasstaggar elevtexter i svenska och svenska som andraspråk* ["Så", should it be tagged as an adverb? An examination of how the annotation tool SWEGRAM assigns part-of-speech tags to student texts in Swedish and Swedish as a second language]. Bachelor's thesis, Uppsala University.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý & Vít Suchomel. 2014. The Sketch Engine: Ten years on. *Lexicography* 1. 7–36. DOI: 10.1007/s40607-014-0009-9.

Kincaid, Peter J. & Leroy J. Delionbach. 1973. Validation of the automated readability index: A follow-up. *Human Factors* 15(1). 17–20. https://doi.org/10.1177/001872087301500103.

Kincaid, Peter J., Robert P. Jr. Fishburne, Richard L. Rogers & Brad S. Chissom. 1975. *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*. https://stars.library.ucf.edu/cgi/viewcontent.cgi?article=1055&context=istlibrary.

Laurence, Anthony. 2011. *Antconc : a learner and classroom friendly, multi-platform corpus analysis toolkit*. https://www.laurenceanthony.net/software/antconc/.

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard & David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for computational linguistics (acl) system demonstrations*, 55–60. DOI: 10.3115/v1/P14-5010.

McCallum, Andrew K. 2002. *MALLET: A machine learning for language toolkit*. http://mallet.cs.umass.edu.

McLaughlin, Harry G. 1969. Smog grading: A new readability formula. *Journal of Reading*. 639–646. http://www.jstor.org/stable/40011226.

Megyesi, Beáta, Jesper Näsman & Anne Palmér. 2016. The Uppsala corpus of student writings: Corpus creation, annotation, and analysis. In *Proceedings of the tenth international conference on language resources and evaluation (LREC'16)*, 3192–3199. Portorož, Slovenia: European Language Resources Association (ELRA). https://aclanthology.org/W17-0216/.

Megyesi, Beáta, Anne Palmér & Jesper Näsman. 2019. *Swegram: Annotering och analys av svenska texter* [SWEGRAM: Annotation and analysis of Swedish texts]. Department of Linguistics & Philology, Uppsala University, Sweden. https://uu.diva-portal.org/smash/record.jsf?pid=diva2%3A1326439&dswid=9822.

Megyesi, Beáta & Rex Ruan. 2024. *Swegram 2.0: Guidelines to annotation and analysis of English and Swedish texts*. Department of Linguistics, Stockholm

University, Sweden. https://www.diva-portal.org/smash/get/diva2:1908711/FULLTEXT01.pdf.

Melin, Lars & Sven Lange. 2000. *Att analysera text: Stilanalys med exempel* [Analyzing text: Stylistic analysis with examples]. Studentlitteratur.

Mitchell, Marcus P., Beatrice Santorini & Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19(2). 313–330. https://aclanthology.org/J93-2004.

Näsman, Jesper, Beáta Megyesi & Anne Palmér. 2017. SWEGRAM: A web-based tool for automatic annotation and analysis of Swedish texts. In *Proceedings of the 21st nordic conference on computational linguistics*, 132–141. Gothenburg, Sweden: Association for Computational Linguistics. https://aclanthology.org/W17-0216/.

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty & Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the tenth international conference on language resources and evaluation* (*LREC'16*), 1659–1666. Portorož, Slovenia: European Language Resources Association (ELRA). https://aclanthology.org/L16-1262/.

Östling, Robert. 2016. Shallow learning for sequence tagging. In *6th Swedish language technology conference* (*SLTC16*), *Umeå*, *Sweden*. https://people.cs.umu.se/johanna/sltc2016/abstracts/SLTC_2016_paper_10.pdf.

Östling, Robert. 2018. Part of speech tagging: Shallow or deep learning? *Northern European Journal of Language Technology* 5. 1–15. https://doi.org/10.3384/nejlt.2000-1533.1851.

Palmér, Anne & Siri Hussenius. 2022. Vad säger de nationella proven i årskurs 6 om elevers pågående skrivutveckling? [What do the national tests in year 6 reveal about students' ongoing writing development?] In *Fjortonde nationella konferensen i svenska med didaktisk inriktning – didaktiska perspektiv på språk och litteratur i en globaliserad värld.* (*smdi 14*) [The fourteenth national conference on Swedish with a didactic focus – didactic perspectives on language and literature in a globalized World.] Malmö universitet.

Pettersson, Eva, Beáta Megyesi & Joakim Nivre. 2013. Normalisation of historical text using context-sensitive weighted Levenshtein distance and compound splitting. In *Proceedings of the 19th Nordic conference of computational linguistics* (*Nodalida 2013*), 163–179. https://aclanthology.org/W13-5617/.

Pilán, Ildikó. 2018. *Automatic proficiency level prediction for intelligent computer-assisted language learning*. University of Gothenburg, Sweden. http://hdl.handle.net/2077/55895.

Pilán, Ildikó & Elena Volodina. 2016. Classification of language proficiency levels in Swedish learners' texts. In *Proceedings of Swedish language technology conference*. https://people.cs.umu.se/johanna/sltc2016/abstracts/SLTC_2016_paper_7.pdf.

Schmidt, Thomas & Kai Wörner. 2014. Exmaralda. In Jacques Durand, Ulrike Gut & Gjert Kristoffersen (eds.), *Handbook on corpus phonology*, 402–419. Oxford University Press. http://ukcatalogue.oup.com/product/9780199571932.do.

Scott, Mike. 2016. *WordSmith Tools Version 7*. Stroud: Lexical Analysis Software. https://lexically.net/wordsmith/version7/.

Volodina, Elena & Sofie Johansson Kokkinakis. 2012. Introducing the Swedish Kelly-list, a new lexical e-resource for Swedish. In *Proceedings of the eight international conference on language resources and evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2012/pdf/264_Paper.pdf.

Wittenburg, Peter, Hennie Brugman, Albert Russel, Alex Klassmann & Han Sloetjes. 2006. ELAN: A professional framework for multimodality research. In Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odijk & Daniel Tapias (eds.), *Proceedings of the fifth International Conference on Language Resources and Evaluation (LREC'06)*. Genoa, Italy: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2006/pdf/153_pdf.pdf.

List of abbreviations

ARI	Automated Readability Index
BNC	British National Corpus
CEFR	Common European Framework of Reference for Language
CFEATS	Language-specific Features
CLI	Coleman Liau Index
CSV	Comma-Separated Values
DEPREL	Dependency Relation
HEAD	Syntactic Head
INCSC	Incidence Score
KELLY	KEywords for Language Learning for Young and Adults Alike
LIX	Läsbarhetsindex (readability index)
NLP	Natural Language Processing

NLTK	Natural Language Toolkit
NORM	Normalized Token
Ovix	Ordvariationsindex (word variation index)
PoS	Part-of-Speech
SMOG	the Simple Measure of Gobbledygook
SUC	Stockholm Umeå Corpus
TTR	Type-Token Ratio
UD	Universal Dependency
UFEATS	Universal Features
UPOS	Universal Part-of-Speech
XML	Extensible Markup Language
XPOS	Language-specific Part-of-Speech

Corresponding author

Beáta Megyesi
Department of Linguistics
Stockholm University
beata.megyesi@ling.su.se

11 A practical guide to the Swedish L2 lexical profile

Therese Lindström Tiedemann
University of Helsinki

David Alfter
University of Gothenburg

Elena Volodina
University of Gothenburg

Vocabulary is a fundamental aspect of any language since without words you cannot communicate, nor learn other aspects of a language, such as grammar or pronunciation. The Swedish L2 profile offers many ways in which researchers can explore the vocabulary which learners can produce and are expected to understand at different proficiency levels. It also provides a foundation for innovative ways of teaching Swedish, for instance, through Computer Assisted Language Learning (CALL) and Data Driven Learning (DDL).

In this chapter we show how the lexical part of SweL2P can be used to explore the vocabulary growth of language learners both receptively and productively in a step-by-step overview. Starting from a bird's eye view of vocabulary in course books and learner essays we show how to zoom in on some specific aspects of vocabulary, choosing adjectives as an example. We use SweL2P to show how adjectives occur in course books and how they appear in learners' texts – comparing the lexis in both, but also showing the potential to explore the way learners acquire vocabulary more broadly. Finally, we present how results in SweL2P can be easily compared to other Swedish corpora.

1 *Introduction*

Vocabulary is a fundamental aspect of any language since without words you cannot communicate, nor can you learn other aspects of a language, such as grammar or pronunciation, without words to practise. The Swedish Second Language (L2) profile (SweL2P; Volodina, Alfter & Tiedemann 2024) is an open resource for Swedish as a second language which presents information about the occurrence of lexis, grammar and morphology in Swedish L2 course books and learner essays side by side in relation to the levels of the

HUM INFRA Therese Lindström Tiedemann, David Alfter & Elena Volodina. 2025. A practical guide to the Swedish L2 lexical profile. In Gerlof Bouma, Dana Dannélls, Dimitrios Kokkinakis & Elena Volodina (eds.), *Huminfra handbook: Empowering digital and experimental humanities* (NEALT Proceedings Series 59), 355–403. University of Tartu Library. DOI: 10.58009/aere-perennius0180 © The authors, CC BY 4.0

Common European Framework of Reference for languages (CEFR; Council of Europe 2020). It offers many ways in which researchers from different disciplines, but also language teachers, assessors and material designers, can explore what learners can produce and what they are expected to understand at different proficiency levels. SweL2P provides a new perspective that can have a substantial impact on language teaching and language learning, as well as on linguistic research.

In this chapter we show how the lexical part of SweL2P can be used to explore the vocabulary growth of language learners both receptively and productively in a parallel fashion which facilitates comparisons, something which according to usage-based perspectives can be important in relation to language learning (cf. Davidson et al. 2008). After a bird's eye view of vocabulary in course books (receptive) and learner essays (productive) we show how to zoom in on specific aspects of vocabulary in two case studies, to explore adjectives further for research purposes, or to support the teaching of Swedish as a second language. We also show how links in the resource lead the user to the empirical data in Korp,[1] Språkbanken Text's corpus management and search platform (Borin et al. 2025). There the user can explore the corpus data more closely and make comparisons to the results in other Swedish corpora, e.g. possible input such as newspapers.

The lexical SweL2P provides new opportunities for empirical research, but also new ways to digitally enhance usage-based language learning and teaching, e.g. through Data driven learning (DDL; see e.g., Gilquin & Granger 2010, Warren 2016, Boulton & Vyatkina 2021), while also showing a promising potential for future use in NLP applications such as Computer assisted language learning (CALL; see e.g., Mohsen et al. 2024) and automatic assessment.

The chapter starts with a section about the research context (Section 2). This is followed by a description of the SweL2P resource including the corpora used to create SweL2P (Section 3). In Section 4 we focus on the lexical profile, after which Section 5 explores some research and teaching scenarios briefly. We then continue with two case studies with guidelines on how to use SweL2P. First, we explore adjectives by frequencies in relation to linguistic research (Section 6), and then we present some ideas on how the subprofiles for adjectival declension, comparison and MWE can be useful in teaching (Section 7). After this we introduce some downstream NLP tasks for which the lexical profile can be used and we include links to downloadable datasets associated with the lexical profile (Section 8), before we end with some final remarks and future prospects (Section 9).

1 https://spraakbanken.gu.se/korp/

2 *Research context*

Second language acquisition (SLA) can be studied in various ways. Since the late 1980s it has become more and more common to try to study second language acquisition through corpora containing learners' own production. This type of research is nowadays often referred to as Learner corpus research (LCR). When learner corpora were introduced it was already evident that first language (L1) corpora could be of much use in relation to second language teaching, e.g. in designing teaching materials. However, as Granger (2002: 21) points out L1 corpora can only "provide valuable information on the frequency and use of words, phrases and structures" they cannot say anything about "the difficulty they [=the words] present for learners in general or for a specific category of learners". Specialists on vocabulary teaching have been claimed to agree that "both frequency and difficulty have to be taken into account" in language teaching (Granger 2002: 22) and to some extent it has been shown that frequency in corpora correlates to the likelihood that learners will use words (Davidson et al. 2008: 138). Learner corpora can provide empirical evidence of what learners might find difficult by studying what they do not use, and hence might not have learnt, or might be avoiding, as well as whether they use vocabulary items and vocabulary types in a manner somehow different to L1 speakers of the language.

In this chapter we first summarise some of the research done in connection to usage-based approaches and data driven learning (DDL) (Section 2.1) and vocabulary profiling (Section 2.2). In section 2.3 we introduce Contrastive interlanguage analysis (CIA) which has been used quite frequently in learner corpus research to compare L1 and L2, but also to compare the language of learners with different L1 backgrounds. Finally, section 2.4 presents some previous research on Swedish adjectives as background to the case studies in section 6.

2.1 *Usage-based approaches and data driven learning (DDL)*

This type of acquisition research fits particularly well with usage-based theories of language acquisition (e.g. Davidson et al. 2008, Wulff 2020, Römer 2023). In usage-based linguistics a prominent idea is that language learning is affected by language input as well as by learner output and general cognitive abilities. The type of language which we encounter as input and use ourselves is seen as likely to affect what we learn and when we learn things. This makes it highly relevant to be able to make comparisons between learner language and first language usage, but also between different proficiency levels.

In order to test the hypothesis that usage affects learning, it is important to have comparative data for different language usage. This is one of the reasons why we believe it is important to be able to compare course-book usage at different CEFR-levels to learners' production at different levels. In order to do this as well as possible we would need access to information about the course books which the learners who wrote the essays in the learner corpus used and preferably also what else they read and what they listened to, etc. This kind of information is not easy to come by and, instead, the SweL2P project chose to use a course-book corpus (COCTAILL, Volodina, Pilán, Eide et al. 2014) consisting of several course books per CEFR level, aimed at adult learners of Swedish, as a way of ascertaining what learners are expected to be able to understand (receptive knowledge). At the same time, this corpus also gives important insights into the input which learners normally get. To also be able to gauge the productive abilities of learners at different proficiency levels, we used a learner corpus for Swedish as a second language, (SweLL-pilot, Volodina, Pilán, Enström et al. 2016a, Volodina 2024). Both corpora are introduced further in Section 3.1.

2.2 *Vocabulary profiling*

Profiling has been described as the "identification and definition of developmental stages" (Granfeldt & Ågren 2014) and was initially used in relation to first language development and language disorders (Crystal et al. 1976) but it has also become asssociated with second language research (see e.g. Clahsen 1985, Pienemann & Mackey 1993, Granfeldt et al. 2005, Keßler & Liebner 2011, Hawkins & Filipović 2012, Granfeldt & Ågren 2014).

For Swedish important work has also been done on a typologically-based lexical profile based on translation corpora, in the work of Viberg (1990, 1992, 2006b,a, 2013), with an aim to support L2 research. This perspective lists the most common words in a certain part of speech and/or in a particular semantic field by frequency and also studies how much of the vocabulary that is made up of the most frequent words in a particular group under study. Based on the idea that differences between languages can result in learning challenges, typological profiling can be of interest in relation to language acquisition. For instance, previous research has shown that Swedish nuclear verbs are "favored at early stages of second language acquisition" (Viberg 2006b: 127). Viberg defines this type of lexical profile as "an account of the distinctive character of its [=the language's] lexical structure in relation to other languages" (Viberg 2006b: 103).

2.3 *Contrastive interlanguage analysis (CIA and CIA²)*

Granger has shown how different corpora can be used to compare the production from different groups, e.g. learners with different first languages or learners and L1 speakers or expert users. This is a method which she has called Contrastive interlanguage analysis (CIA, Granger 1996, 2015), and in its revised form CIA². CIA and CIA² build on the ideas from contrastive analysis with the help of corpora. In CIA there is an emphasis on comparing like with like, e.g. only foreign language learners with other foreign language learners, not second language learners with foreign language learners, only texts of the same genre etc (e.g. Granger 1996: 44). To compare interlanguage with first language production, special L1 corpora were compiled in relation to the International Corpus of Learner English (Granger 1996: 44), something which we have not had a chance to do in the SweL2P project. Lately, it has also become common to compare advanced learners to "proficient" student writing "regardless of native speaker status" (Granger 2015: 12), as well as to make comparisons with large L1 reference corpora such as Corpus of Contemporary American English (COCA) (cf. Granger 2015).

Granger (2015) revised the method due to the criticism of CIA. The criticisms primarily questioned native speaker language as a target for learners. In the revised form, CIA², Granger emphasises variation both in the learner language and in the native/expert language. However, it is worth to bear in mind that Granger (2009) noted that "all the studies that compare learners of different proficiency levels are in fact based on an underlying L1 norm" (Granger 2015: 13–14). Tenfjord et al. (2006: 101–2) has also pointed out that comparison to a target language can be part of an objective method to study various phenomena in the learner language.

The data which we have gathered in SweL2P can be easily used to make comparative studies, e.g., between different proficiency levels, between learner language and course-book language, according to CIA². Furthermore, links to Korp (see Section 6.2) fascilitate comparison of the SweL2P data to corpora from different text types and varieties.

2.4 *Learning Swedish adjectives*

One aspect of language that clearly develops from early on in second language acquisition is the use of adjectives. Swedish adjectives are also both lexically and grammatically complex which is another reason why they can be interesting to observe over the different proficiency levels. Their numbers increase, they change semantically (cf. e.g. Axelsson 1994), as well as

Table 1: The most frequent adjectives in Löhndorf's (2021) study of L1 Swedish.

Grade 3	Grade 5	Grade 9	Grade 11/12
arg 'angry'	*bra* 'good'	*bra* 'good'	*stor* 'big'
rädd 'afraid'	*stor* 'big'	*rädd* 'afraid'	*bra* 'good'
glad 'happy'	*liten* 'small/little'	*hel* 'whole'	*många* 'many'
osäker 'insecure'	*rolig* 'amusing'	*stor* 'big'	*olika* 'different'
stor 'big'	*kul* 'fun'	*trött* 'tired'	*annan* 'other'
liten 'small/little'	*viktig* 'important'	*liten* 'small/little'	*ny* 'new'
jätterädd 'very afraid'	*fin* 'nice'	*ny* 'new'	*kränkt* 'insulted'
vanlig 'common/usual'	*gammal* 'old'	*mörk* 'dark'	*hel* 'whole'
nästa 'next'	*ny* 'new'	*fler* 'more'	*olika* 'different'
olika 'different'	*andra* 'other'	*många* 'many'	*fler* 'more'

morphologically. Morphologically, adjectives are interesting to study both in relation to word formation and to inflection.

The fact that Swedish adjectives are inflected both inside the noun phrase (Sw. *attribut*, for modifiers inside the noun phrase) and as predicatives (Sw. *predikativ*) on the clausal level, can prove challenging to learners. In addition, some adjectives are more common in a premodifying position inside the noun phrase whereas others are more common as predicatives.

In Allén's *Frequency Dictionary of Present-Day Swedish* (Sw. Nusvensk frekvensordbok) (Allén 1970, 1971) adjectives make up 6% of all tokens and 10% of all lemmas. Adjectives have also been shown to be much less common than nouns in the vocabulary of young L1 speakers (Löhndorf 2021: 79). Löhndorf also shows that children learning Swedish as a first language first tend to use adjectives to modify concrete nouns, only later increasing their use with more abstract meanings. She lists the ten most common adjectives in her materials from grades 3, 5, 9, 11/12 (see Table 1) which can be of interest for comparisons with L2 Swedish.

Young first language (L1) speakers have been shown to use adjectives mainly as predicatives at first (Löhndorf 2021: 82). Later they start to use adjectives more in premodifying position in noun phrases, and extended noun phrases are used more by proficient L1 speakers. In the L1 writing of upper-secondary students with good marks, adjectival pre-modifiers (Sw. attribut) make up a substantial part of all adjectives (cf. Hultman & Westman 1977). The amount of simple vs extended noun phrases have also been shown to correlate to whether texts represent typical spoken language or written language (cf. Collberg 2021) and to the information density in the text, lower in speech, higher in written texts especially in more formal genres.

Table 2: The most frequent adjectives in Axelsson's 1994 study of spoken L2 Swedish.

Adjectives	Proportion of all adjectives (%)
bra 'good'	10
svår 'difficult'	7
stor 'large'	4
bättre 'better'	3
svensk 'Swedish'	3
liten 'small'	3
lång 'long'	2
olika 'different'	3
inte bra 'not good'	2
ny 'new'	1
lätt 'easy'	2
halv 'half'	2
gammal 'old'	2
	45

Furthermore, according to Löhndorf (2021: 88), Ravid & Levie (2010) have shown that girls have higher adjective density, at least in Hebrew.

Axelsson (1994) investigated the order of acquisition of adjectives in Swedish as a second language in relation to semantic fields and the frequency of occurrence, but she also studied adjectival agreement. Her definition of adjectives relies primarily on "semantic meaning and syntactic function" (Axelsson 1994: 112) unlike the Swedish Academy Grammar (SAG, Teleman et al. 1999) which tends to put morphology first in word class classification. But like SAG, Axelsson (1994: 113) classifies adjectives used as adverbs (or rather adverbials) as adjectives. In addition, participles which are used as adjectives and certain pronouns are also classed as adjectives to be in line with Allén (1970, 1971), e.g. *olika* 'different', *själv* 'self'.

For her study, Axelsson collected learner data in the form of interviews, rather than written data as in SweL2P. The most common adjectives in her spoken data are listed in Table 2 and any adjectives that were less common were used by less than 20 individual learners (Axelsson 1994: 117). She emphasises that the adjectives used were affected by the topics discussed (Axelsson 1994: 118), but we can see that there is an interesting overlap with the data from L1 speakers in grade 3, 5, 9 and 11/12 in Löhndorf's study presented in Table 1.

If we compare these adjectives to the ones which Goddard & Karlsson (2008) have suggested should be included in a Swedish version of the Natural

semantic metalanguage (NSM) semantic primes, based on the work on prime words done by Anna Wierzbicka (e.g. Wierzbicka 1996) and Charles Goddard (e.g. Goddard 2012) only three are the same. The semantic primes include the adjectives *bra* 'good', *dålig* 'bad', *stor* 'big', *liten* 'small', *nära* 'near', *avlägsen* 'far'. In her master dissertation, Arle (2018) found that the concrete adjectives *dålig* 'bad', *stor* 'big', *liten* 'little' were more common in easy language versions of the novels she studied, than in the original texts, hence showing some similarity between easy language novels and learner language.

3 *The Swedish L2 profile*

The Swedish Lexical Profile is part of a bigger Swedish L2 Profile (SweL2P, Volodina, Alfter & Tiedemann 2024), which is described in an online hands-on manual by Volodina and Lindström Tiedemann.[2] The SweL2P was created as part of the research project *Development of lexical and grammatical competences in immigrant Swedish* and released in 2023. Below, we introduce the parent resource, SweL2P, and its user interface briefly (for more information, please see e.g. Volodina, Alfter & Tiedemann 2024) and then we focus on the lexical profile with all of its subparts in Section 4.[3]

3.1 *Swedish L2 profile – An overview*

The Swedish Second Language Profile (SweL2P) features the following profiles:

- a lexical profile, organised into subprofiles by words, multiword expressions, adjectival declensions and adjectival and adverbial structure;
- a grammatical profile, including noun patterns and verb patterns (Lindström Tiedemann et al., in preparation);
- a morphological profile, organised into word families and morpheme families (Volodina et al. 2021, Volodina, Ali Mohammed et al. 2022, Volodina, Mohammed et al. 2024)[4] (cf. Sköldberg et al. 2019).

SweL2P is empirical in nature since it is based on data from two corpora: COCTAILL – a corpus consisting of course books used for teaching Swedish

2 spraakbanken.github.io/L2_profiles/SwedishL2profiles, last updated in April 2023.

3 This section is based on Volodina, Alfter & Tiedemann (2024), which has been invited in an extended form as a chapter to this handbook

4 Additional information in the form of guidelines for the morphological analysis are available https://docs.google.com/document/d/1G5PEfeDEKg4dAZaupj6FmUUWBGiegiqagzXgTA3cDSY/edit?usp=sharing

to L2 students (Volodina, Pilán, Eide et al. 2014); and SweLL-pilot – a corpus of learner essays written by L2 learners of Swedish (Volodina, Pilán, Enström et al. 2016a, Volodina 2024). Texts in both corpora have been annotated with CEFR (Council of Europe 2020) levels by experts, starting from A1 (beginner) to C1 (advanced). There are also a few essays from the C2 level but no course books. COCTAILL has been used to get an approximation of the vocabulary and grammar L2 learners meet when reading, and therefore what they are expected to understand *receptively*. SweLL-pilot has been used to get an approximation of the vocabulary and grammar L2 learners are able to produce actively when writing, and therefore represents learners' *productive* writing abilities.

Within a usage-based approach one normally does not believe in clear boundaries between lexis and grammar and even though it may look as though we have clearly divided the SweL2P into lexis, grammar and morphology, this is not quite the case. The lexical profile naturally overlaps with the morphological profile which focuses primarily on word formation (e.g. derivation, compounding), whereas inflectional aspects are only rarely included since we have based the morphological profile on the lemma and base form. The only reason there are some inflected forms is that multi-word expressions (MWEs) may contain lexicalised inflected forms (e.g. *på allas läppar* 'on everyone's lips') and because a lemma-final morpheme which changes in other inflected forms in the morphological paradigm has been seen as an inflectional morpheme, (e.g. *flicka, flickor* 'girl, girls', where the *-a* in the singular form has been annotated as an inflectional ending).[5]

To further illustrate the fuzzy borders between lexis and grammar we also see that there are grammatical aspects under the lexical profile since there we can search by part of speech (POS), gender, nominal declension, and verbal conjugation.

3.2 *Graphical user interface*

The graphical user interface for browsing the SweL2P[6] has some features that are shared by all of its modules and subprofiles, such as a filter for CEFR levels, a possibility to enter a search item/word (except in the grammatical profile) and an option to see frequencies and samples from receptive and productive data for one's search. Some other features are specific for a specific (sub)profile.

5 For further information about the morphological annotation principles please see https://docs.google.com/document/d/1G5PEfeDEKg4dAZaupj6FmUUWBGiegiqagzXgTA3cDSY/edit?usp=sharing.

6 https://spraakbanken.gu.se/larkalabb/svlp

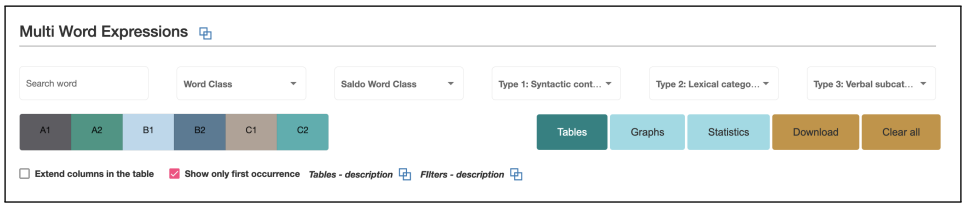


Figure 1: Filters for Multiword Expressions (Lexical profile, SweL2P)

CEFR level	Word	Lemgram	Sense	Word Class	Saldo Word Class	Receptive	Productive
A1	jeans	jeans..nn.1	jeans..1	Noun (NN)	Noun (nn)	0.77 (2)	0.00 (0)
A1	kläder	kläder..nn.1	kläder..1	Noun (NN)	Noun (nn)	3.09 (8)	7.32 (3)
A1	shorts	shorts..nn.1	shorts..1	Noun (NN)	Noun (nn)	1.16 (3)	0.00 (0)
A2	makaroner	makaroner..nn.1	makaroner..1	Noun (NN)	Noun (nn)	0.44 (3)	0.00 (0)

Figure 2: Table view for Sen*Lex (Lexical profile, SweL2P), filtered for 'always plural' nouns

Filters appear at the top of the page, providing a set of filters for each subprofile, e.g. MWE *Types 1-3* in Figure 1. The resource can be explored using several views: Table, Graphical and Statistical.

The Table view (Figure 2) lists all items which fit the filter(s) chosen. Columns contain descriptive information, among others, a clickable word (e.g. *jeans* in Figure 2) that opens a link to an entry with this item at `https://svenska.se/`, as well as a manual morphological analysis from the project and an earlier analysis from the Swedish Academy Glossary (SAOL, *Svenska Akademiens ordlista*) and the Contemporary Dictionary of the Swedish Academy (SO, *Svensk ordbok utgiven av Svenska Akademien*). In addition, this view also presents receptive and productive frequencies (both relative and absolute) based on the corpora mentioned above. If you click on the frequencies it opens a corpus search tool, Korp (Borin et al. 2025), containing the actual occurrences in the (Swedish) corpora we use. The receptive corpus (COCTAILL) is open to all users whereas the productive corpus (SweLL pilot) is restricted for ethical reasons.

The range of the columns depends on the profile – notably, for the grammatical profile, "pattern" is listed instead of "word" and here relative frequencies are currently given in terms of occurrences per 100 sentences rather than in relation to tokens.[7]

7 For more information e.g. about the choice of relative frequency, please see Lindström Tiedemann et al. (in preparation).

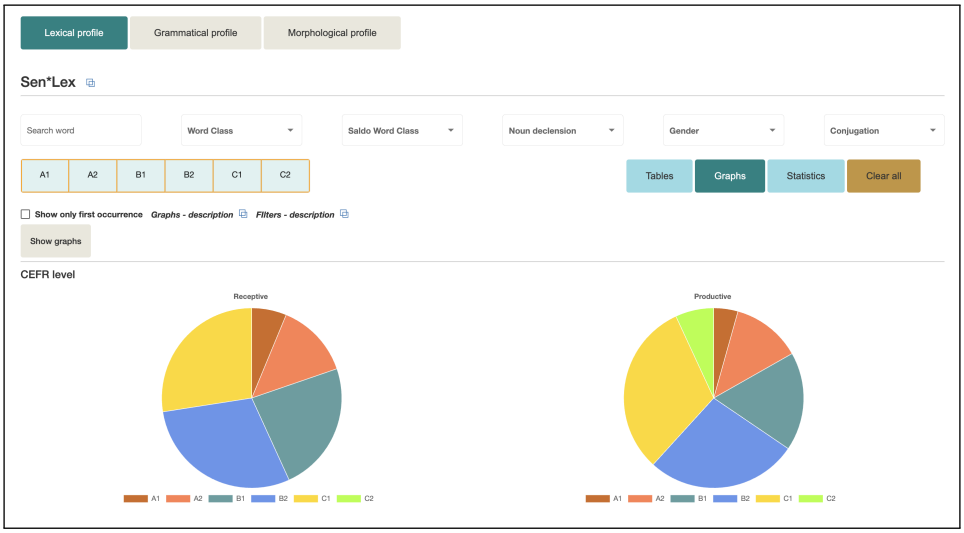


Figure 3: Graph view over Sen*Lex across CEFR levels (Lexical profile, SweL2P)

Graphical and Statistical views summarise the statistics and distribution of various features for the current selection in the two sources, receptive and productive. For example, Figure 3 shows a graph view of Sen*Lex (i.e. the whole word list in the lexical profile, see further Section 4.1) across CEFR levels, including all occurrences. The pie charts show you the proportion of lemmas which are represented at each CEFR-level.

In Figure 4, instead of graphs and distributions, we see counts in terms of types (cf. lemmas), tokens and type-token ratios (TTRs) per filter category so that we can study a statistical breakdown of each selection contrasting receptive and productive competences. In Figure 4 the number of types and tokens in Sen*Lex, in the lexical profile, are listed per CEFR-level (A1–B2 is visible in the figure) and across word classes (showing adjectives and adverbs in the figure).

The entire dataset or filtered data selection can be downloaded from the website as an excel file or as a csv file by clicking on the button to the right of the Statistics tab, "Download Statistics".

4 *The Swedish lexical L2 profile – resource description*

The SweL2P resource allows the user to choose different ways to approach the lexis. Under the heading Lexical profile, the following subprofiles can be accessed: adjectival declension, adjectival and adverbial structure, multi word expressions and Sen*Lex (cf. Figure 5).

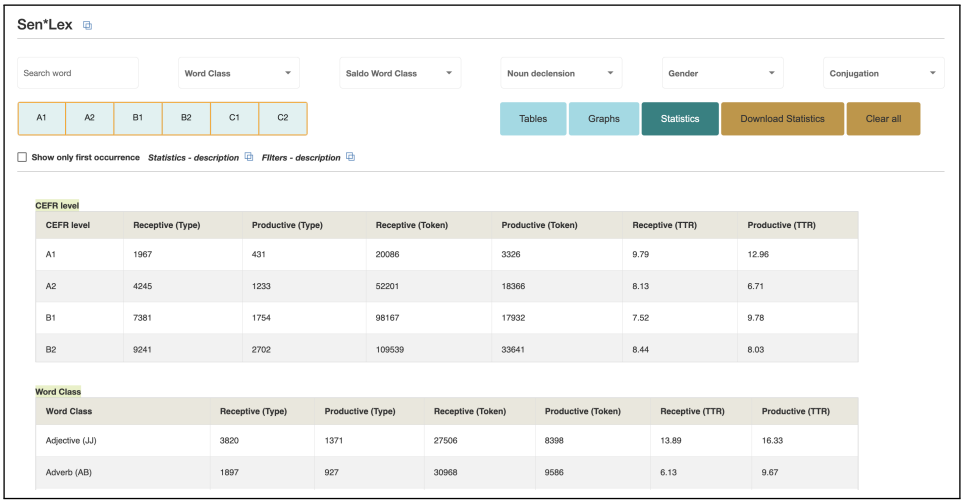


Figure 4: Statistical view showing entries per CEFR level A1–B2 and entries for two word classes (Lexical Profile, SweL2P)

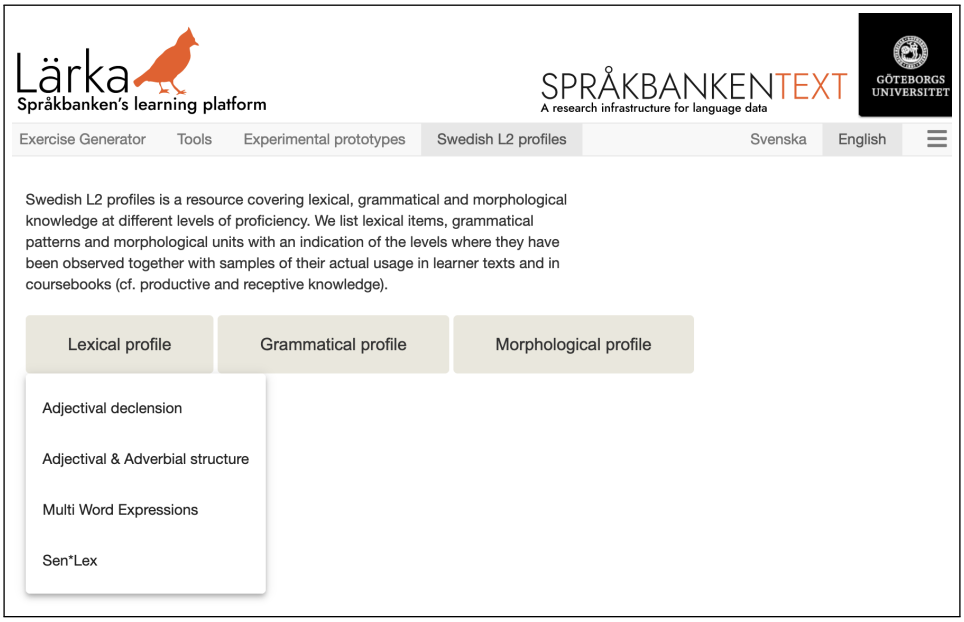


Figure 5: Overview of the possible selections under the Lexical profile in the SweL2P graphical user interface.

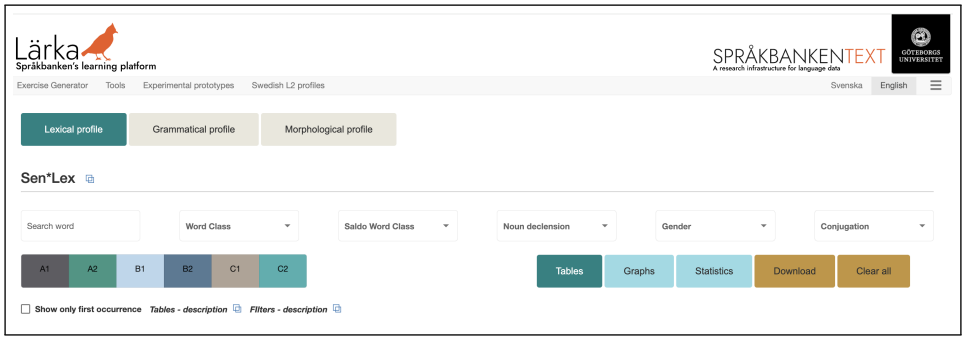


Figure 6: Buttons on the right for selection of the three different viewing options: Tables, Graphs or Statistics
(SweL2P)

The whole lexical profile is, in fact, based on Sen*Lex which is a sense-based word list which we introduce further in Section 4.1. However, rather than including everything as filters in the word list we provide some separate subprofiles for ease of access.

For each of the subprofiles of the Swedish Lexical L2 profile the user can choose various settings which will filter the results and provide different viewing alternatives. The views are presented as buttons on the right-hand side (see Figure 6) and the default setting is the table view, the alternatives being graphs and statistics. On this side the user also has the possibility of choosing to download the results, or to clear all the previous settings in that part of the profile.

Each part of the profile also offers a choice of showing only first occurrences. This is checked by default so that you only see instances according to their first level of occurrence. However, since for vocabulary this is currently based only on the registered first occurrence in course books it is often better to uncheck this.

4.1 *Sen*Lex – a CEFR-graded sense-based vocabulary list*

Traditional word frequency lists – and especially corpus-derived word lists – are typically based on form, grouping all inflected forms of a word under the same lemma, and making no distinction between different senses of a word. From a second language learning perspective, this is problematic for at least two different reasons: first, not all word forms are learned at the same proficiency level (Capel 2010, 2012, 2015). For example, learners typically first learn simple verb forms such as the Swedish present tense, before learning periphrastic forms such as the Swedish perfect or conditional

forms (Philipsson 2007, Yamaguchi et al. 2022). Second, polysemous words are typically subsumed under the same lemma – and by extension they share the same frequency in traditional wordlists. However, not all meanings of polysemous words are learned at the same proficiency level (Alfter, Cardon et al. 2022).

In order to address the polysemy problem, we have compiled new sense-based versions of SVALex (François et al. 2016) and SweLLex (Volodina, Pilán, Llozhi et al. 2016). The original corpora used to compile these lists were re-tagged with Sparv (Borin et al. 2016, Hammarstedt et al. 2022), including automatic word sense disambiguation. This allows us to separate different senses of words and assign them separate frequency values (Alfter 2021: 31–32). We call these sense-based lists SenSVALex and SenSweLLex, or Sen*Lex to refer to both lists simultaneously. In the online SweL2P resource, the two are available as a joint list and hence called Sen*Lex in the graphical user interface.

To prepare the data, we combined automatic processing (Borin et al. 2016, Nieto Piña 2019) and manual annotation (Volodina et al. 2021, Lindström Tiedemann et al. 2024a, Lindström Tiedemann et al., in preparation).

The manual annotation of Sen*Lex included various categories such as morphological analysis (Volodina et al. 2021[8], cf. Sköldberg et al. 2019), multi-word classification (Lindström Tiedemann et al. 2024a), and adjective and adverb type using the dedicated lexicographic annotation interface Legato (Alfter et al. 2019). In order to alleviate the human effort, the data was also linked to Lexin (Hult et al. 2010), Saldo-Morphology (Borin et al. 2013), and Swesaurus (Borin & Forsberg 2014) to (semi-)automatically enrich the data. Out of 15 categories, 8 were automatically enriched. In a pilot study where annotators checked 100 randomly sampled items, it was found that the automatic annotations for nominal declension type, nominal gender, and verb conjugation type were of sufficiently high quality that these categories were not manually checked in Legato. However, at a later stage further checks were carried out of all nouns and verbs in relation to these categories and some corrections were made before merging the information with Sen*Lex in the SweL2P. Some of the remaining automatically enriched categories were manually checked through Legato (adjective/adverb structure, adjectival declension) and others are still pending further manual annotation (synonyms, topics/domain, transitivity). Potential additional categories for manual annotation were also discussed but have not yet been included.

8 Additional information in the form of guidelines for the morphological analysis are available as https://docs.google.com/document/d/1G5PEfeDEKg4dAZaupj6FmUUWBGiegiqagzXgTA3cDSY/edit?usp=sharing

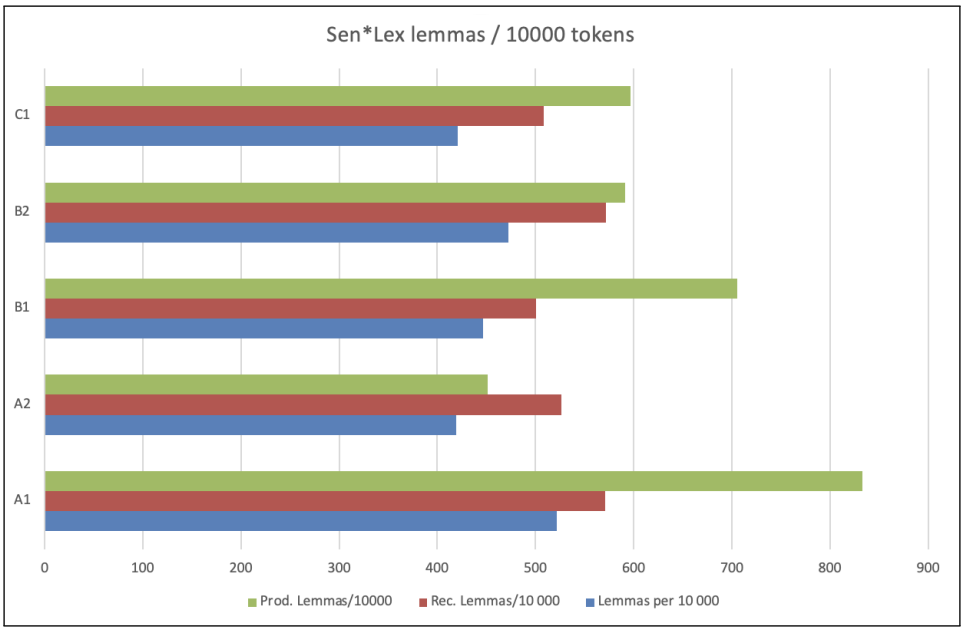


Figure 7: Comparison of the number of lemmas in relation to the total number of tokens in Sen*Lex presented as a total as well as separately for receptive and productive data.

4.2 *General overviews from the lexical profile in SweL2P*

Using the lexical SweL2P the user can choose to download all items (or first occurrences only) and use this data to see how many different lemmas there are per level and also how frequent they are in receptive (L2 Swedish course books based on the COCTAILL corpus) and productive (L2 Swedish learner essays based on the SweLL pilot corpus) data. As usual for such a comparison it is important to do this as relative frequencies, i.e. in relation to the data size of the respective data sets, if we want to compare the two datasets or different CEFR-levels. This is important since the data sizes at the different levels and in the two corpora as a whole are different.

Figure 7 shows that at the group level the number of lemmas in relation to the token frequency (cf. type/token ratio) is quite stable in the material as a whole, between 400–500+/ 10000 tokens at each CEFR level. Similarly, in the receptive data there are no big differences really between the different levels, the frequency is between 500–600 / 10 000 tokens. But as one would expect, the productive data shows clear differences between the levels. Interestingly, and at first somewhat surprisingly, there is a much higher number of lemmas per 10000 tokens at A1 (c. 800 / 10000 tokens). This means more lemmas

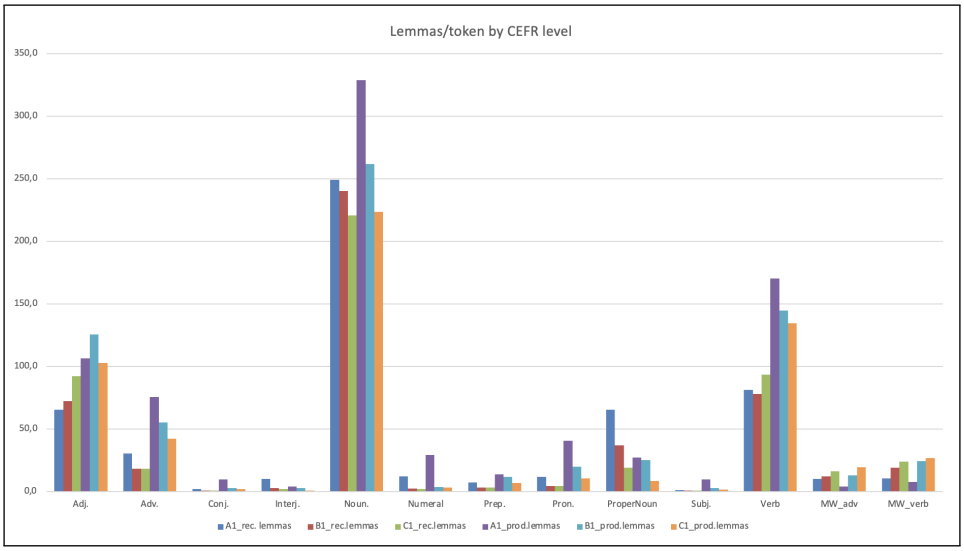


Figure 8: Lemmas per POS in relation to token frequency per CEFR level based on SweL2P

that are rarely used. The number of lemmas in relation to tokens is also quite high at B1-level (c. 700 / 10000 tokens). At A2-level it is only c. 450 and at B2 and C1 level nearly 600 and hence very similar to the receptive data. It could be that this is a result of the number of tasks which have been used to collect the data for the different levels, or it could be a result of words that are misused or misspelt, but this needs further analysis.

If we inspect the lemmas according to Saldo word class (SaldoPOS) (cf. Borin et al. 2013), we see that nouns are the most common word class, followed by verbs and adjectives (Figure 8). Nouns become less common in both receptive and productive texts at higher CEFR levels. Verbs remain fairly stable in receptive data but become less common at higher proficiency levels in productive data. Interestingly, we can also see that according to the relative frequency, learners use more verb lemmas, i.e. they vary their verb usage more at the group level, than the verb proportion used in the receptive texts. Adjectives, unlike nouns and verbs, show an increase in lemma proportion and hence vary more at higher CEFR levels in the receptive data (at the group level, i.e. based on all course books). However, in the productive data there is first an increase followed by a drop and A1 and C1 levels are found to be almost equal. However, notably, the proportion of adjectives is higher in productive texts than in receptive data which is something that should be explored further. It is quite possible that this is related to the topics of the texts.

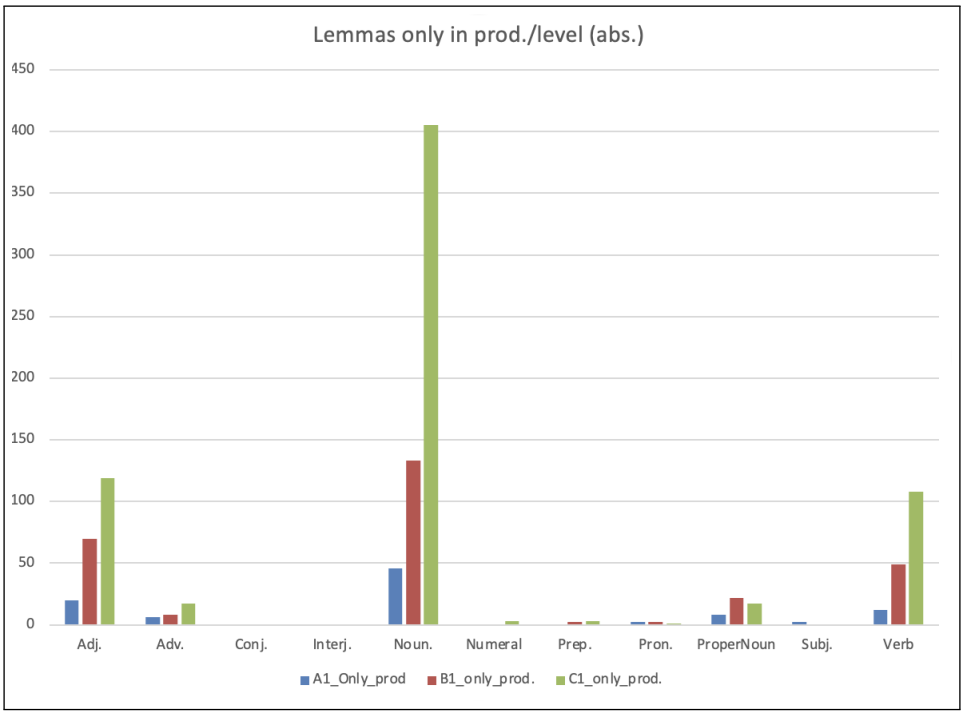


Figure 9: Lemmas exclusive to the productive data at that level based on SweL2P

Nouns, verbs, adjectives and adverbs all show higher lemma proportions in the learner production than in the receptive data at the levels A1, B1 and C1 (Figure 8), except nouns at C1-level which are used as much in receptive data as in productive data. A closer look at the word lists from the SweL2P show that this is clearly also shown in the number of lemmas which only occur in the productive data set at a certain level (Figure 9). This poses the question of how these lemmas were learnt. Were they learnt based on their frequency in course books at the level before this or from other texts? One way of exploring this issue further is by looking at whether vocabulary items appear to be core items or more peripheral at a certain level, in the sense of whether it is something that is common to many course books and/or learners (core) or rather something that is only covered by some and might not be necessary for everyone to know at that level (peripheral). This is something which we have explored in Volodina, Alfter & Tiedemann (2022).

5 *Research and teaching use scenarios*

The SweL2P resource is currently primarily useful for researchers and teachers of the Swedish language. Researchers can use the resource either:

1. to identify new research questions in a bottom-up manner by comparing receptive (course books) and productive (learner essays) occurrences across CEFR levels, or
2. to search for data based on a hypothesis, e.g. based on SLA theories or previous research on Swedish as a second language.

Teachers can similarly use the resource in at least two different ways:

1. to identify potentially challenging constructions for learners at a certain proficiency level, for instance by looking for structures which are rare in the course books and comparing this to what they have noticed that the learners find challenging. Based on this teachers have more information to use when they decide how to treat these structures in their teaching,
2. to plan data driven learning exercises where the teacher uses the resource to find concordance examples in corpora. Since the teacher can do this by CEFR-level there should be potential to find appropriate examples from course books and learner texts that could be adapted for classroom exercises (bearing in mind copyright as well as citation restrictions for learner essays), or the teacher can go one step further and look for similar examples in other corpora (cf. Section 6.2).

We now present a case study of how the lexical profile can be used in relation to linguistic research on adjectives (Section 6), followed by a case study in relation to language teaching where we look at adjectives in relation to their declension and adjectival MWEs (Section 7).

6 *Case study 1: linguistic research with the lexical SweL2P including hands-on guidelines*

Previous research (see Section 2.4) has shown that adjectives are used more at higher proficiency levels and that their semantic fields change (Axelsson 1994). We showed above (Section 4.2) that adjectives are the third most common POS in our data.

In this case study we look at the frequency of adjectives across CEFR levels and compare their receptive and productive frequencies based on

course-book data and learner essays, respectively. We start by looking at adjectives through Sen*Lex (Section 6.1) and compare the most common adjectives to previous studies. After this we explore how to access the actual occurrences in the corpora we have used, and how we can compare the results to other Swedish corpora (Section 6.2). In this way, we show how we can make a comparison similar to that in Axelsson (1994) but with access to actual corpus data for comparison rather than a static frequency list based on one genre, such as Allén (1970, 1971) which was based on newspapers. Furthermore, we also know that all of the data has been annotated in exactly the same way since the corpora are all annotated with the same pipeline (Sparv, Borin et al. 2016). This way, our method can potentially be used to compare learner language to different genres of first language writing and different potential conventions and/or norms (cf. Section 2.3).

In order to use the SweL2P you go to https://spraakbanken.gu.se/larka/svlp. The lexical profile is selected by clicking on the box Lexical profile and then selecting one of the alternatives from the drop-down menu: Adjectival declension, Adjectival-adverbial structure, Multi word expressions or Sen*Lex (see Figure 5). All of these are based on Sen*Lex (see Section 4.1 for more information), however, as described earlier, Sen*Lex has been augmented through manual annotation.

6.1 *Using Sen*Lex*

Sen*Lex (see Section 4.1) has a number of different filtering options which allow the user to get a bird's eye view of the lexis in course books and/or learner essays at different CEFR levels or to focus on a particular word class, gender, nominal declension, or verbal conjugation across all CEFR levels or to focus only on one or two (See Figure 10 and Table 3 for more details).

Sen*Lex can either be browsed directly online through the SweL2P resource, or the user can choose to download all items in the word list or all that have been filtered by clicking on Download on the righthand side (see Figure 10), which is usually better for researchers, in particular.

As an example of how Sen*Lex can be used we here use SweL2P to study the frequency of adjectives at different CEFR levels and compare these not only between the receptive and the productive side in the resource but also with previous research for both L2 and L1 Swedish (see Axelsson 1994, Löhndorf 2021, respectively). Zooming in further on adjectives we can compare the most frequent adjectives at each level in the receptive and productive data. We can do this either inside the resource or by downloading data from the resource as described above. Let us take a closer look at the most common adjectives in Sen*Lex for three levels (A1, B1, C1).

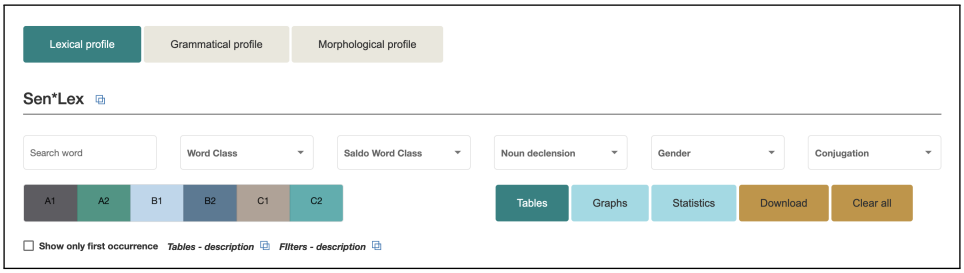


Figure 10: Filtering options in the Sen*Lex wordlist in the lexical profile, SweL2P.

Table 3: Filters and filtering options in Sen*Lex

Filters (Menus)	Nr categories	Filtering options
Word class	22	acc. to the SUC taxonomy, (Källgren et al. 2006: Table 12)
Saldo word class	37	acc. to the word classes used as part of lemgrams in Saldo, see https://spraakbanken.gu.se/resurser/saldo/taggmangd
Noun declension	6	*-or* (*blomma, blommor* 'flower, flowers'); *-ar* (*älv, älvar*, 'river, rivers'); *-er* (*dans, danser; bok, böcker* 'dance, dances; book, books'); *-r* (*sko, skor* 'shoe, shoes'); *-n* (*bo, bon* 'nest, nests'); - (*bär, bär* 'berry, berries')
Gender (nouns only)	6	always plural, no gender (*avgaser* 'exhaust'); neuter (*ett ägg, ett år* 'egg, year'); common (*en älv, en bil* 'river, car'); common or neuter (*bacon* 'bacon'); no gender (*gång på gång* 'time and again'); not applicable (*Afrika* 'Africa')
Conjugation (verbs only)	6	*-ar, -as* (*-ade(s)*); *-er, -es* (*-de(s), -te(s)*); *-r, -s* (*-dde(s)*); strong verbs (ablaut -); irregular (*dra – drog* 'tug – tugged'); no conjugation (refers to all non-verbal word classes)

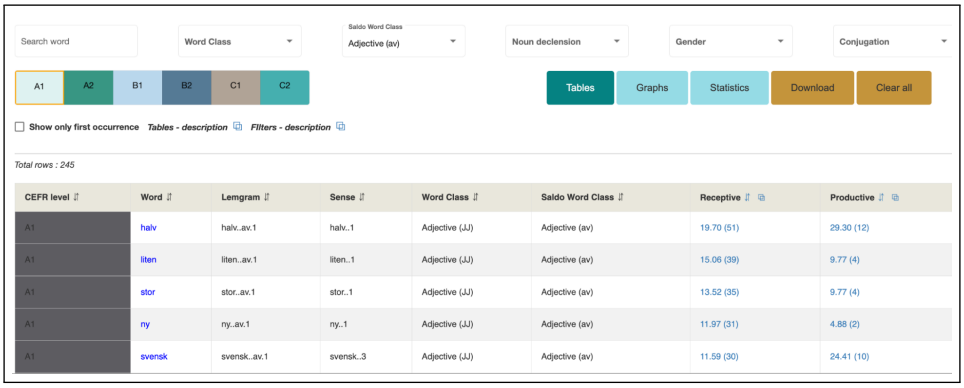


Figure 11: Most common adjectives in receptive data according to Saldo word class at A1-level, SweL2P

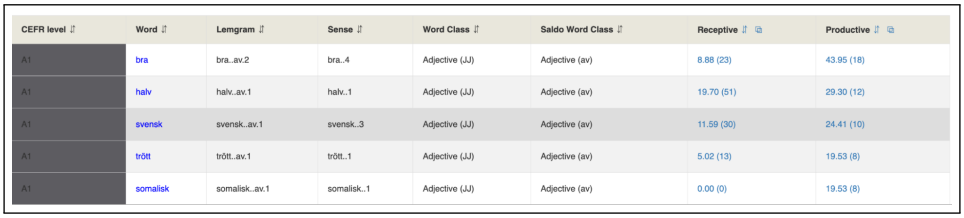


Figure 12: Most common adjectives in productive data according to Saldo word class at A1-level, SweL2P

Choosing the filters *Saldo word class: Adjective* and *CEFR-level: A1* we can then click on the arrows at the top of the columns marked receptive or productive to order the hits according to the least or the most common in either category (cf. the columns at the right in Figure 11). The most common adjectives at A1-level in the receptive data are *halv* 'half', *liten* 'little', *stor* 'big', *ny* 'new' and *svensk* 'Swedish'. In the productive data there are few overlaps among the most common adjectives, as there *bra* 'good', *halv* 'half', *svensk* 'Swedish', *trött* 'tired', *somalisk* 'Somali' are the most common adjectives (Figure 12). The differences here are clearly partly due to the limited topics which A1-level learners are asked to write about and hence do not necessarily give a very good idea of their lexical proficiency. It also makes it clear that adjectives connected with national, cultural or linguistic background are likely to be learnt quite early which also relates well to previous research (Axelsson 1994), and is in accordance with the A1 "can-do" statement (Council of Europe 2020).

For comparison, let us compare the most frequent adjectives at all CEFR levels in the productive data in SweL2P to the 13 adjectives listed as the most common by Axelsson (1994) based on her spoken learner data (Table 4).

Table 4: The most frequent adjectives in productive data in SweL2P as compared to Axelsson (1994).

	Axelsson (1994)	CEFR level in SweL2P				
		A1	A2	B1	B2	C1
1	*bra* 'good'	**bra** 'good'	**bra** 'good'	**olik** 'different'	*god* 'good'	**stor** 'large'
2	*svår* 'difficult'	**halv** 'half'	**svensk** 'Swedish'	*viktig* 'important'	*viktig* 'important'	*psykisk* 'psychological'
3	*stor* 'large'	**svensk** 'Swedish'	**olik** 'different'	**bra** 'good'	**stor** 'large'	*olik* 'different'
4	*bättre* 'better'	*somalisk* 'Somali'	*snäll* 'kind, nice'	**ny** 'new'	*andra* 'other'	*viktig* 'important'
5	*svensk* 'Swedish'	*trött* 'tired'	*glad* 'happy'	**stor** 'large'	**bra** 'good'	*andra* 'other'
6	*liten* 'small'	**gammal** 'old'	**gammal** 'old'	*god* 'good'	**olik** 'different'	*hög* 'high'
7	*lång* 'long'	**olik** 'different'	**ny** 'new'	*andra* 'other'	**ny** 'new'	**bra** 'good'
8	*olika* 'different'	**liten** 'small'	**liten** 'small'	*hel* 'whole'	*hel* 'whole'	*arbets-relaterad* 'work-related'
9	*inte bra* 'not good'	**stor** 'large'	*rolig* 'fun(ny)'	*glad* 'happy'	*stressad* 'stressed'	*stress-relaterad* 'stress-related'
10	*ny* 'new'	**svensk** 'Swedish'	**stor** 'large'	*ensam* 'alone'	*negativ* 'negative'	*stressad* 'stressed'
11	*lätt* 'easy'	*andra* 'other'	**halv** 'half'	**svår** 'difficult'	**svår** 'difficult'	**ny** 'new'
12	*halv* 'half'	*dålig* 'bad'	*fin* 'pretty'	*dålig* 'bad'	*arbetslös* 'unemployed'	*arbetslös* 'unemployed'
13	*gammal* 'old'	*engelsk* 'English'	*hel* 'whole'	**gammal** 'old'	*viss* 'certain'	*karolinsk* 'Carolingian'

Items that occur at two adjacent levels have been shaded. The items that are also on the Axelsson list have been bolded in Table 4. We can see that 7–8 of the items at A1 are also present in Axelsson even though her data is from spoken interviews. *Svensk* appears twice in the A1 frequency list but with different meanings according to the annotation.[9] Eight of the A2 items are also on Axelsson's list and to a large extent they are the same as on A1.

9 In the sense disambiguation, lemmas are linked to different senses in Saldo (Borin et al. 2013). Svensk 1 is the adjective linked to Sweden and hence nationality, while Svensk 3 is linked instead to Swedish as a language.

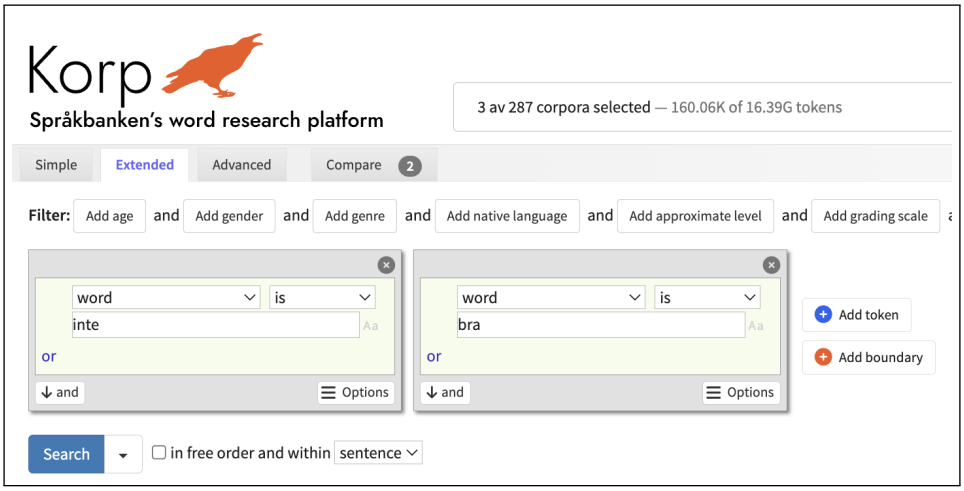


Figure 13: Querying the SweLL-pilot for occurrences of *inte bra* 'not good' in Korp

On B1 there are only 6 items which are the same as in Axelsson and this is where *svår* 'difficult' makes an entrance among the top-13 in the written learner essays. At B2 the similarity is 5 items, at C1 only 3 items. Clearly worth noting is also the fact that *dålig* 'bad' is relatively more common than in the earlier study, and as we saw in Section 2.4 *dålig* 'bad' is considered a semantic prime in Goddard & Karlsson (2008). It was also a word that was more common in easy language versions of novels than in the originals according to Arle (2018).

Instead of *dålig* 'bad' Axelsson (1994) included *inte bra* 'not good'. This has not been included in Sen*lex as an adjectival entry of its own, nor have any other combinations of the negation *inte* 'not' and an adjective. This is because it has not been seen as a MWE but rather as two independent single words. Nevertheless, in Korp we can query the SweLL pilot corpus (see Figure 13) for the collocation *inte bra*. This shows that the collocation occurs on all levels but it is much less common than *bra* is on its own, although, in comparison to *olika* 'different' and *ny* 'new', i.e. the items next to *inte bra* 'not good' on Axelsson's list, the frequency is similar to *olika* 'different' at A1, but at the other levels *olika* is much more common. *Ny* 'new' is less common at A1, similar at A2 and after that more common than *inte bra* 'not good'.

Out of the top five receptive and productive adjectives in SweL2P (see Figure 14, 15) three are the same and the two productive adjectives which are not among the most common receptive ones are clearly related to the type of tasks which forms the basis of the data: *psykisk* 'psychological' is a word that appears most in the tasks about stress which form an important

CEFR level	Word	Lemgram	Sense	Word Class	Saldo Word Class	Receptive	Productive
C1	stor	stor..av.1	stor..1	Adjective (JJ)	Adjective (av)	22.82 (359)	30.46 (160)
C1	andra	andra..av.1	andra..2	Adjective (JJ)	Adjective (av)	16.27 (256)	18.66 (98)
C1	ny	ny..av.1	ny..1	Adjective (JJ)	Adjective (av)	11.63 (183)	11.61 (61)
C1	olik	olik..av.1	olik..1	Adjective (JJ)	Adjective (av)	10.49 (165)	23.80 (125)
C1	liten	liten..av.1	liten..1	Adjective (JJ)	Adjective (av)	9.98 (157)	8.57 (45)

Figure 14: Most common adjectives in receptive data according to Saldo word class at C1-level, SweL2P

CEFR level	Word	Lemgram	Sense	Word Class	Saldo Word Class	Receptive	Productive
C1	stor	stor..av.1	stor..1	Adjective (JJ)	Adjective (av)	22.82 (359)	30.46 (160)
C1	psykisk	psykisk..av.1	psykisk..1	Adjective (JJ)	Adjective (av)	0.25 (4)	25.89 (136)
C1	olik	olik..av.1	olik..1	Adjective (JJ)	Adjective (av)	10.49 (165)	23.80 (125)
C1	viktig	viktig..av.1	viktig..1	Adjective (JJ)	Adjective (av)	8.33 (131)	21.70 (114)
C1	andra	andra..av.1	andra..2	Adjective (JJ)	Adjective (av)	16.27 (256)	18.66 (98)

Figure 15: Most common adjectives in productive data according to Saldo word class at C1-level, SweL2P

part of the data (cf. Caines & Buttery 2018). A task effect in relation to the task about stress is also visible in other words which show a high frequency such as *stressad* 'stressed', *stressrelaterad* 'stress-related' and also *arbetslös* 'unemployed' and *psykisk* 'psychological'. Similarly, the frequency of *viktig* 'important' is affected not by topic but by the fact that tasks at this proficiency level usually consist of argumentative essays.

6.2 *Browsing the data in Korp and comparing with reference corpora*

Axelsson (1994) used Allén (1970, 1971) for comparisons, a frequency word list (*Nusvensk frekvensordbok*, 'Frequency Dictionary of Present-Day Swedish') based on newspaper texts, since at the time that was the best resource for statistical comparisons. Nowadays, thanks to Språkbanken Text[10] and the Language Bank of Finland,[11] we have access to many different Swedish corpora which have been annotated in the same way and which can provide excellent statistical and qualitative data for comparison through Korp (Borin et al. 2012). This gives us better potential to empirically predict what learners can learn at the different stages of acquiring Swedish as Davidson et al. (2008: 138) have claimed that "a broad sample of many different use domains should be used" if we wish to try to predict learner usage.

10 https://spraakbanken.gu.se/
11 https://www.kielipankki.fi/

If we click on the number in either the receptive or the productive column in the SweL2P this will take us to the data in Korp, where we can see how the words have been used in the context of at least one sentence. The receptive data is open to everyone, whereas the productive data is restricted. The receptive data can only be accessed at the sentence level, whereas the productive data can also be accessed as full texts by clicking on a link under text attributes, to the right in Korp.

Let us look at the word *psykisk* 'psychological, mental' and see how it was used. In COCTAILL, *psykisk* appears in sentences such as (1a) and (1b) and it only appears in one course book.

(1) a. Jag har faktiskt också rätt – förutsatt att jag inte lider av en så allvarlig psykisk sjukdom att jag kvalificerat mig för inlåsning på en psykiatrisk klinik – att krossa mitt huvud mot marken.
'Actually, I also have the right – as long as I am not afflicted by a serious psychological disease which has qualified me to be locked up at a psychiatric clinic – to beat my head to pieces against the ground.' (Språkporten – svenska som andraspråk 1 2 3)

b. Barn behöver närvarande föräldrar – inte bara fysiskt utan också andligt och psykiskt.
'Children need parents who are present – not only physically but also spiritually and psychologically.'
(Språkporten – svenska som andraspråk 1 2 3)

All of the productive examples of *psykisk* at the C1 level are from learner essays on the topic of *Stress in current society*, e.g. (2a) and (2b).

(2) a. Det är inte bara vår fysiska hälsa som vi måste oroa sig för utan också psykiska.
'It is not only our physical health that we should worry about, but also our psychological [health].' (T100TT1)

b. Socialstyrelsen har i sin studie bevisat att stress på jobbet kan vara anledningen till psykisk sjukdom.
'The National Board of Health and Welfare have proven in their study that stress at work can be a reason for mental disease.'
(T95TT1)

Using the same query as the one that opens in Korp when we choose to click on the number in one of the columns, we can remove the CEFR-level (Figure 16) and then pick another corpus which we would like query

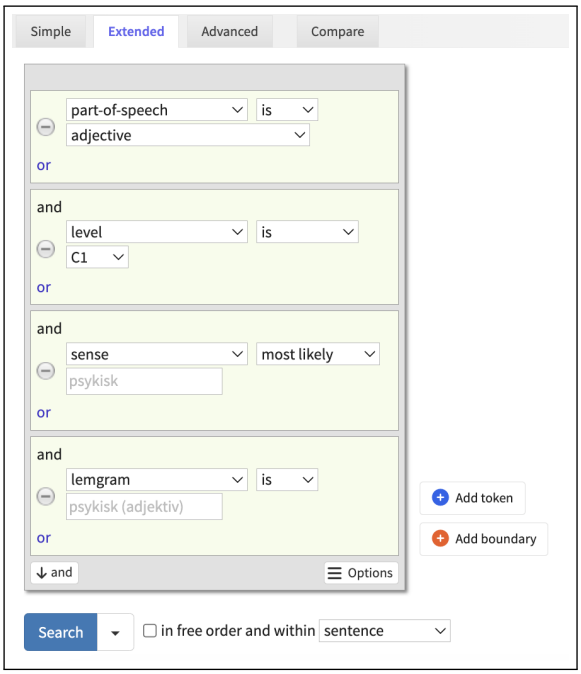


Figure 16: The adjustable query which opens in Korp.

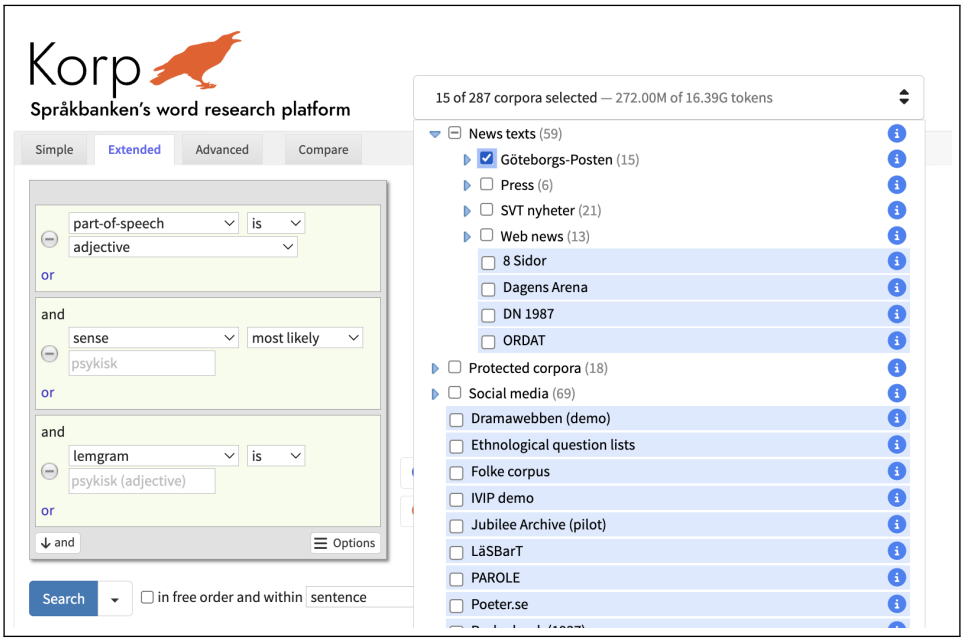


Figure 17: Choosing a new corpus to query in Korp

by using the drop-down corpus menu (Figure 17), e.g. news texts – Göteborgsposten.[12] The statistics from the search for *psykisk* as an adjective show a relative frequency 30.4 per 1 million token in *Göteborgsposten*, compared to 25.9 / 10 000 token, or 2589 / 1 million token, which is highest in student writings (C1 level). However, this word is nowhere near as common in course books, where the highest relative frequency is 0.26 / 10 000 tokens at B2 and 0.25 / 10 000 tokens at C1. Hence, learners who are using this are using it more than is common in either course books or newspapers. It seems unlikely that they have learnt the word only based on the course book, and the comparison with newspaper texts indicates possible overuse. This could possibly mean that they have copied the prompt, and it is hard to say how well they understand what the word means. To investigate this we would first of all need to have a closer look at the way learners use the word. This is something one can do if granted access to the restricted SweLL-pilot corpus.

7 *Case study 2: the lexical profile when teaching adjectives*

Teachers can of course also be interested in browsing the lexical frequencies as we showed in the previous section where we focused on research. We will now look at another aspect of the lexical profile which can interest both teachers and researchers, but this time focusing on a teacher perspective.

When teaching adjectival inflection, it can be of interest to know how often students are likely to read and also use adjectives from the different declensions, since according to usage-based theories this may affect how easily different forms can be learnt. In this section we therefore show how adjectives can be browsed by declension (Section 7.1), but also according to how they are used in comparisons (Section 7.2). Similarly, it can also be of interest to have more information in relation to the teaching of adjectival MWEs, and therefore we finish with an example of how to examine adjectival MWEs through the MWE profile (Section 7.3).

7.1 *The adjectival declension profile*

Teaching adjectives to L2 Swedish learners involves teaching them to inflect the adjectives correctly according to the gender and number of the noun, as well as whether the noun phrase is definite or indefinite. In addition, the inflection also depends on the syntactic position since the inflection is

12 Göteborgsposten, also known as GP, is a newspaper on the west coast of Sweden in the town of Gothenburg.

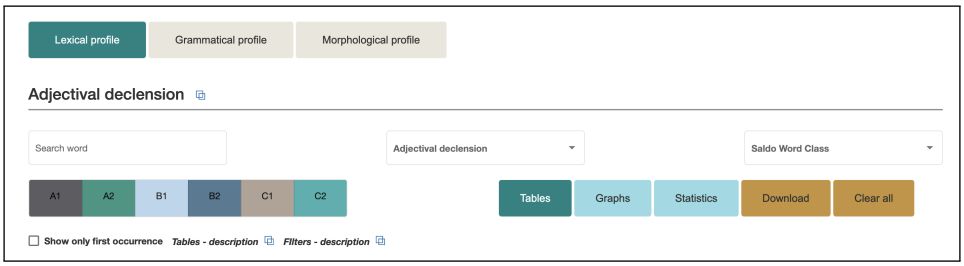


Figure 18: Adjectival declension and its filtering options, SweL2P

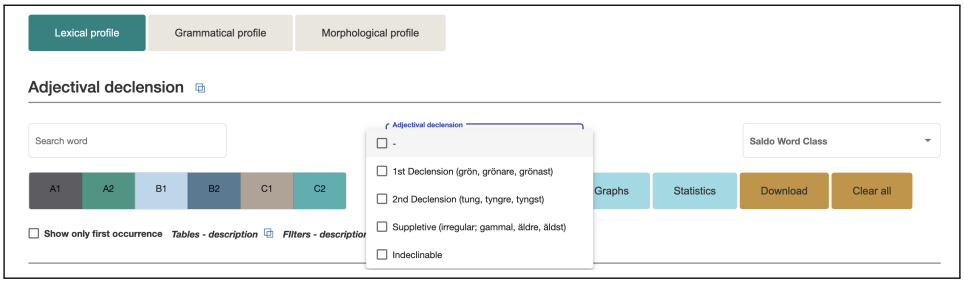


Figure 19: The adjectival declension filtering options in the subsection Adjectival declension, SweL2P

not quite the same when the adjective is part of a noun phrase (nominal modifier, Sw. *attribut*) as when it is an adjectival phrase on the clausal level (predicative, Sw. *predikativ*).

This part of the lexical profile is based on all adjectives in Sen*Lex (according to the part-of-speech tag). These adjectives have been manually annotated with information regarding declension in our annotation tool, Legato (Alfter et al. 2019).

The subprofile can be filtered according to CEFR-levels through the buttons on the left-hand side (see Figure 18), just as all other subprofiles in the lexical profile, cf. the MWE subprofile in Lindström Tiedemann et al. (2024a) and the morphological profile in Volodina, Alfter & Tiedemann (2024).

It is possible to choose only to view first occurrences or all items, where first occurrences are based on first occurrence in course books. The user can choose to search for a specific adjective, or filter based on adjectival declension (see Figure 19), or Saldo word class (see Figure 20). Saldo word class might seem counter-intuitive seeing as these are items which have been classed as adjectives. However, we have two different types of POS-annotation and they sometimes yield different results.

If, for instance, we select A1 and graphs (see Figure 21) we can see the proportion of adjectives of different declensions in both course books

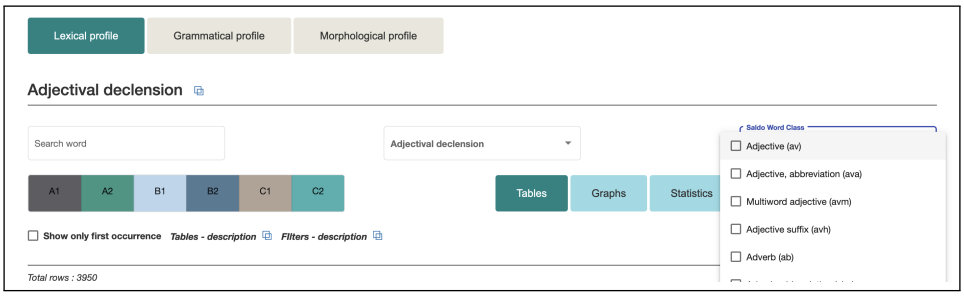


Figure 20: The Saldo word class filtering options in the subsection adjectival declension, SweL2P



Figure 21: Adjectival declension. Filter selected: A1. View selected: Graphs, showing all occurrences, SweL2P.

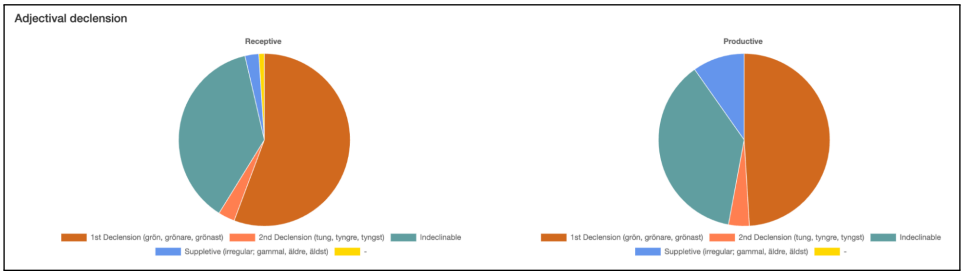


Figure 22: Adjectival declensions at A1-level, SweL2P

(receptive, left-hand side) and learner texts (productive, right-hand side) from the A1-level (see Figure 22).

This shows that the regular first declension *grön, grönare, grönast* 'green, greener, greenest' is clearly the most common in both receptive and productive texts, but quite closely followed by indeclinable adjectives (e.g. *svensk* 'Swedish', *alfabetisk* 'alphabetic'; see Figure 22).

With this much exposure to the first declension adjectives in course-book input and also the high level of usage in the productive texts, we might hypothesise that learners may be able to use the first declension quite well. However, the second most common group is nearly as common and consists

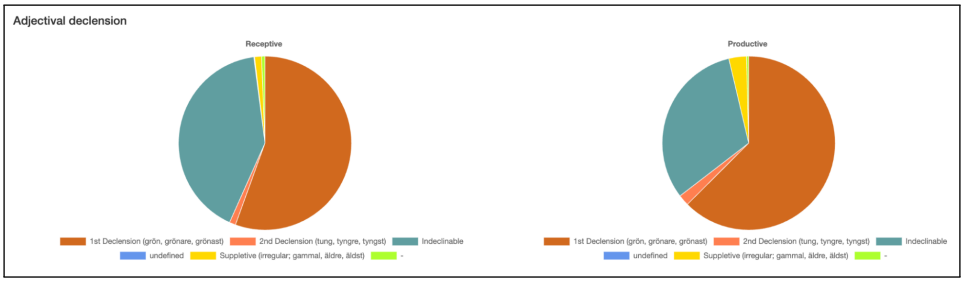


Figure 23: Adjectival declensions at B1-level, SweL2P

of indeclinable adjectives. Hence, learners meet both of these in nearly equal proportion, and that might complicate learning to distinguish the two. Furthermore, it must be noted that this only gives us information about the lemmas and their declension and not how often the learners actually read or use the comparative and superlative forms from each declension. If a teacher has seen that their students seem to have some problems with adjectival inflection, seeing these frequencies they can decide to try to provide more examples of the different declensions and additional exercises.

At B1 level (Figure 23) we see that the first declension continues to be the most common in both datasets, however, its proportion in the productive data increases. Suppletive adjectives can be seen to decrease clearly in the productive data and less clearly in the receptive data. At C1 level, finally, the proportion of first declension and indeclinable adjectives seems fairly stable in comparison to B1.

Interestingly, suppletive adjectives such as *dålig, sämre, sämst* 'bad, worse, worst' are more common in the productive texts than in the course books at A1 level. However, the suppletive adjective proportion has been reduced somewhat in the productive data at C1 level. Nevertheless, as noted for the A1 level, all levels would require more information regarding the contexts and inflectional forms in which the adjectives are used to determine how much exposure the learners get to the different patterns which the inflectional paradigms show.

Notably the adjectives which are listed as indeclinable are only indeclinable as far as comparison goes, and even in this sense they might sometimes be metaphorically compared (e.g. *Kan man bli **svenskare** än så?* 'Can one become **more Swedish** than that?'). Just like adjectives from the first and the second declension, these "indeclinable" adjectives also have to agree with the nominal correlate in terms of gender and number, and with the noun phrase for definiteness: *dålig, dåligt, dåliga* 'bad'. The first declension has very regular inflection to show agreement with the nominal correlate *grön,*

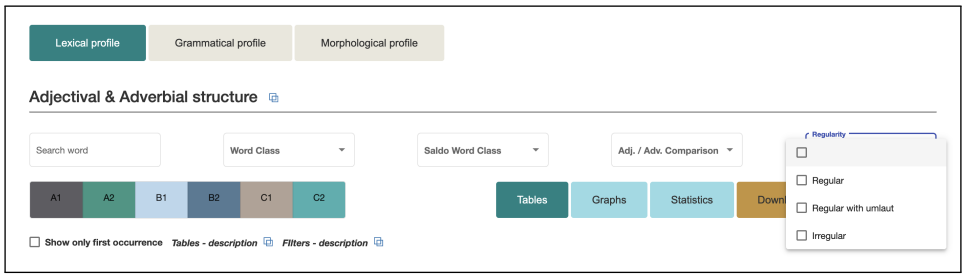


Figure 24: Filtering options under regularity in the adjectival and adverbial structure subsection, SweL2P

grönt, gröna 'green'. The same is true of the second declension: *tung, tungt, tunga* 'heavy'.

Sadly, the current view in the lexical profile does not tell us how common the different inflectional forms are and previous research has shown that learners tend to overuse non-inflected forms. To assess the difficulty of learning the difference of the first declension and the indeclinable adjectives, we need to take a closer look at the frequency and use of the different inflectional forms. We can do this by querying the corpora as shown above (Section 6.2).

Since suppletive adjectives are so common on the lower proficiency levels, it is very important that adjectival inflection and comparison is not only taught on the basis of the regular first declension, but includes suppletive adjectives from the start. It is also important to consider how often inflected suppletive verbs are used in the texts that learners read and whether noticing (Schmidt 1990) techniques might be needed to help learners learn the different ways in which adjectives can be inflected to show comparison.

7.2 *The adjectival and adverbial structure profile*

Adjectives can also be browsed through the option called Adjectival and adverbial structure (see Figure 24). Like adjectival declension, this has been based on all lemmas which have been part-of-speech tagged as adjectives, but this also includes all adverbs, and the lemmas have been manually annotated in Legato (Alfter et al. 2019) according to the type of comparison that can be used with them: morphological (*-are, -ast* '-er, -est') or periphrastic (*mer, mest* 'more, most'). In addition, manual annotations also marked whether the morphological inflection is regular. This means that this profile to some extent overlaps with the subprofile for adjectival declension (see Section 7.1), although, it provides other opportunities for filtering and includes two parts of speech.

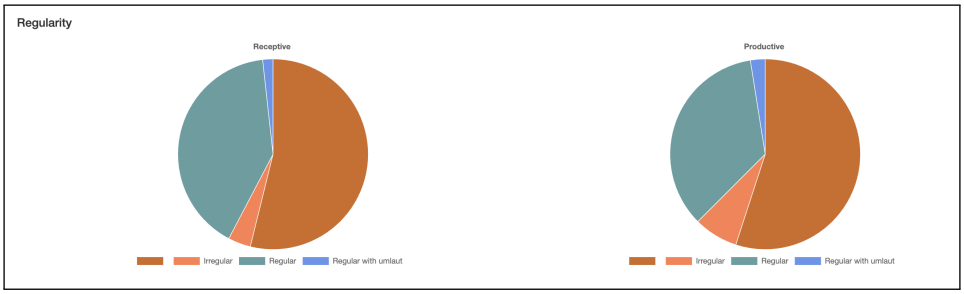


Figure 25: Adjective regularity at A1-level including whether stem change with umlaut, SweL2P

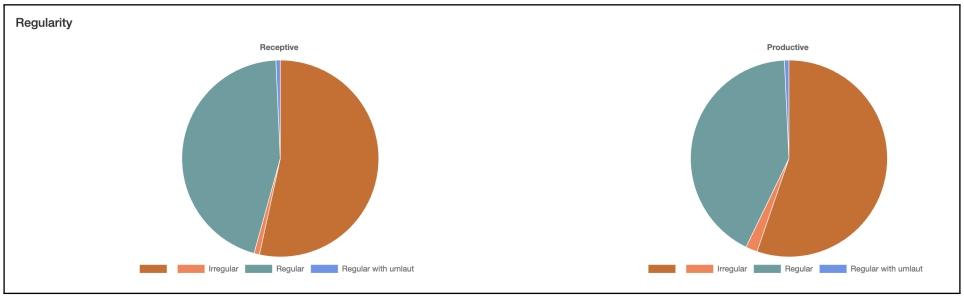


Figure 26: Adjective regularity at C1-level including whether stem change with umlaut, SweL2P

In this subprofile it is also possible to filter adjectives that have regular inflection with a stem change with umlaut such as *stor, större, störst* 'big, bigger, biggest'. These adjectives are clearly relatively rare at all levels (see Figures 25 and 26). It is important in teaching to remember how rare these types of adjectives are in comparison with others, since this is likely to make it more difficult to learn when to use the umlaut characters when writing Swedish. Initial studies on Swedish as a second language in Finland have shown that Swedish spelling connected with the umlaut letters, ä, ö, can pose challenges to learners even if their first language contains the same characters (Lindström Tiedemann et al. 2024b).

7.3 *The multiword expression subprofile and the study of adjectival MWEs*

For more advanced learners it can be of interest (see e.g. Pawley & Syder 1983, Wray 2002) to practise multiword expressions as they have been proven to be important for idiomatic language use (Paquot 2019). In the subprofile called

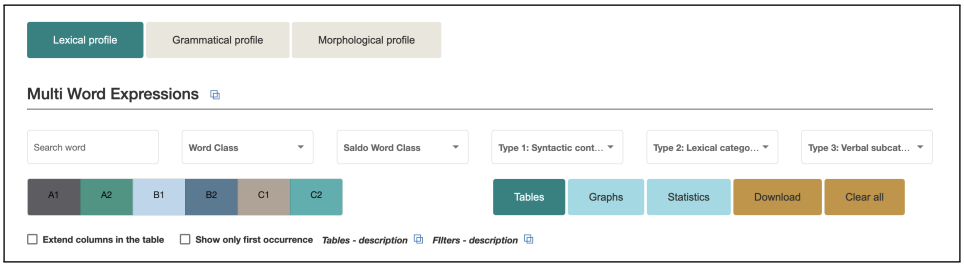


Figure 27: Filtering options for multiword expressions, SweL2P



Figure 28: Automatically annotated adjectival MWEs according to the Saldo word class filter over CEFR levels, SweL2P

Multiword expressions (MWEs), you also find adjectival information.[13] This subprofile provides a possibility to filter automatically annotated multiword expressions (MWEs) based on the manual annotation according to different types (1–3) as well as the two different POS tagsets (see Figure 27).

We can filter the MWEs by automatically annotated adjectival MWEs (see Figure 28). This shows us that the frequency of adjectival MWEs varies at the different CEFR-levels and that there are clear differences between the receptive and productive data.

The first MWE filter (Type 1) separates MWEs that can be non-contiguous, i.e. that allow insertion of other words due to word order restraints or other factors, e.g. *fatta beslut* 'make decision'), from contiguous ones, i.e. the ones that cannot be separated, e.g. *Vita Huset* 'White House'. There are no non-contiguous adjectival MWEs in the SweL2P (Figure 29).

13 The manual annotation of multiword expressions has been carefully described in Lindström Tiedemann et al. (2024a).

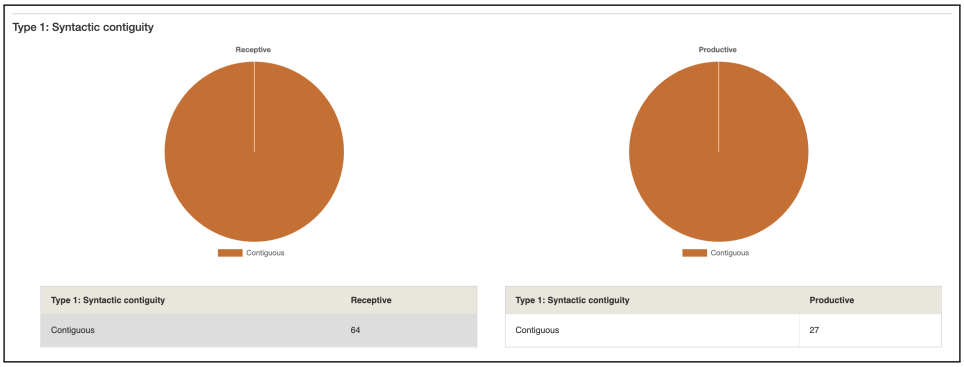


Figure 29: Contiguity of adjectival MWEs in SweL2P

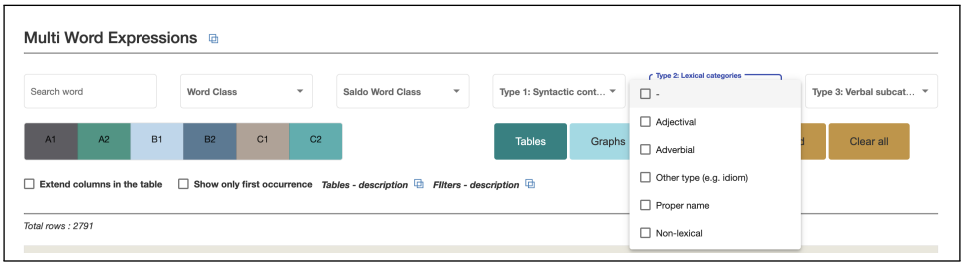
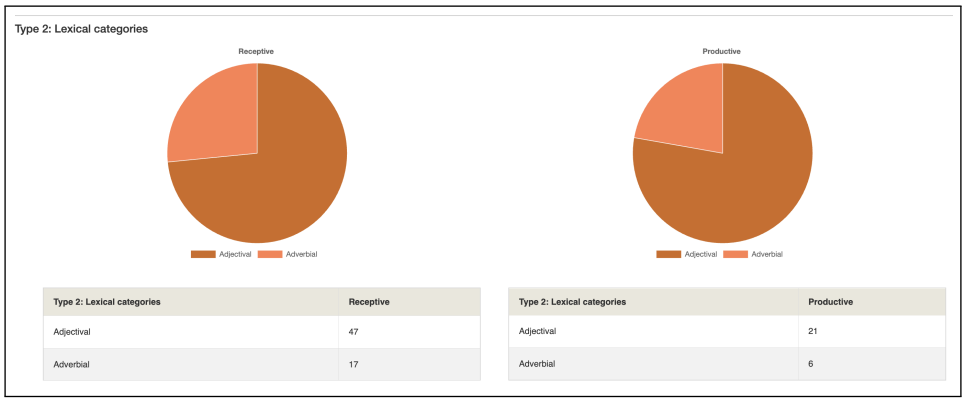


Figure 30: Lexical subcategories under Type 2 filter, SweL2P

The second filter (Type 2, Figure 30) includes subcategorisation by lexical class: adjectival, adverbal, interjections, nominal, non-lexical, other types (i.e. idioms), proper names, verbal according to manual annotation (for more information see Lindström Tiedemann et al. 2024a). Even though we have checked the reliability of the automatic annotation (Volodina, Alfter, Lindström Tiedemann et al. 2022), in many ways this is an important way to provide users with a possibility to see that different analyses may be possible in certain contexts. By comparing the Saldo word class and the manual annotation under type 2, we can also check the reliability of the automatic annotation of MWEs by Sparv. We see that there is a certain difference but this does not necessarily mean that Sparv should be seen as unreliable with regards to adjectival MWEs, but rather that the classification can depend on the guidelines for POS categorisation. Still in relation to teaching it is valuable to see that some of the adjectival MWEs are used primarily in typically adjectival functions, wheras others are used more in adverbial functions (Figure 31) and therefore categorised as adverbial MWEs in the manual annotation. This might cause challenges to learners and should be studied more closely.



Type 2: Lexical categories	Receptive
Adjectival	47
Adverbial	17

Type 2: Lexical categories	Productive
Adjectival	21
Adverbial	6

Figure 31: Adjectival MWEs according to the manual annotation of lexical categories, SweL2P

We now select (1) Saldo word class: Multiword adjective and (2) Type 2: adverbial, so that we can see which MWEs that are used adverbially. The first to appear is sadly a mistaken MWE annotation by Sparv, *i sin* 'in his/her/their' (A2, B1, C1, C2), which is not used as an MWE. Instead this is the preposition *i* 'in', and the pronoun *sin* 'his, her, their'. However, from the B2-level there are several adjectival MWEs which are in fact commonly used adverbially: e.g. *i högsta hugg* 'at the ready', *i blodet* 'in the blood', *i sken* 'bolting, running away', *i tur* 'in turn', *till hands* 'at hand'. These are often not completely transparent or compositional and they can often pose challenges even to advanced learners (cf. Alfter, Tiedemann et al. 2022, Lindström Tiedemann et al. 2024a).

The final MWE filter (Type 3) categorises verbal MWEs further into subcategories, such as support verb MWEs, particle verbs and reflexive verbs and hence will not be treated here any further.

8 *Downstream NLP tasks with the lexical profile*

The SweL2P resource is also useful for research and development within the area of Natural Language Processing (NLP). Historically, Sen*Lex and its predecessors – SVALex and SweLLex – have been extensively used for multiple downstream NLP tasks, which we shortly outline below: Development of machine learning algorithms for text and sentence classification, including CEFR-grading of learner essays, readability classification of reading comprehension texts and complexity prediction on a sentence level (e.g. Pilán, Volodina & Zesch 2016, Pilán, Volodina & Borin 2016). Linguistic fea-

tures, such as graded vocabulary lists, have been used for training predictive systems. One of the most important insights from this work is that lexical features have proven to be the most reliable predictors among a variety of other linguistic features (Pilán & Volodina 2018). These experiments have resulted in an online demo-tool for classification of texts (Volodina, Pilán & Alfter 2016). Recently, the work on essay classification has been resumed, this time relying on the Large Language Models (LLMs) where linguistic features are of less importance (Muñoz Sánchez, Alfter et al. 2024, Muñoz Sánchez, Dobnik et al. 2024).

Automatic complexity prediction of single lexical items (Alfter & Volodina 2018), where the two graded vocabulary lists (SVALex and SweLLex) have been used to train automatic systems to classify, at which level a learner can be expected to know a word (receptively or productively). Experiments on different approaches to predicting the difficulty of a lexical item (especially important for items that are not in Sen*Lex and hence have not been linked to a CEFR level through course books or learner essays) have resulted in an experimental suite of tools, CEFR-tools (Alfter 2021). Lately, the abilities of Language Language Models to predict lexical complexity have also been tested, where SVALex and similar graded lists have been used as a test set (Alfter 2024).

The same vocabulary lists for Swedish have been very important input for two methodological studies on crowdsourcing linguistic annotations from second language learners. In one study the object of annotation was multiword expressions (Alfter, Tiedemann et al. 2022), which was later analysed more qualitatively in Lindström Tiedemann et al. (2022). Another study focused on single lexical items (Volodina, Alfter & Tiedemann 2022). In both cases the task was to rank the items from easy to difficult using comparative judgment settings (best-worst scaling). A very encouraging conclusion from the two experiments was that, when annotation is solicited in a comparative setting, second language learners are as reliable in predicting the difficulty of lexical items as are experts, such as language teachers and assessors.

Another very promising direction, where the lexical profile in SweL2P with its lexicographically enriched annotation has extreme potential, is the development of teaching materials, such as vocabulary exercises for language learners within the context of Intelligent Computer-Assisted Language Learning (ICALL), similar to Volodina, Pilán, Borin et al. (2014), Volodina & Pijetlovic (2015), Alfter et al. (2018). In the future, there is a lot to be explored. For example: using the SweL2P, we can filter relevant items of interest (e.g. the most frequent adjectives at the A1 level, productively and receptively) and extract sentences containing those adjectives.

Using the Lärka[14] facilities, we can easily create stepwise vocabulary training:

1. First, learners work with gapped sentences, where the target adjectives have been removed, trying to fill in the gaps with appropriate adjectives (training productive knowledge), see the mock-example below.

```
TASK: Fill in the missing adjectives
-----------------------------------
Du är en [b.......] fotograf !
Klockan är [h.......] åtta på kvällen .
Det finns en [g.......] soffa , två fåtöljer och ett bord i
    vardagsrummet .
```

2. Second, they listen to the sentences and fill in the gaps based on what they can hear (training listening and spelling).

```
TASK: Listen to the sentence and fill in the word you hear (an
    adjective)
-----------------------------------
[play >] Du är en [.......] fotograf !
[play >] Klockan är [.......] åtta på kvällen .
[play >] Det finns en [.......] soffa , två fåtöljer och ett
    bord i vardagsrummet .
```

3. Finally, they get a list of missing adjectives ("word bank") that should be matched with the gaps (training graphical form recognition, as well as grammatical and semantic constraints).

```
TASK: Use one of the words to fill in the gaps. One word is
    ``odd'' :
halv , bra , gammal , nästa
-----------------------------------
Du är en [.......] fotograf !
Klockan är [.......] åtta på kvällen .
Det finns en [.......] soffa , två fåtöljer och ett bord i
    vardagsrummet .
```

4. Each of the exercise types can be used to log reaction times (or similar functionality) to permit use of the generated exercises for experimental research into cognitive processes and learning.

14 A language learning platform developed at Språkbanken Text, Sweden: https://spraakbanken.gu.se/larka

It is worth mentioning that we are collecting available scripts that could be reused by other researchers or students, in a github repository.[15] Among others, we provide a script for the generation of frequency lists from COCTAILL and SweLL essays (or resources with similar annotation); for interlinking such frequency lists with other lexicographic information; for generation of best-worst scaling mini-sets of four items using a redundancy-reducing algorithm. New scripts are contiuously being added, therefore, the current description might not provide full coverage.

All of the datasets that are relevant for the Swedish (Lexical) Profile are available for download from Språkbanken's resource webpage and are citable as resources:

- COCTAILL (Volodina & Pilán 2014)
- SweLL-pilot (Volodina, Pilán, Enström et al. 2016b)
- SVALex (Volodina, Pilán & François 2016)
- SweLLex (Volodina & Pilán 2016)
- Sen*Lex (Alfter et al. 2023)
- L2Lex-Adj (Lindström Tiedemann et al. 2023a)
- L2Lex-AdjAdv (Lindström Tiedemann et al. 2023b)
- Swe-MWELex (Lindström Tiedemann et al. 2023c)

9 *Final remarks*

The lexical part of the Swedish L2 profile (SweL2P) provides many opportunities for teachers and language researchers to explore L2 Swedish from different perspectives in innovative empirical ways. It can be used as a way to identify possible needs for further teaching and for new teaching materials, including functioning as a way to design DDL exercises. Furthermore, users can use SweL2P for research and to identify new research questions in a data driven manner. In addition to this, the Swedish L2 profile can be used to provide new CALL/ICALL applications which are sensitive to CEFR levels.

The lexical part of the Swedish L2 profile has been carefully designed in a way that ascertains its comparability with reference corpora available through Korp. Since the corpus queries used for the Swedish L2 profile are completely open, users can reuse them on other corpora which have been annotated in the same way and compare L2 usage and L2 course-book usage to the use in a variety of L1 genres, thus supporting e.g. CIA analyses.

15 https://github.com/spraakbanken/L2_profiles

The Swedish L2 profile is open to everyone and unlike other similar resources the development team has made sure that the user does not only have access to information about what sources the different lists are based on, but is also given the opportunity to access the empirical data. Our data is openly available, apart from the data from the learner texts which require further authentication. The fact that SweL2P facilitates comparison with other Swedish corpora means a great potential in connection with research and teaching. It can help us compare frequencies, contexts and also find examples that can be used in designing exercises for teaching. This gives the resource a more descriptive angle than previous profiles such as the English Vocabulary Profile which may appear more prescriptive and which only provides occasional examples from the empirical data. It also makes our profile easier to use as a basis for further research.

It is a fact that languages and research domains that can boast rich data collections are subject to more empirical and data-intensive research (Perc 2014, Søgaard 2022). This makes us believe that now, with the Sen*Lex vocabulary list and the lexical profile in the Swedish Second Language Profile (SweL2P) available for research and development, the field of Swedish as a second language, research on L2 Swedish as well as ICALL for Swedish and other related research fields will receive a boost.

In the future, we are expecting both short-term and long-term impact from the Swedish Second Language Profile on the fields of Swedish as a Second Language, Learner Corpus Research (nationally and internationally), and NLP- and AI-based approaches to L2 Swedish, both with respect to research and teaching.

Acknowledgments

Work on the Swedish L2 Profile has been supported by a research grant from the Swedish Riksbankens Jubileumsfond *Development of lexical and grammatical competences in immigrant Swedish*, P17-0716:1 (2018–2021). Work on the article for the first author has been supported by the University of Helsinki (Finland); the last author has been supported by Nationella språkbanken and HUMINFRA, both funded by the Swedish Research Council (2018–2024, contract 2017-00626; 2022–2024, contract 2021-00176) and their participating partner institutions.

References

Alfter, David. 2021. *Exploring natural language processing for single-word and multi-word lexical complexity from a second language learner perspective*. Data Lingvistica 31, University of Gothenburg.

Alfter, David. 2024. Out-of-the-box graded vocabulary lists with generative language models: Fact or fiction? In *Proceedings of the Workshop on Natural Language Processing for Computer-Assisted Language Learning (NLP4CALL 2024)*.

Alfter, David, Lars Borin, Ildikó Pilán, Therese Lindström Tiedemann & Elena Volodina. 2018. From language learning platform to infrastructure for research on language learning. In *CLARIN Annual Conference 2018*, 53.

Alfter, David, Rémi Cardon & Thomas François. 2022. A dictionary-based study of word sense difficulty. In Rodrigo Wilkens, David Alfter, Rémi Cardon & Núria Gala (eds.), *Proceedings of the 2nd Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI) within the 13th Language Resources and Evaluation Conference*, 17–24. Marseille, France: European Language Resources Association. https://aclanthology.org/2022.readi-1.3.

Alfter, David, Therese Lindström Tiedemann & Elena Volodina. 2019. LEGATO: A flexible lexicographic annotation tool. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, 382–388.

Alfter, David, Therese Lindström Tiedemann & Elena Volodina. 2023. *Sen*Lex (updated: 2023-04-20)*. [*Data set*]. DOI: 10.23695/b417-1j26.

Alfter, David, Therese Lindström Tiedemann & Elena Volodina. 2022. Crowdsourcing relative rankings of multi-word expressions: Experts versus non-experts. *Northern European Journal of Language Technology* 7(1).

Alfter, David & Elena Volodina. 2018. Towards single word lexical complexity prediction. In *Proceedings of the thirteenth workshop on innovative use of NLP for Building Educational Applications*, 79–88.

Allén, Sture. 1970. *Nusvensk frekvensordbok baserad på tidningstext* [Frequency dictionary of present-day Swedish]. Vol. 1: *Graford. Homografkomponenter.* [Graphic words. Homograph components]. Stockholm: Almqvist & Wiksell.

Allén, Sture. 1971. *Nusvensk frekvensordbok baserad på tidningstext* [Frequency dictionary of present-day Swedish]. Vol. 2: *Lemman* [Lemmas]. Stockholm: Almqvist & Wiksell.

Arle, Solveig. 2018. *Hur används primord för att skriva lättläst?: En undersökning av tre återberättade romaner*. University of Helsinki. (MA dissertation).

Axelsson, Monica. 1994. *Noun phrase development in Swedish as a second language: A study of adult learners acquiring definiteness and the semantics and morphology of adjectives*. Stockholm University. (Doctoral dissertation).

Borin, Lars & Markus Forsberg. 2014. Swesaurus; or, the Frankenstein approach to Wordnet construction. In *Proceedings of the Seventh Global Wordnet Conference*, 215–223.

Borin, Lars, Markus Forsberg, Martin Hammarstedt, Louise Holmer & Arild Matsson. 2025. 10 korp: Språkbanken's word research platform. In *Sixty years of Swedish computational lexicography*. Dana Dannélls, Kristian Blensenius & Lars Borin (eds.). De Gruyter. 175–194. DOI: 10.1515/9783111577234-010.

Borin, Lars, Markus Forsberg, Martin Hammarstedt, Dan Rosén, Roland Schäfer & Anne Schumacher. 2016. Sparv: Språkbanken's corpus annotation pipeline infrastructure. In *The Sixth Swedish Language Technology Conference (SLTC), Umeå University*, 17–18.

Borin, Lars, Markus Forsberg & Lennart Lönngren. 2013. SALDO: a touch of yin to WordNet's yang. *Language Resources and Evaluation* 47(4). 1191–1211.

Borin, Lars, Markus Forsberg & Johan Roxendal. 2012. Korp-the corpus infrastructure of Språkbanken. In *LREC*, vol. 2012, 474–478.

Boulton, Alex & Nina Vyatkina. 2021. Thirty years of data-driven learning: Taking stock and charting new directions over time. *Language Learning and Technology* 25(3). 66–89.

Caines, Andrew & Paula Buttery. 2018. The effect of task and topic on opportunity of use in learner corpora. In *Learner corpus research: New perspectives and applications*. Vaclav Brezina & Lynne Flowerdew (eds.). London & New York: Bloomsbury Publishing Academic. 5–27.

Capel, Annette. 2010. A1–B2 vocabulary: Insights and issues arising from the English Profile Wordlists project. *English Profile Journal* 1(1). 1–11.

Capel, Annette. 2012. Completing the English Vocabulary Profile: C1 and C2 vocabulary. *English Profile Journal* 3. 1–14.

Capel, Annette. 2015. The English Vocabulary Profile. In Julia Harrison & Fiona Barker (eds.), *English Profile in practice*, 9–27. Cambridge University Press.

Clahsen, Harald. 1985. Profiling second language development: A procedure for assessing L2 proficiency. In *Modelling and assessing second language acquisition*. Kenneth Hyltenstam & Manfred Pienemann (eds.). (Multilingual Matters 18). Clevedon: Multilingual Matters. 283–331.

Collberg, Philippe. 2021. Grammatik i skolan: Till hjälp för skrivande? In Johan Brandtler & Mikael Kalm (eds.), *Nyanser av grammatik: Gränser, mångfald, fördjupning*. Studentlitteratur.

Council of Europe. 2020. *Common European Framework of Reference for Languages: Learning, teaching, assessment. Companion volume with new descriptors*. Council of Europe Publishing.

Crystal, David, Paul J Fletcher & Michael Garman. 1976. *The grammatical analysis of language disability: A procedure for assessment and remediation*. London: Edward Arnold.

Davidson, Douglas J, Peter Indefrey & Marianne Gullberg. 2008. Words that second language learners are likely to hear, read, and use. *Bilingualism: Language and Cognition* 11(1). 133–146.

François, Thomas, Elena Volodina, Ildikó Pilán & Anaïs Tack. 2016. SVALex: a CEFR-graded lexical resource for Swedish foreign and second language learners. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 213–219.

Gilquin, Gaëtanelle & Sylviane Granger. 2010. How can data-driven learning be used in language teaching? In Michael McCarthy & Anne O'Keeffe (eds.), *The Routledge Handbook of Corpus Linguistics*, 359–370. New York: Routledge.

Goddard, Cliff. 2012. Semantic primes, semantic molecules, semantic templates: Key concepts in the NSM approach to lexical typology. *Linguistics* 50(3). 711–743.

Goddard, Cliff & Susanna Karlsson. 2008. Re-thinking THINK in contrastive perspective: Swedish vs. English. In Cliff Goddard (ed.), *Cross-linguistic semantics*, 225–240. Philadelphia & Amsterdam: John Benjamins Publishing Company.

Granfeldt, Jonas & Malin Ågren. 2014. SLA developmental stages and teachers' assessment of written French: Exploring Direkt Profil as a diagnostic assessment tool. *Language Testing* 31(3). 285–305.

Granfeldt, Jonas, Pierre Nugues & Emil Persson. 2005. Direkt Profil: A system for evaluating texts of second language learners of French based on developmental sequences. In *43rd Annual Meeting of the Association of Computational Linguistics*, 53–60.

Granger, Sylviane. 1996. From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In Aijmer K. (ed.), *Languages in contrast. Text-based cross-linguistic studies*, 37–51. Lund University Press: Lund. http://hdl.handle.net/2078.1/75847.

Granger, Sylviane. 2002. A bird's eye view of learner corpus research. In Sylviane Granger, Joseph Hung & Stephanie Petch-Tyson (eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, 3–33. Second Language Acquisition & Foreign Language Teaching. Amsterdam & Philadelphia: Benjamins.

Granger, Sylviane. 2009. The contribution of learner corpora to second language acquisition and foreign language teaching. *Corpora and language teaching* 33. 13–32.

Granger, Sylviane. 2015. Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research* 1(1). 7–24. DOI: 10.1075/ijlcr.1.1.01gra.

Hammarstedt, Martin, Anne Schumacher, Lars Borin & Markus Forsberg. 2022. *Sparv 5 user manual*. Göteborg.

Hawkins, John A & Luna Filipović. 2012. *Criterial features in L2 English: Specifying the reference levels of the Common European Framework*, vol. 1. Cambridge University Press.

Hult, Ann-Kristin, Sven-Göran Malmgren & Emma Sköldberg. 2010. Lexin - a report from a recycling lexicographic project in the north. In Anne Dykstra & Tanneke Schoonheim (eds.), *Proceedings of the XIV Euralex International Congress* (*Leeuwarden, 6-10 July 2010*), 800–809. Leeuwarden/Ljouwert, The Netherlands: Fryske Akademy. https://euralex.org/category/publications/euralex-leeuwarden-2010/.

Hultman, Tor & Margareta Westman. 1977. *Gymnasistsvenska*. Liber-Läromedel.

Källgren, Gunnel, Sofia Gustafson-Capková & Britt Hartmann. 2006. *Manual of the Stockholm Umeå Corpus version 2.0*. Department of Linguistics, Stockholm University. https://spraakbanken.gu.se/parole/Docs/SUC2.0-manual.pdf.

Keßler, Jörg-U & Mathias Liebner. 2011. Diagnosing L2 development: Rapid profile. In Manfred Pienemann & Jörg-U Keßler (eds.), *Studying processability theory: An introductory textbook*, 133–148. Amsterdam & Philadelphia: John Benjamins Publishing Company.

Lindström Tiedemann, Therese, David Alfter, Yousuf Ali Mohammed, Daniela Helena Piipponen, Beatrice Silen & Elena Volodina. 2024a. Multiword expressions in Swedish as a second language: Taxonomy, annotation, and initial results. In *Multiword expressions in lexical resources*. Voula Giouli & Verginica Barbu Mititelu (eds.). (Phraseology and Multiword Expressions). Germany: Language Science Press. 309–348. DOI: 10.5281/zenodo.10998647.

Lindström Tiedemann, Therese, David Alfter & Elena Volodina. 2022. CEFR-nivåer och svenska flerordsuttryck. In Siv Björklund, Marianne Haagensen Bodil Nordman & Anders Westerlund (eds.), *Svenskan i Finland 19: Föredrag vid den nittonde sammankomsten för beskrivningen av svenskan i Finland, Vasa den 6–7 maj 2021*, 218–233. Svensk-Österbottniska Samfundet.

Lindström Tiedemann, Therese, David Alfter & Elena Volodina. 2023a. *L2LexAdj* (*updated: 2023-04-20*).[*Data set*]. DOI: 10.23695/z8wj-1r23.

Lindström Tiedemann, Therese, David Alfter & Elena Volodina. 2023b. *L2LexAdjAdv* (*updated: 2023-04-20*).[*Data set*]. DOI: 10.23695/3kcx-s405.

Lindström Tiedemann, Therese, David Alfter & Elena Volodina. 2023c. *Swedish MWELex* (*updated: 2023-04-20*).[*Data set*]. DOI: 10.23695/352q-wa92.

Lindström Tiedemann, Therese, Yousuf Ali Mohammed & Elena Volodina. In preparation. *Swedish grammar profiling for empirical L2 research and teaching*.

Lindström Tiedemann, Therese, Daniela Piipponen, Lisa Södergård & Erik Axelson. 2024b. *Normalisering som analysredskap för andraspråkssvenska*. Unpublished presentation at Svenskan i Finland 21.

Löhndorf, Simone. 2021. *Development of adjectival use and meaning structures in Swedish students' written production*. Lund University. (Doctoral dissertation).

Mohsen, Mohammed Ali, Sultan Althebi, Rawan Alsagour, Albatool Alsalem, Amjad Almudawi & Abdulaziz Alshahrani. 2024. Forty-two years of computer-assisted language learning research: A scientometric study of hotspot research and trending issues. *ReCALL* 36(2). 230–249.

Muñoz Sánchez, Ricardo, David Alfter, Simon Dobnik, Maria Irena Szawerna & Elena Volodina. 2024. Jingle BERT, Jingle BERT, frozen all the way: Freezing layers to identify CEFR levels of second language learners using BERT. In *Proceedings of the Workshop on Natural Language Processing for Computer-Assisted Language Learning* (*NLP4CALL 2024*).

Muñoz Sánchez, Ricardo, Simon Dobnik, Maria Irena Szawerna, Therese Lindström Tiedemann & Elena Volodina. 2024. Did the names I used within my essay affect my score? Diagnosing name biases in automated essay scoring. In *Proceedings of the workshop on computational approaches to language data pseudonymization* (*CALD-pseudo 2024*), 81–91. https://aclanthology.org/2024.caldpseudo-1.10/.

Nieto Piña, Luis. 2019. *Splitting rocks: Learning word sense representations from corpora and lexica*. Data Lingvistica 30, University of Gothenburg.

Paquot, Magali. 2019. The phraseological dimension in interlanguage complexity research. *Second language research* 35(1). 121–145.

Pawley, Andrew & Frances Hodgetts Syder. 1983. Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In Jack C. Richards & Richard W. Schmidt (eds.), *Language and communication*, 191–226. London: Longman.

Perc, Matjaž. 2014. The Matthew effect in empirical data. *Journal of The Royal Society Interface* 11(98). 20140378. DOI: 10.1098/rsif.2014.0378.

Philipsson, Anders. 2007. *Interrogative clauses and verb morphology in L2 Swedish: Theoretical interpretations of grammatical development and effects of different elicitation techniques*. Centrum för tvåspråkighetsforskning. (Doctoral dissertation).

Pienemann, Manfred & Alison Mackey. 1993. An empirical study of children's ESL development and Rapid Profile. *ESL development: Language and literacy in schools* 2. 115–259.

Pilán, Ildikó & Elena Volodina. 2018. Investigating the importance of linguistic complexity features across different datasets related to language learning. In *Proceedings of the workshop on linguistic complexity and natural language processing*, 49–58.

Pilán, Ildikó, Elena Volodina & Lars Borin. 2016. Candidate sentence selection for language learning exercises: From a comprehensive framework to an empirical evaluation. *Traitement Automatique des Langues* 57(3). 67–91.

Pilán, Ildikó, Elena Volodina & Torsten Zesch. 2016. Predicting proficiency levels in learner writings by transferring a linguistic complexity model from expert-written coursebooks. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers*, 2101–2111.

Ravid, Dorit & Ronit Levie. 2010. Hebrew adjectives in later language text production. *First Language* 30(1). 27–55. DOI: 10.1177/0142723709350529.

Römer, Ute. 2023. Usage-based approaches to Second Language Acquisition vis-à-vis Data-Driven Learning. *TESOL Quarterly*.

Schmidt, Richard W. 1990. The role of consciousness in second language learning. *Applied linguistics* 11(2). 129–158.

Sköldberg, Emma, Louise Holmer, Elena Volodina & Ildikó Pilán. 2019. State-of-the-art on monolingual lexicography for Sweden. *Slovenščina 2.0: empirične, aplikativne in interdisciplinarne raziskave* 7(1). 13–24.

Søgaard, Anders. 2022. Should we ban English NLP for a year? In *Proceedings of the 2022 conference on empirical methods in natural language processing*, 5254–5260.

Teleman, Ulf, Staffan Hellberg & Erik Andersson. 1999. *Svenska akademiens grammatik*. Svenska Akademien & Norstedts ordbok.

Tenfjord, Kari, Paul Meurer & Knut Hofland. 2006. The ASK corpus: A language learner corpus of Norwegian as a second language. In *LREC'06*, 1821–1824.

Viberg, Åke. 1990. Svenskans lexikala profil. In Erik Andersson & Marketta Sundman (eds.), *Svenskans beskrivning 17*. Åbo Akademi.

Viberg, Åke. 1992. Tvärspråklig lexikologi med svenskan i centrum. *Nordiske studier i leksikografi* 1.

Viberg, Åke. 2006a. The typological profile of Swedish – an introduction. *STUF - Language Typology and Universals* 59. 3–11.

Viberg, Åke. 2006b. Towards a lexical profile of the Swedish verb lexicon. *STUF - Language Typology and Universals* 59. 103–129.

Viberg, Åke. 2013. Seeing the lexical profile of Swedish through multilingual corpora. The case of Swedish åka and other vehicle verbs. In Karin Aijmer & Bengt Altenberg (eds.), *Advances in Corpus-based Contrastive Linguistics: Studies in Honour of Stig Johansson*. John Benjamins.

Volodina, Elena. 2024. On two SweLL learner corpora–SweLL-pilot and SweLL-gold. In *Proceedings of the Huminfra Conference (HiC 2024)*, 83–94. DOI: 10.3384/ecp205012.

Volodina, Elena, David Alfter, Therese Lindström Tiedemann, Maisa Susanna Lauriala & Daniela Helena Piipponen. 2022. Reliability of automatic linguistic annotation: native vs non-native texts. In *Selected papers from the CLARIN Annual Conference 2021*.

Volodina, Elena, David Alfter & Therese Lindström Tiedemann. 2022. Crowdsourcing ratings for single lexical items: a core vocabulary perspective. *Slovenščina 2.0: empirične, aplikativne in interdisciplinarne raziskave* 10(2). 5–61.

Volodina, Elena, David Alfter & Therese Lindström Tiedemann. 2024. Profiles for Swedish as a second language: lexis, grammar, morphology. In *Proceedings of the Huminfra Conference (HiC 2024)*, 10–19. DOI: 10.3384/ecp205002.

Volodina, Elena, Yousuf Ali Mohammed & Therese Lindström Tiedemann. 2022. Lyxig språklig födelsedagspresent from the Swedish word family. In Elena Volodina, Dana Dannélls, Aleksandrs Berdicevskis, Markus Forsberg & Shafqat Virk (eds.), *Live and Learn: Festschrift in honor of Lars Borin*. GU-ISS-2022-03, 153–160. Department for Swedish, Multilingualism, Language Technology, University of Gothenburg.

Volodina, Elena, Yousuf Ali Mohammed & Therese Lindström Tiedemann. 2021. CoDeRooMor: A new dataset for non-inflectional morphology studies of Swedish. In *Proceedings of the 23rd Nordic conference on computational linguistics (NoDaLiDa)*, 178–189.

Volodina, Elena, Yousuf Ali Mohammed & Therese Lindström Tiedemann. 2024. Swedish word family resource: Construction, applicability, strengths and first experiments. *ITL-International Journal of Applied Linguistics*. DOI: 10.1075/itl.22026.vol.

Volodina, Elena & Dijana Pijetlovic. 2015. Lark trills for language drills: Text-to-speech technology for language learners. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, 107–117.

Volodina, Elena & Ildikó Pilán. 2014. *COCTAILL.* [*Data set*]. DOI: 10.23695/v8xs-p564.

Volodina, Elena & Ildikó Pilán. 2016. *SweLLex.* [*Data set*]. DOI: 10.23695/6h3v-zw25.

Volodina, Elena, Ildikó Pilán & David Alfter. 2016. Classification of Swedish learner essays by CEFR levels. *CALL communities and culture–short papers from EUROCALL* 2016. 456–461.

Volodina, Elena, Ildikó Pilán, Lars Borin & Therese Lindström Tiedemann. 2014. A flexible language learning platform based on language resources and web services. In *LREC*, 3973–3978.

Volodina, Elena, Ildikó Pilán, Stian Rødven Eide & Hannes Heidarsson. 2014. You get what you annotate: a pedagogically annotated corpus of coursebooks for Swedish as a second language. In *Proceedings of the 3rd Workshop on NLP for Computer-Assisted Language Learning*. Linköping University Press.

Volodina, Elena, Ildikó Pilán, Ingegerd Enström, Lorena Llozhi, Peter Lundkvist, Gunlög Sundberg & Monica Sandell. 2016a. Swell on the rise: Swedish learner language corpus for European reference level studies. In *Proceedings of the 10th Language Resources and Evaluation Conference* (*LREC*). Portorož, Slovenia.

Volodina, Elena, Ildikó Pilán, Ingegerd Enström, Lorena Llozhi, Peter Lundkvist, Gunlög Sundberg & Monica Sandell. 2016b. *SweLL-pilot.* [*Data set*]. DOI: 10.23695/6h3v-zw25.

Volodina, Elena, Ildikó Pilán & Thomas François. 2016. *SVALex, v.01.* [*Data set*]. DOI: 10.23695/rwrq-1w38.

Volodina, Elena, Ildikó Pilán, Lorena Llozhi, Baptiste Degryse & Thomas François. 2016. SweLLex: second language learners' productive vocabulary. In *Proceedings of the joint workshop on NLP for computer assisted language learning and NLP for language acquisition*, 76–84.

Warren, Martin. 2016. Introduction to data-driven learning. In Fiona Farr & Liam Murray (eds.), *The Routledge Handbook of Language Learning and Technology*, 337–347. London & New York: Routledge.

Wierzbicka, Anna. 1996. *Semantics: Primes and universals*. Oxford: Oxford University Press.

Wray, Alison. 2002. *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Wulff, Stefanie. 2020. Usage-based approaches. In Nicole Tracy-Ventura & Magali Paquot (eds.), *The Routledge Handbook of Second Language Acquisition and Corpora*, 175–188. New York & Oxford: Routledge.

Yamaguchi, Nami, David Alfter, Kaori Sugiyama & Thomas François. 2022. Towards a verb profile: distribution of verbal tenses in FFL textbooks

and in learner productions. In *Proceedings of the 11th workshop on natural language processing for computer-assisted language learning* (*NLP4CALL 2022*), 123–142.

List of abbreviations

CALL Computer-Assisted Language Learning
CEFR Common European Framework of Reference for Languages
CIA Contrastive Interlanguage Analysis
COCA Corpus of Contemporary American English
COCTAILL COrpus of Coursebook Texts As Input for Language Learning
DDL Data Driven Learning
ICALL Intelligent Computer-Assisted Language Learning
L1 first language, cf. mother tongue
L2 second language, i.e. non-native language
L2Lex-Adj a language resource containing a L2 list of adjectives with associated information used as an input to a subsection of Swedish L2 lexical profile devoted to declension of adjectives
L2Lex-AdjAdv a language resource containing a L2 list of adjectives and adverbs with associated information used as an input to a subsection of Swedish L2 lexical profile devoted to adverbial and adjectivar structure
MWE Multi-Word Expressions
NLP Natural Language Processing
NSM Natural Semantic Metalanguage
POS part of speech
SAG Swedish Academic Grammar
SAOL Svenska Akademiens Ordlista
Sen*Lex sense-based lexicon for L2 Swedish
SenSVALex sense-based SVALex
SenSweLLex sense-based wordlist of Swedish as a second language based on learner essays (SweLL-pilot)
SLA Second Language Acquisition
SO Svenska ordbok
SVALex wordlist of Swedish as a second language based on course books (COCTAILL)
SweL2P Swedish Second Language Profile

SweLL	corpus of learner essays, Swedish Learner Language
SweLLex	wordlist of Swedish as a second language based on learner essays (SweLL-pilot)
Swe-MWELex	a language resource containing an L2 list of mult-word expressions with associated information used as an input for a subsection of Swedish L2 lexical profile devoted to multi-word expressions
TTR	type-token ratio

Corresponding authors

Therese Lindström Tiedemann
Department of Finnish, Finno-Ugrian and Scandinavian Studies
University of Helsinki
therese.lindstromtiedemann@helsinki.fi

Elena Volodina
Språkbanken Text
Department of Swedish, Multilingualism, Language Technology
University of Gothenburg
elena.volodina@svenska.gu.se

David Alfter
Department of Literature, History of Ideas, and Religion
University of Gothenburg
david.alfter@lir.gu.se

12 Unlocking Swedish historical material through OCR and HTR

Dana Dannélls
University of Gothenburg

Robin Kurtz
National Library of Sweden

Erik Lenas
Swedish National Archives

Viktoria Löfgren
Swedish National Archives

This chapter presents some of the efforts made by three national institutions and the challenges each institution encountered while advancing Optical Character Recognition (OCR) and Handwritten Text Recognition (HTR) technologies for Swedish historical material. It introduces the resources, models, and tools that can be used to refine computational approaches for improving OCR and HTR processing, which, in turn, could enhance text- and data-driven research.

1 *Introduction*

The technology that allows libraries and cultural institutions to convert analog data into digital format, enabling searches by character, word, or phrase, is called Optical Character Recognition (OCR) for printed material and handwritten text recognition (HTR) for handwritten material.

To date, all digitized collections in Sweden have been processed using OCR/HTR models trained on materials with varying layouts and quality. As a result, the digitization process has introduced noisy data into the material. Enhancing these models and, consequently, improving the quality of the processed material relies heavily on the availability of diverse, high-quality Swedish language resources, such as corpora, annotated datasets, lexicons, computational tools, and language models. Language resources play a crucial role in the analysis and interpretation of linguistic information in printed or written form (Baron & Rayson 2009, Rigaud et al. 2019,

HUM INFRA Dana Dannélls, Robin Kurtz, Erik Lenas & Viktoria Löfgren. 2025. Unlocking Swedish historical material through OCR and HTR. In Gerlof Bouma, Dana Dannélls, Dimitrios Kokkinakis & Elena Volodina (eds.), *Huminfra handbook: Empowering digital and experimental humanities* (NEALT Proceedings Series 59), 405–431. University of Tartu Library. DOI: 10.58009/aere-perennius0181 © The authors, CC BY 4.0

Ehrmann et al. 2022) Although not as large as English, Swedish has a considerable amount of advanced language technology resources and tools.[1] Particularly over the last decade, advances in transformer techniques and fine-tuning methods —together with open-source implementations for pre-training general language models — have made cutting-edge Swedish large language models and large-scale datasets available.

Owing to these rapid developments and recognizing the advantages they have for improving the results of OCR and HTR processes, a workshop was held at Gothenburg University in the spring of 2024 to discuss the challenges that are associated with the latest techniques. The workshop brought together representatives from Språkbanken Text, KBLab, the datalab of the National Library of Sweden (Kungliga biblioteket – KB) and the AI lab at the National Archives of Sweden (Riksarkivet).

The workshop aimed to showcase the recently developed Swedish tools and resources and to discuss the challenges associated with them. Participants also explored strategies to address these challenges in the ongoing work on OCR and HTR for historical Swedish documents. In this paper, we summarize the tools and datasets that have been presented at the workshop. To set the stage, we first provide a short overview of the groups that participated in the workshop.

1.1 *Språkbanken Text*

Språkbanken Text is a leading research infrastructure for Swedish,[2] and a national center for the collection and refinement of language resources, primarily Swedish text corpora and lexicon resources. These resources are freely available to the general public for research, teaching, and other purposes, either through inclusive platforms such as Korp (Borin et al. 2025) and Strix (Forsberg et al. 2025).

Part of the infrastructure work is to develop computational tools for linguistic analysis to enrich the resources with linguistic data equipped from structured lexical resources. With rich linguistic resources, new possibilities for data-driven humanities research become feasible, from large-scale diachronic studies to fine-grained discourse analysis.

Most of the resources at Språkbanken Text are texts, some of them are historical resources from 1600 and onwards. They have often been added to the repository after being scanned and digitized elsewhere without any corrections of the erroneous character sequences or words in the text, which decreases the quality of the linguistic analysis, and in turn, restricts the

1 Swedish is the 5th largest Wikipedia by article count according to meta Wikipedia.
2 https://spraakbanken.gu.se/en

information retrieved from the text. Consequently, improving post-OCR text quality has centered on leveraging rich lexical resources (Dannélls et al. 2021a) and on the ongoing development of novel methods. The latest method presented by Löfgren & Dannélls (2024) has been added as a plugin to Mink,[3] a data platform that allows users to upload language data to a local computer and access them via Strix,[4] a document-centered search platform.

1.2 *KBLab — The National Library of Sweden*

The National Library of Sweden[5] collects, preserves, and gives access to almost everything that is published in Sweden, including books, newspapers, radio, TV, and more. Its collections currently hold over 18 million items, which are continuously digitized to make them more accessible.

KBLab[6] was founded in 2019 as the library's *datalab* to give researchers the possibility to do large-scale quantitative research, to curate data maintained by the National Library, and to train models on the data maintained by the library to be used by academia, governmental organizations, and the industry (Börjeson et al. 2024).

A new initiative at the National Library, KB, aims to leverage Artificial Intelligence (AI) and Machine Learning (ML) to enhance the accessibility of its collections, using, for example, speech-to-text models to make the enormous radio collections easily searchable via text, or text-vision models to enable searching for specific images based on descriptions.[7] With this in mind it is especially important to improve the OCR quality for newspapers for services such as tidningar.kb.se.

1.3 *The National Archives*

The Swedish National Archives[8] has collected and preserved the archives and data from government authorities since the early 17th century. Large-scale digitization is an ever ongoing project, and as of now, the National Archives has scanned about 270 million document images, which represents about 5 percent of the entire archives.

A large proportion of these documents are handwritten. The National Archives started working with HTR in 2019. Today, this work is continued

3 https://spraakbanken.gu.se/mink/
4 https://spraakbanken.gu.se/strix/
5 https://www.kb.se/
6 https://www.kb.se/in-english/research-collaboration/kblab.html
7 A demo for searching through a collection of postcards can be accessed here https://lab.kb.se/bildsok/
8 https://riksarkivet.se

by the newly established AI lab, which aims to scale up the extraction of searchable text from scanned documents through HTR and OCR. This makes the archive materials available for data-driven research in a way that was previously impossible. The AI lab also works with creating databases and search indexes for frequently requested documents using HTR and OCR, which streamlines the work of archivists, and makes case-handling much more effective.

The challenges of large-scale HTR/OCR projects are not only theoretical, but to a large extent also practical. The National Archives are currently investing in hardware, infrastructure and software development for scaling up the successful transformation of scanned documents to digital text or structured databases.

2 *Research problem and relevance to Digital / Experimental Humanities*

Historical digitized collections are large – on the order of a hundred million pages – and unique cultural heritage datasets containing information that rarely is conveyed in other media types. For humanities scholars, this material holds significant value, offering insight into societal contexts and historical descriptions. The original images underlying the collections are characterized by low document quality: complex layouts, a mixture of typefaces, headings in various sizes, images, enriched with unique typographical features, illustrations, and decorative elements. The challenges exhibited by different types of material are illustrated in Figure 1. To add to the complexity, because documents have been scanned using different systems with diverse ML techniques, leading to varying quality results over the years. In turn, the poor quality of the processed material makes it inaccessible to humanities scholars, who are unable to locate what they need because the vast range of sources and materials available for research remains out of reach for digital humanities work.

3 *OCR and HTR workflow*

The input to an OCR/HTR process is an image or a set of images. The output of the process is the automatically extracted text, specifically, the extracted characters, words and sentences, represented in the image.

Several commercial and open-source OCR and HTR software are offered on the market, with the most prominent ones being Abbyy FineReader,[9]

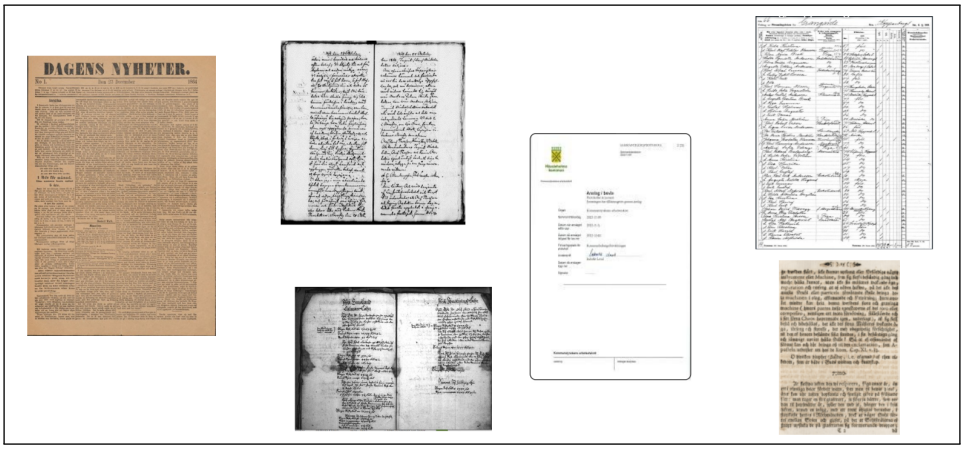


Figure 1: Historical digitized documents with varying complexities

Transkribus,[10] Tesseract,[11] Calamari,[12] and OCRopus.[13] Despite this, almost all libraries in Sweden are still using the commercial Abbyy FineReader and recently Transkribus as their core for OCR/HTR processing. Each software tool has its own built-in intelligence, which is applied slightly differently across various digitization workflows. Due to advances in digitization techniques, OCR results from pages processed a decade ago often differ from those processed more recently (Holley 2009). This is partially because older software versions were used in earlier digitization efforts.

A typical OCR/HTR process consists of four to five sequential steps, as shown in Figure 2: (1) Data/image processing, including binarization and normalization operations such as correcting skew lines and removing noise; (2) Segmentation following the image layout structure; (3) Feature extraction for simplifying the image data by detecting colours, objects and pixel points within the image object bounding box; (4) Character recognition using a language model; and (5) Post-correction of the output results from the language model. Over the years, researchers faced challenges in obtaining accurate results in each step, but thanks to advances in AI methods, some of the steps are no longer a challenge. However, since each step depends on the previous one, error propagation remains. In the following, we will describe each of the steps separately.

9 https://help.abbyy.com/en-us/
10 https://www.transkribus.org/
11 https://github.com/tesseract-ocr/tesseract
12 https://github.com/Calamari-OCR/calamari
13 https://github.com/ocropus

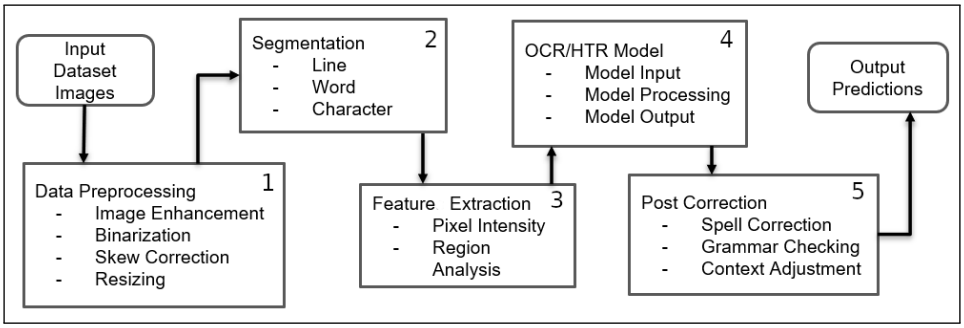


Figure 2: Stages of OCR/HTR processing

3.1 *Input images/dataset*

To understand the complexity of the task, consider the various formats of the scanned document provided as the input image to the OCR/HTR engine, as can been seen in Figure 1. Documents are ranging from single-column documents to multi-column layouts, different line lengths, and even complex table formats. Unfortunately, there is no single model capable of handling this diversity; instead, in the next preprocessing step each model must be trained to accommodate a specific input format.

3.2 *Data preprocessing*

The goal of this step is to improve the quality of the source document by addressing the damage caused by paper aging, usage, and wear and tear, as well as removing noise from faded characters, ink splodges, and bleed–through.

Depending on the quality of the source, an appropriate strategy, covering everything from simple binarization to advanced deep-learning techniques, must be chosen. The research field that is dedicated to solving these particular issues, is known as document analysis systems (Sfikas & Retsinas 2024). Notably, since 2020, the field has made significant breakthroughs thanks to advancements in transformer-based architectures (Wolf et al. 2020), which have made OCR and HTR systems less sensitive to the quality of the source.

3.3 *Layout analysis*

Layout analysis or segmentation, is the identification of boundaries between columns, lines, words, and characters. Things that affect the results of this step are small margins between lines and columns, or mixed fonts on the same page. The method for segmenting an image or a document

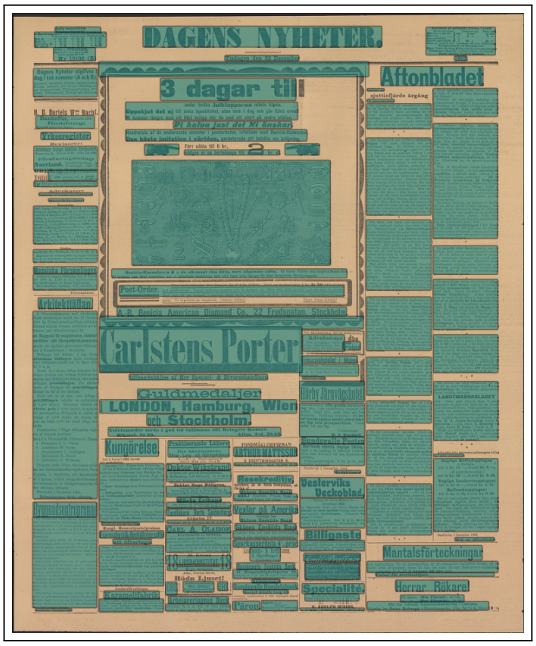


Figure 3: A newspaper page marked with identified bounding boxes

depends heavily on whether the document is structured, unstructured or semi-structured. For example, a regular book with one block of text per page containing a header and the page number is much easier to segment than a multi-column layout with images or headlines covering multiple columns as in Figure 3.

Similar to the previous step, this process can be entirely language-independent and is often examined within research focusing on document-level analysis (Sfikas & Retsinas 2024).

3.4 *Feature extraction*

The feature extraction stage is applied to capture the characteristics of characters and symbols in the image. It involves analyzing and extracting either numerical or structural attributes that describe the shape, structure, and distribution of pixels in characters. For example, a numerical feature extraction of the character <B> can result in a distribution of pixels as shown in Figure 4, and have a horizontal and a vertical feature vector that summarize the distribution that is: [4 2 2 4 2 2 4] [7 3 3 3 4].

1	**1**	**1**	**1**	0
1	0	0	0	**1**
1	0	0	0	**1**
1	**1**	**1**	**1**	0
1	0	0	0	**1**
1	0	0	0	**1**
1	**1**	**1**	**1**	0

Figure 4: The distribution of pixels of "B"

Off-the-shelf OCR and HTR systems apply different methods to train character models for recognizing mixed type faces (i.e. Fraktur, Schwabacher) and language specific characters, such as 'ä', 'ö', and 'å'. Usually, these systems do not have enough data to train character models for historical text, and consequently, they struggle to recognize uncommon characters such as rare ligatures because these characters were absent in the training material. Another problematic aspect is typically caused by limited predictability and generality, difficulty in distinguishing between similar characters and sensibility to translation Therefore, accurate character models are crucial for better system performance.

3.5 *OCR and HTR models*

OCR and HTR models are trained to recognize characters and words and classify them accordingly. The training is done with the help of a set of examples, where both the internal value (the image) and the target value (the text) are given. Based on the examples, the system makes various predictions from unseen images.

When training OCR and HTR models, the most important ingredient is the data that the model will be trained and tested/evaluated on. Training can be done on (semi-)automatically annotated data, which is data that was automatically annotated but additionally proofread by human annotators or annotated data generated by some model with high confidence. Evaluation generally involves human annotators who read through the source documents and label relevant information in different parts of the document. As this process is usually very costly, annotated datasets are rare and small. Different techniques and approaches to increasing the amount of data for training and evaluating OCR and HTR models have been proposed, for an overview and further reading see Agarwal & Anastasopoulos (2024).

3.6 *Post-correction*

Post-correction refers to the process applied to OCR or HTR output after the text has been scanned and processed. It is commonly used to correct residual or newly introduced processing errors. A straightforward approach to post-correction involves word replacement, in which edit-distance methods are applied to generate a list of candidate words, and the erroneous word is replaced with the closest match from a dictionary. This method can be further refined by incorporating word frequency data from large corpora to weight the probability of each candidate word, thereby increasing the likelihood of selecting the most accurate replacement.

Another approach to post-correction is context-based, meaning that the text sequence is not predicted greedily character by character, but is instead constructed by predicting the most probable token based on its preceding or surrounding context (Nguyen et al. 2021).

The post-correction step is typically integrated into a complete OCR/HTR system, but it can also be extended by incorporating an additional model to correct any remaining errors. In recent work on Swedish OCR for historical text, Löfgren & Dannélls (2024) implemented a character-based post-OCR method that significantly reduced both character and word error rates.

3.7 *Output predictions*

The output text from OCR/HTR is generally structured, often encoded in XML or ALTO files, and follows the hierarchical layout identified by the model during layout analysis (see Section 3.3). For multi-page documents, the output includes separate sections for each page, which in turn contain nested elements such as text blocks, columns, lines, or smaller units. Each element in this hierarchy typically includes metadata about its position in the document—such as page number—or its exact coordinates on the page (bounding box). It also specifies the type of content it represents, such as text, image, or table. While XML is a common format for this structured output, specialized formats like hOCR are also frequently used.

3.8 *Evaluation*

Evaluation is not part of the core OCR/HTR processing workflow, but it is a crucial step in assessing the performance of any system. To evaluate OCR and HTR systems, their output is compared to a reference—i.e., the correct text. Two standard metrics are commonly used to quantify errors: Character Error Rate (CER) and Word Error Rate (WER).

CER and WER are calculated by determining the number of edits—substitutions (S), deletions (D), and insertions (I)—needed to transform the output to the reference, by the number of characters/words in the reference (N):[14]

$$\text{CER/WER} = \frac{S + D + I}{N}.$$

While WER gives us a good estimate about the final performance of the OCR/HTR system, CER helps to get insight into specific character errors, such as regularly missing diacritics.

4 *Data and models for OCR/HTR processing*

OCR and HTR models for converting image to text are trained primarily on language data. To improve the accuracy of these models more data and in particular, reference data is required.[15] When the models have been trained and applied to new data, the results have to be evaluated against manually corrected versions of the data.

4.1 *Språkbanken Text*

As already mentioned, Språkbanken Text collects, annotates, and preserves language data, in particular written text. In this section we describe the datasets that are available for training and evaluating Swedish OCR models. All data has been released under CC BY license.[16]

Svensk Fraktur

The data is a collection of 199 predominantly Fraktur,[17] printed between approximately 1626 and 1816, that was collected within the project "A free cloud service for OCR"; a collaboration between Språkbanken Text and the Gothenburg University Library (Borin et al. 2016). The transcriptions were manually annotated with typographic information such as font changes and ligatures by a company that specializes in double-keying processes.[18]

14 Note that the error rate is not necessarily between one and zero.

15 Reference data or gold standard or ground-truth data – all terms are used interchangeably – is the correct version of the automatically processed data, manually curated, ideally by more than one human annotator.

16 https://spraakbanken.gu.se/resurser?s=OCR&language=All

17 Fraktur is a type of script used for writing European languages that originated in the 12th century.

18 Double-keying is a verification method to ensure accuracy, it involves entering the same data twice – usually by two different people and then comparing the two entries.

Table 1: An overview of the Swedish datasets with manual transcriptions

Dataset	OCR system	Time period	Characters
Svensk Fraktur	OCRopus	1626–1816	282 432
Then swänska Argus	OCRopus	1732–1734	259 468
Literature	Abbyy	1836–2001	7 266 866
Newspapers	Tesseract	1818–2018	6 956 748
	Abbyy		6 928 171
	Abbyy-Tesseract		6 922 088

Then Swänska Argus

The second source of blackletter data is *Then Swänska Argus*, which was prepared within the same project as *Swedish fraktur* (Borin et al. 2016). This dataset consists of 25 issues of *Then swänska Argus*, a periodical by Olof von Dalin published between 1732 and 1734 (Språkbanken Text 2020). It is significantly older than the newspaper texts, but the language is casual and simple for its time, and it is often cited as the first example of Late Modern Swedish (*yngre nysvenska*).

The transcription is based on a transcription made at Uppsala University, with some modification by Borin et al. (2016). The original *Then swänska Argus* has been republished at least once, and the transcription does not appear to be directly tied to a specific printing. As a result, the transcription does not contain any line breaks. In order to align it with the OCR output, line breaks were inserted in the transcription. The line breaks were inserted such that the sum of the CER of each individual line was minimized.

Literature

The literature dataset consists of 79 titles of Swedish literature printed between 1836 and 2001. In total, it contains about seven million characters, making it roughly the same size and from the same period as the newspapers (see Table 1). The OCR quality is generally much higher than the newspapers', most likely because of higher paper quality and simpler page layout. The data was provided as XML files by the Swedish Literature Bank, *Litteraturbanken*.[19]

19 https://litteraturbanken.se/

Table 2: The 10 most common character errors made by the single model (left) and after voting using an ensemble of 5 models (right).

Single model				Conf. voted models			
GT	Pred.	Count	Percentage	GT	Pred.	Count	Percentage
ä	a	124	4.06	ä	a	118	5.43
ö	o	95	3.11	å	ä	58	2.67
å	ä	67	2.19	ö	o	52	2.39
__	_	53	1.73	″	″	46	2.12
u	n	45	1.47	å	a	34	1.56
ä	å	39	1.28	n	u	31	1.43
″	″	38	1.27	ſ	f	29	1.33
f	ſ	37	1.21	ä	å	22	1.01
i	l	33	1.08	r	t	22	1.01
				f	ſ	22	1.01

GT: Ground Truth, the true character; Pred.: the predicted character; Count: the number of occurrences; Percentage: the percentage of the total number of errors. Note that GT "__" and Pred. "_" means a space has been deleted

Newspapers

The newspaper collection is partially a free digital reference resource featuring content from newspapers published between 1818 and 2018 (Dannélls et al. 2019). It is a subset of the largest newspaper collection at KB called Kubhist (see Section 4.2). A total of 200 digitized newspapers were selected from Kubhist, with one newspaper chosen per year. These newspapers were carefully curated to represent common variations in layout and typography. From each newspaper, the second and fourth pages were selected, resulting in a collection of 400 pages, each was segmented down to the paragraph level. Similar to Svensk Fraktur, this dataset was manually transcribed by an external agency which specializes in double-keying. Additionally, manual annotations were conducted at both the document and paragraph levels by two undergraduate students with backgrounds in language technology and linguistics. Due to copyright limitations, only a subset of the collection, spanning up to 1906 is freely available.

Reference data enables the analysis of OCR models' performance and helps to identify what types of errors OCR systems tend to make. Brandt Skelbye & Dannélls (2021) ran a qualitative analysis of the OCR errors in a selection of newspaper editions from 1818–1848, and listed the 10 most common character errors made by OCR models trained on Swedish historical data in Calamari (see Table 2).

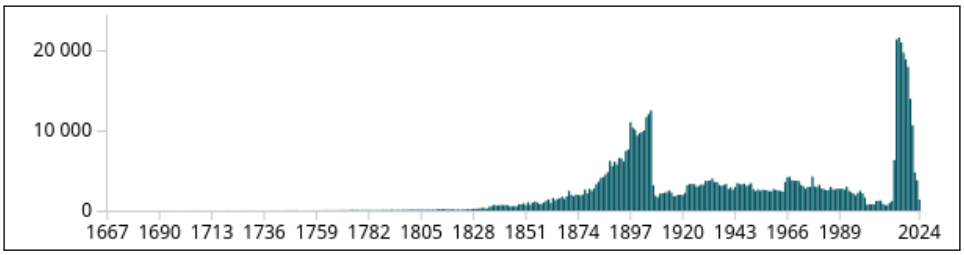


Figure 5: Misleading frequencies for the search term *örn* on tidningar.kb.se

Among the most common character errors, we find a mixture of loss of diacritics and ligatures such as for *'a' 'ä' 'å'*, or the mixing up of single letters. However, OCR errors are not limited to single characters. Another common error occurs in certain letter combinations such as *rn*, which are identified as *m* or vice versa. As such, searching for articles about *örn* 'eagles' in Swedish newspapers (see Figure 5) becomes next to impossible due to mix-ups with the more frequent *om* 'if', showcasing the need for better OCR systems for Swedish.

4.2 *National Library of Sweden*

The by far largest collection of digitized data at the National Library of Sweden is its newspaper collection, ranging from 1645 to today. The printed versions are delivered to the national library where they are digitized, and OCR processed with a commercial OCR system based on Abbyy FineReader. While KBLab has redone the OCR processing for some smaller collections, such as the proceedings of the Swedish parliament with a more recent version of Tesseract to increase the OCR quality, overhauling the complete newspaper collection is out of scope due to its sheer size (nearly 40 million pages currently available via tidningar.kb.se).

Another challenge with the digitization of newspapers lies in their layout. Extracting a specific article from one page requires finding all text-boxes belonging to the article, and determining their order. While newspaper layouts typically follow a left-to-right and up-down reading order, it is not always clear whether the text wraps around a picture or whether the picture serves as a barrier between articles, where the article actually starts and ends, and finally whether some sections belong to a different article, are an extension, or are something else such as advertisements (Rekathati 2021).

Any improvement on the OCR quality itself greatly increases this collection's value and avoids the need to apply fixes on the data as in Malmsten et al. (2020). While it is important for a cultural heritage institution to preserve

Table 3: Performance of the two-OCR engine and the two post-OCR correction models based on byT5

Model	CER (%)	WER (%)
Two-OCR	2.93	12.05
Löfgren & Dannélls (2024)	1.92	7.41
KBLab	1.57	6.23

the layout of newspapers, it is preferable for training language models and the analysis of the collection in general to be able to extract complete articles.

Post-OCR correction

The first step in improving the OCR quality of KB's newspaper collections was to follow Löfgren & Dannélls (2024) and train a new post-OCR correction model using a similar set-up. This model uses only newspaper data and is trained on a Nvidia DGX A100 with 8 GPUs, for much longer (75 iterations over the training data compared to 3), using a different learning rate scheduler, optimizer, and other related hyperparameters. The previous model by Löfgren & Dannélls (2024) reduced the WER by 39% compared to the output of the two-OCR engine by Dannélls et al. (2021b). The new model's improved capabilities manage to further reduce the WER by another 16%.

Table 3 shows the performance of the two-OCR model (Dannélls et al. 2021b), the post-OCR correction model based on byT5 (Xue et al. 2022) by Löfgren & Dannélls (2024), and the updated version trained at KBLab on the annotated newspaper dataset described in Section 4.1 using the same train-test split.

This relatively simple post processing step, can greatly enhance the usefulness of the data making for example search and word frequencies much more reliable. One caveat of this model is that it is trained on errors for older versions of Abbyy FineReader and Tesseract; this means that current versions of OCR engines potentially make new mistakes that are not learned by the post-OCR correction model. To further improve performance, we plan to update the training data (i) with automatic annotations by newer models, (ii) by employing the crowd-sourced annotations by Project Runeberg,[20] (iii) as well as input images generated from text.

20 https://runeberg.org/

4.3 *The National Archives*

The National Archives regularly publishes open datasets on its website[21] for use in different contexts. An especially valuable data set for training HTR models is *Göteborgs poliskammare*. It is the outcome of the crowdsourcing project described by Dick et al. (2024), in which 25 000 pages of 19th century handwritten police records were transcribed with the help of volunteers. The volunteers assisted with both creating an initial training set for an HTR model, and correcting the transcriptions made by the fine-tuned model. In total, this project yielded around 300 000 lines with high-quality transcriptions.

In addition to this data, the AI lab has collected datasets from within the National Archives and in collaboration with other institutions. The AI lab's latest HTR model, called *The Swedish Lion*[22] is trained on over one million lines of historical Swedish text from the years 1600–1900. Unfortunately, the training set cannot be published by the National Archives in its entirety, but roughly half the training set (566 000 lines) is available for download via the data and model sharing platform Huggingface.

Starting in 2025, HTR transcriptions produced by the Swedish Lion will be regularly published on the National Archives' website to be read alongside the original images. The transcriptions will be also available to download as plain text and Alto XML for further analysis.

5 *Research / use scenarios*

By releasing ground-truth data and OCR/HTR models, we seek to support better studies and user applications such as:

- Track linguistic and semantic change over time
- Improve search results in historical documents
- Identify statistical and semantic patterns
- Run analyses in large amounts of text and generate new research questions
- Translate historical texts into modern language for increased accessibility
- Integrate digital documents into chat applications

In the next section we present two use cases that describe how the user can utilize our tools to access text from a digital or analog image or document.

21 https://riksarkivet.se/psidata/

22 Available as an open-source model and as a model within the HTR platform Transkribus.



Figure 6: Search results for *gata* in Kalmar newspaper from 1872-06-15 in KB's online service

6 *Case studies*

After providing all the background information about OCR/HTR technology, the data and models available let us now look at two case studies that describe how the user can work with material that requires digitization, OCR or HTR processing.

6.1 *Case study 1: Identifying place names in newspaper collections*

In this use case, we assume the researcher aims to identify all place names in a set of newspaper collections available through KB's digital archive.[23] The collections have already been scanned, OCR processed, and made accessible online (see also Chapter X in this volume).

23 KB's digitized newspapers archive: https://tidningar.kb.se/

However, if the researcher's interest instead lies in a collection of printed novels that have not yet been digitized, the process begins differently. In such cases, the researcher must contact the digitization services at their university library,[24] requesting the material to be scanned and OCR processed. Most Swedish university libraries rely on the Abbyy FineReader OCR engine for this task. Although the libraries typically run the most recent model, update frequencies vary, meaning that OCR results can differ between institutions. Moreover, the quality of the source material—including print clarity, layout, and preservation state—has a major impact on OCR accuracy.

Once digitized, the next step is to identify place names. Unfortunately, KB's digital archive does not currently support named-entity search. As a workaround, researchers can narrow their search to terms that commonly appear in Swedish street names. For example, in Swedish, many street names end with *gata* 'street'. When we search for this term the search results in one single hit (see Figure 6). Upon closer inspection, however, we find that *gata* actually appears six times on the page, with three occurrences containing OCR errors: *Fiskaregatan, Storgatan, Norra Långgatan* (*OCR: Norra Länggatan*), *Storgatan, Nya Hamngatan* (*OCR: Nya Hanillgatail*), and *Westerlånggatan* (*OCR: Westcrlänggatani*). These discrepancies likely stem from either limitations in the search interface or errors introduced during OCR processing.

To perform a more refined analysis, researchers can use Strix, Språkbanken Text's document-centric platform for text exploration. Strix can be executed via Mink, Språkbanken Text's data processing platform, which allows users to upload their own collections, for example, digitized newspaper data, and run analytical workflows on the data.

There are three main ways to obtain data for analysis:

1. Request digitized files directly from KB, though access to copyrighted materials may be restricted.[25]

2. Download the image files and have them digitized by a university library.

3. Perform local OCR processing using Tesseract. For documents written in Fraktur, we recommend using Språkbanken Text's character model, specifically optimized for historical Swedish print (see Section 7).[26]

24 Many Swedish university libraries offer free digitization services, see for example Gothenburg University Library https://www.ub.gu.se/sv/tjanster-och-stod/digitala-kopior-ur-samlingarna

25 Access to copyrighted material may be restricted by KB.

26 In the future, Tesseract processing will be directly integrated into Mink

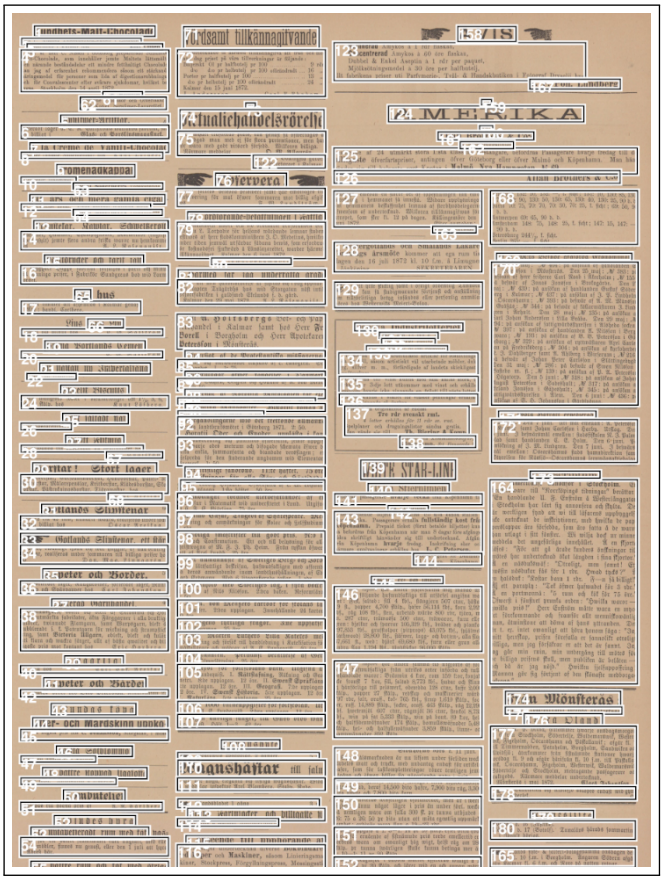


Figure 7: Segmentation results of Kalmar newspaper from 1872-06-15

The OCR process typically produces XML files that describe the document's page layout, including segmentation and annotation data. These files can be visualized to reconstruct how the OCR engine segmented the page, see Figure 7.[27] In this example, segments were numbered to indicate the order in which the OCR engine processed them (Dannélls et al. 2021b). It is important to note that this logical reading order is not always intuitive, and mis-segmentation often contributes to OCR errors.

Alongside XML output, OCR systems also generate plain text files containing recognized text (see Figure 8). Here, one of the text segments from a Kalmar newspaper page dated 1872-06-15 is shown side by side with its OCR

27 The University of Salford (Manchester) offers a set of tools for layout visualization and evaluation https://www.primaresearch.org/tools/PerformanceEvaluation

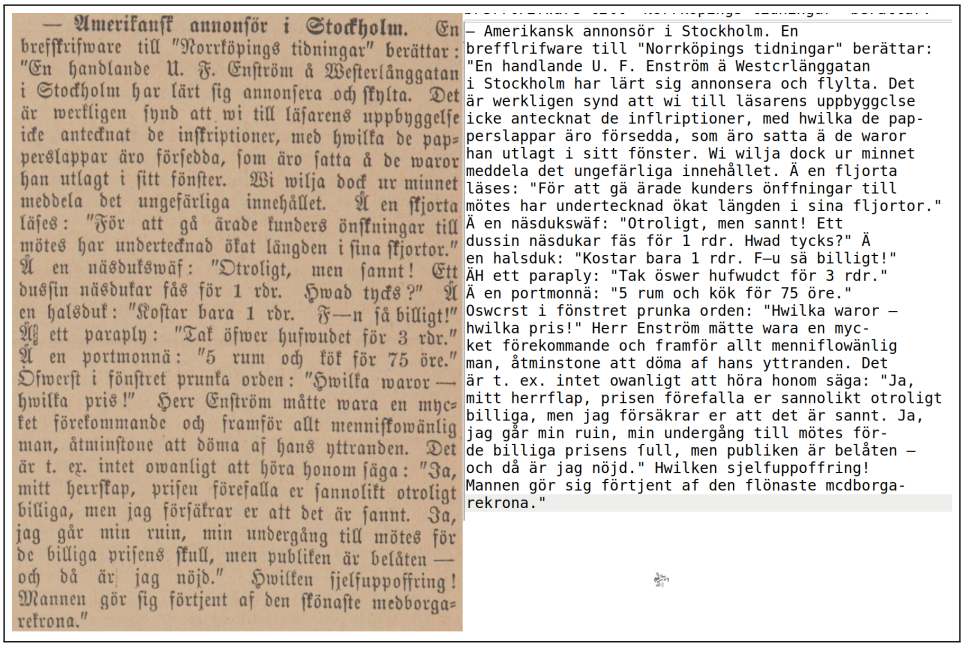


Figure 8: OCR transcription of segment 164 in Kalmar newspaper dated 1872-06-15. The image is shown on the left and the OCR output in plain text appears on the right

transcription. Before performing further searches on the text, it is advisable to identify and visualize OCR errors. To do so, we uploaded the OCR text files into Mink.

After uploading, the text is processed using Språkbanken Text's annotation pipeline, which includes post-OCR correction (Löfgren & Dannélls 2024). Once this step is done, Strix can launched locally to closely examine OCR errors and search through the processed material. Figure 9 shows how the results appear in Strix: on the left, the original OCR text; on the right, the corrected version after post-processing. Words highlighted in yellow mark differences between the two versions. For instance, the post-processing pipeline successfully corrected the street name *Westerlånggatan*, although some OCR errors remain unresolved.

This case study thus demonstrates a complete workflow—from identifying a research question (street names in newspapers), through OCR acquisition and quality control, to post-correction and close reading—using the integrated tools available via Språkbanken Text.

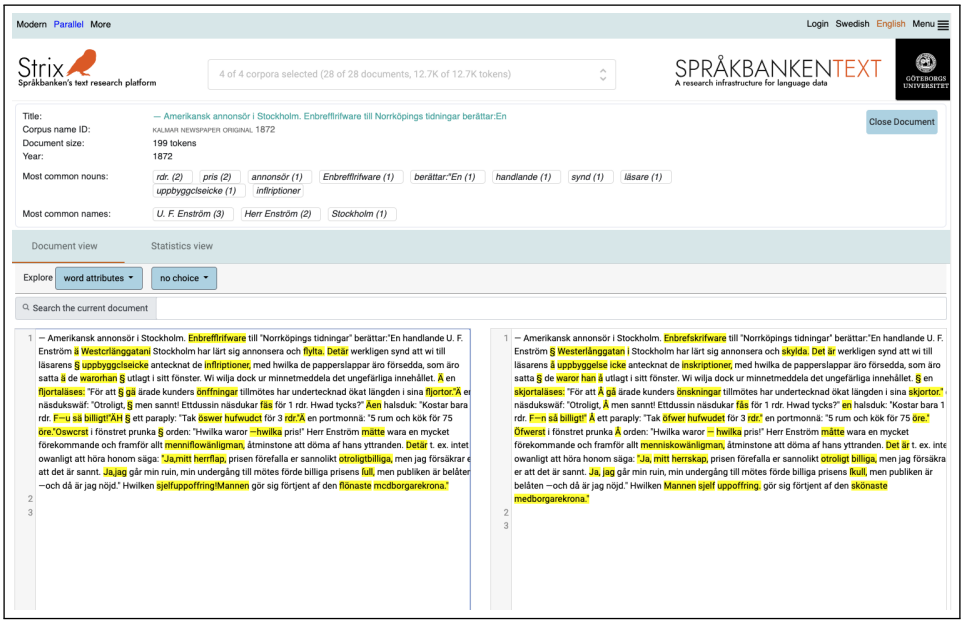
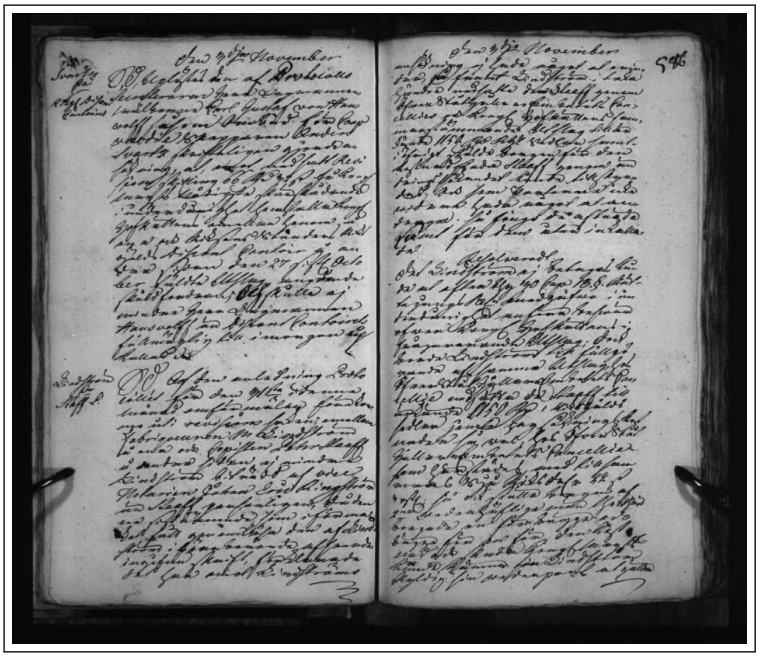


Figure 9: Post-OCR correction of segment 164 in Strix. OCR results before post-processing appear on the left and Post-OCR on the right

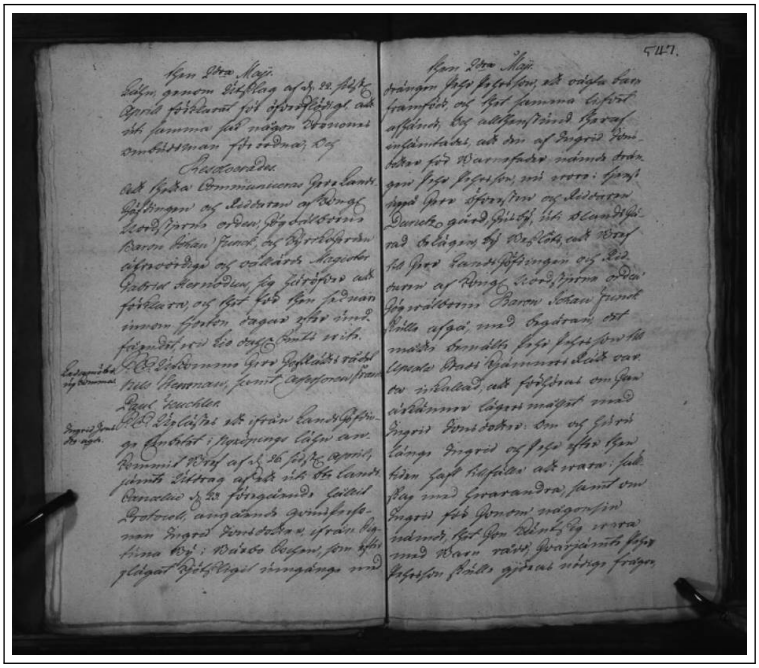
6.2 *Case study 2: Transcribing the archive of Svea Court of Appeal*

As the first step in the National Archives' effort in large-scale HTR, an ongoing project at the Archives aims to automatically transcribe the archive of Svea Court of Appeal (Swedish: *Svea Hovrätt*). The scanned archive contains over a million scanned images of handwritten court records from the 18th and early 19th centuries. Figure 10 shows two typical examples of images from the archive: free running text, in different hands, laid out in two-page spreads with occasional marginalia.

Maybe the first thing that comes to mind is developing the actual HTR model. For this you usually employ some type of neural network. The experience at the Swedish National Archives suggests that transformer architectures operating on a text line level give the best results (Li et al. 2023), even though interesting research exists on doing the text recognition on entire document images instead of text line (Kim et al. 2022, Wei et al. 2024). But these models, even though versions that are pre-trained on large quantities of handwritten text are freely available, do not work well out-of-the-box on historical Swedish handwriting. Thus, the first step, and often the most time-consuming one, is to create the training data for the models.



a.



b.

Figure 10: Examples of scanned documents from the archive of Svea Court of Appeal

Since the HTR model operates on the text line level, the document image must be segmented into lines. In this case, we segment each page into regions and then into lines using instance segmentation models.

When the models are trained and evaluated, they need to be run in an effective way and chained together in a pipeline, and this pipeline should export the results in an appropriate output format for downstream use. If you only have one set of models that should be used on a specific material, then writing this pipeline is pretty straight-forward. But if you want a level of generality that allows you to use any type of models, for instance trained with Hugging Face transformers, mmlabs or native pytorch, and use any kind of segmentation scheme you want, with customized evaluation schemes and several possible output formats, along with customizable preprocessing and postprocessing steps, then you need something like an entire HTR package, written specifically for implementing any type of HTR/OCR project effectively on any type of compute infrastructure. This project, at the Swedish National Archives, is called HTRflow, and is designed to achieve just this generality. The National Archives designed HTRflow with Svea Hovrätt as their use-case, but their aim was to maintain it, improve it, and use it as their inference-engine for all our HTR/OCR projects in the foreseeable future, hopefully aided by the open-source community.

The last stage in any HTR/OCR project is presenting the text to the end users, which in the case of Svea Hovrätt are researchers and the interested general public. For this a team within the IT-department wrote a customization of Universal Viewer 4, which allowed the transcribed text to be visible next to the document images, and by integrating the image viewer with The National Archives's search engine it is also possible to search through the text and get the results highlighted in the actual image.

7 *Practical information about resources, models and tools*

All the resources, models, and tools for Swedish presented here are freely available under the CC BY 4.0 license.

Resources

1. UB Fraktur a selection of 199 pages from 1626–1816 https://spraakbanken.gu.se/en/resources/svensk-fraktur-1626-1816.
2. A selection of digitized versions of Swedish newspapers from 1818 to 1870 https://spraakbanken.gu.se/resurser/svenska-tidningar-1818-1870

3. A selection of digitized versions of Swedish newspapers from 1871 to 1906 https://spraakbanken.gu.se/resurser/svenska-tidningar-1871-1906
4. Open source datasets https://riksarkivet.se/psidata
5. A dataset for training HTR models *Göteborgs poliskammare* https://riksarkivet.se/psidata/goteborgs-poliskammare
6. Ground truth HTR datasets https://huggingface.co/Riksarkivet

Models

1. OCR Post-correction model https://huggingface.co/KBLab/swedish-ocr-correction
   Training code of the model https://github.com/kb-labb/post-ocr-correction/tree/main
2. Swedish language OCR models trained with open source OCR software Calamari https://github.com/mskelb/OCR_SB
3. Trained Tesseract Fraktur model for Swedish http://demo.spraakdata.gu.se/ocr/models/clean_natural_140420-00024000.pyrnn.gz
4. HTR model based on the TrOCR architecture https://huggingface.co/Riksarkivet/trocr-base-handwritten-hist-swe-2

Tools

1. Mink https://spraakbanken.gu.se/mink
2. Strix https://spraakbanken.gu.se/strix
3. HTRflow https://huggingface.co/blog/Gabriel/htrflow

Acknowledgements

The research presented here has been enabled by the Swedish national research infrastructure Nationella språkbanken, funded jointly by the Swedish Research Council (2018–2024, contract 2017-00626) and the 10 participating partner institutions. We would like to acknowledge the Swedish national research infrastructure Huminfra, funded for the years 2022-2024 and 2005-2028, contracts 2021-00176 and 2023-00171 respectively, and the participating partner institutions.

References

Agarwal, Milind & Antonios Anastasopoulos. 2024. A concise survey of OCR for low-resource languages. In Manuel Mager, Abteen Ebrahimi, Shruti Rijhwani, Arturo Oncevay, Luis Chiruzzo, Robert Pugh & Katharina von der Wense (eds.), *Proceedings of the 4th workshop on natural language processing for indigenous languages of the Americas (AmericasNLP 2024)*, 88–102. Mexico City, Mexico: Association for Computational Linguistics. DOI: 10.18653/v1/2024.americasnlp-1.10.

Baron, Alistair & Paul Rayson. 2009. Automatic standardization of texts containing spelling variation: How much training data do you need? In Michaela Mahlberg, Victoria González-Díaz & Catherine Smith (eds.), *Proceedings of the corpus linguistics conference (CL2009)*. University of Liverpool, UK: Lancaster E-Prints UK. https://eprints.lancs.ac.uk/id/eprint/42529/.

Borin, Lars, Gerlof Bouma & Dana Dannélls. 2016. *A free cloud service for OCR / En fri molntjänst för OCR*. Tech. rep. University of Gothenburg, Gothenburg.

Borin, Lars, Markus Forsberg, Martin Hammarstedt, Louise Holmer & Arild Matsson. 2025. Korp: Språkbanken's word research platform. In Kristian Blensenius Dana Dannélls & Lars Borin (eds.), *Sixty years of Swedish computational lexicography*, 175–193. Berlin: De Gruyter.

Börjeson, Love, Chris Haffenden, Martin Malmsten, Fredrik Klingwall, Emma Rende, Robin Kurtz, Faton Rekathati, Hillevi Hägglöf & Justyna Sikora. 2024. Transfiguring the library as digital research infrastructure: Making KBLab at the national library of Sweden. *College & Research Libraries* 85(4).

Brandt Skelbye, Molly & Dana Dannélls. 2021. OCR processing of Swedish historical newspapers using deep hybrid CNN–LSTM networks. In Ruslan Mitkov & Galia Angelova (eds.), *Proceedings of the international conference on recent advances in natural language processing (RANLP)*. Held Online: INCOMA Ltd. https://aclanthology.org/2021.ranlp-1.23.

Dannélls, Dana, Lars Björk, Ove Dirdal & Torsten Johansson. 2021a. A two-OCR engine method for digitized Swedish newspapers. In *Selected papers from the CLARIN annual conference 2020, Linköping electronic conference proceedings 180*. Linköping: Linköping University Electronic Press.

Dannélls, Dana, Lars Björk, Ove Dirdal & Torsten Johansson. 2021b. A two-OCR engine method for digitized Swedish newspapers. In *Selected papers from the CLARIN annual conference 2020, linköping electronic conference proceedings 180*. Linköping: Linköping Electronic Conference Proceedings.

Dannélls, Dana, Lars Björk & Torsten Johansson. 2019. Evaluation and refinement of an enhanced OCR process for mass digitisation. *Digital Humanities in the Nordic and Baltic Countries Publications* 2(1). 112–123. DOI: 10.5617/dhnbpub.11085.

Dick, Kasperowski, Johansson Karl-Magnus & Karsvall Olof. 2024. Temporalities and values in an epistemic culture: Citizen humanities, local knowledge, and AI-supported transcription of archives. *Archives & Manuscripts* 51(2). 3–22. DOI: 10.37683/asa.v51.10937.

Ehrmann, Maud, Matteo Romanello, Antoine Doucet & Simon Clematide. 2022. Introducing the hipe 2022 shared task: Named entity recognition and linking in multilingual historical documents. In Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg & Vinay Setty (eds.), *Advances in information retrieval*, 347–354. Cham: Springer International Publishing.

Forsberg, Markus, Dana Dannélls, Lars Borin & Aleksandrs Berdicevskis. 2025. Background: Språkbanken text. In *Sixty years of Swedish computational lexicography*, 161–173. Berlin: De Gruyter.

Holley, Rose. 2009. How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs. *D-Lib Magazine* 15(3/4). https://www.dlib.org/dlib/march09/holley/03holley.html.

Kim, Geewook, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han & Seunghyun Park. 2022. Ocr-free document understanding transformer. In *Computer vision – eccv 2022: 17th european conference, proceedings, part xxviii*, 498–517. Tel Aviv, Israel: Springer-Verlag. DOI: 10.1007/978-3-031-19815-1_29.

Li, Minghao, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li & Furu Wei. 2023. TrOCR: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the thirty-seventh aaai conference on artificial intelligence and thirty-fifth conference on innovative applications of artificial intelligence and thirteenth symposium on educational advances in artificial intelligence* (AAAI'23/IAAI'23/EAAI'23), 13094–13102. AAAI Press. DOI: 10.1609/aaai.v37i11.26538.

Löfgren, Viktoria & Dana Dannélls. 2024. Post-OCR correction of digitized Swedish newspapers with ByT5. In Yuri Bizzoni, Stefania Degaetano-Ortlieb, Anna Kazantseva & Stan Szpakowicz (eds.), *Proceedings of the 8th joint sighum workshop on computational linguistics for cultural heritage, social sciences, humanities and literature* (*latech-clfl 2024*), 237–242. St. Julians, Malta: Association for Computational Linguistics. https://aclanthology.org/2024.latechclfl-1.23.

Malmsten, Martin, Love Börjeson & Chris Haffenden. 2020. *Playing with words at the National Library of Sweden – making a Swedish BERT*. https://arxiv.org/abs/2007.01658.

Nguyen, Thi, Adam Jatowt, Mickaël Coustaty & Antoine Doucet. 2021. Survey of post-OCR processing approaches. *ACM Computing Surveys* 54. 1–37. DOI: 10.1145/3453476.

Rekathati, Faton. 2021. *The KBLab blog: a multimodal approach to advertisement classification in digitized newspapers*. https://kb-labb.github.io/posts/2021-03-28-ad-classification/.

Rigaud, Christophe, Antoine Doucet, Mickaël Coustaty & Jean-Philippe Moreux. 2019. ICDAR 2019 competition on Post-OCR text correction. In *2019 international conference on document analysis and recognition (ICDAR)*, 1588–1593. DOI: 10.1109/ICDAR.2019.00255.

Sfikas, Giorgos & George Retsinas (eds.). 2024. *Document Analysis Systems - 16th IAPR International Workshop, DAS 2024 Proceedings*. Vol. 14994. (Lecture Notes in Computer Science). Springer. DOI: 10.1007/978-3-031-70442-0.

Språkbanken Text. 2020. *Dalin: Then Swänska Argus 1732-1734*. DOI: 10.23695/9z65-nv18.

Wei, Haoran, Chenglong Liu, Jinyue Chen, Jia Wang, Lingyu Kong, Yanming Xu, Zheng Ge, Liang Zhao, Jianjian Sun, Yuang Peng, Chunrui Han & Xiangyu Zhang. 2024. *General OCR theory: towards OCR-2.0 via a unified end-to-end model*. https://openreview.net/forum?id=3LOcwfB4JX.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest & Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the conference on empirical methods in natural language processing: System demonstrations*, 38–45. Online: Association for Computational Linguistics. DOI: 10.18653/v1/2020.emnlp-demos.6.

Xue, Linting, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts & Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics* 10. 291–306. DOI: 10.1162/tacl_a_00461.

List of abbreviations

CER	Character Error Rate
GT	Ground Truth
HTR	Handwritten Text Recognition
KB	Kungliga biblioteket (National Library of Sweden)
NLP	Natural Language Processing
OCR	Optical Character Recognition
WER	Word Error Rate

Corresponding author

Dana Dannélls
Språkbanken Text
Department of Swedish,
Multilingualism, Language
Technology
University of Gothenburg
dana.dannells@svenska.gu.se

13 Tugg: A transcription tool for language documentation

Magnus Ahltorp
Nakajima Koen Research Institute

Harald Berthelsen
Institute for Language and Folklore

Modern language documentation is less about field linguists treating the speakers of a language as rather passive informants, and more about actively involving them in the work as language consultants. This includes handling tools like recording equipment and being able to transcribe language data on their own. Catering to non-expert users of transcription software presents new challenges for a category of tools that traditionally was the exclusive domain of expert users.

We will describe a new transcription tool that aims to be directly usable without any special knowledge of the tool, and only requiring modest amounts of previous computer knowledge for the end user. The tool is geared towards producing fully Leipzig glossing rules compliant transcriptions with a minimum of effort.

This chapter will describe the tool, how to set up the server part of the tool, and practical instructions on how to use it, including suggested workflows. The chapter ends with a brief description of a practical use case and future directions.

1 *Introduction*

Traditionally, the field linguist has treated the speakers of the described language as fairly passive informants. The research has often consisted of a linguist travelling to an area, collecting data, travelling home, and writing a book. The speakers of the language have often not received anything back and have not really learned anything new about their language or gained skills. To use the terminology of Cameron et al., this was *ethical research*, "research on" (Cameron et al. 1992: 14) subjects.

The field of language documentation instead aims to be more reciprocal (Dwyer 2006: 31) and involve the speakers, normally called consultants (Dwyer 2006: 52). Involving the consultants can mean that they should be

HUM INFRA Magnus Ahltorp & Harald Berthelsen. 2025. Tugg: A transcription tool for language documentation. In Gerlof Bouma, Dana Dannélls, Dimitrios Kokkinakis & Elena Volodina (eds.), *Huminfra handbook: Empowering digital and experimental humanities* (NEALT Proceedings Series 59), 433–453. University of Tartu Library. DOI: 10.58009/aere-perennius0182
© The authors, CC BY 4.0

able to handle technical equipment like recording tools and, after some basic linguistic training, also be able to transcribe the material on their own (Mosel 2006: 74). To once again use the terminology of Cameron et al., language documentation aims to be *empowering research*, "research on, for and with" (Cameron et al. 1992: 22) subjects.

Many tools for transcribing speech data are, however, very complicated to work with and export data from. This chapter describes a language-independent tool for manual transcription of sound files that is directly modelled on the Leipzig glossing rules (Max Planck Institute for Evolutionary Anthropology 2015) format in conjunction with a phonetic/phonematic transcription.

Two popular tools commonly used for transcribing speech are Praat (Boersma & Weenink 2022) and Elan (Max Planck Institute for Psycholinguistics, The Language Archive 2024). Praat has many features for analysing speech, making it suitable for users with a need for deep phonetic analysis. It can be configured to work in many different ways, but this also makes it less suited to users that only need to transcribe speech.

Elan is oriented more towards video transcription, and is often used with many annotation layers. A common use case is creating many overlapping types of annotations, such as what is said, eye gaze, and gestures. Both these tools support arbitrary layers.

Tugg is instead specialised only on transcribing speech, with no advanced acoustic analysis and a fixed layer structure, aiming to get the user up to speed as fast as possible with almost no training in the tool itself.

This chapter is not meant as a substitute for an education in either field linguistics or phonetics, so any advice given should only be seen as giving context to the use of Tugg.

2 *Overview of Tugg*

Tugg presents one sound file at a time on a web page that is scrolled in the vertical up-and-down manner that most users are familiar with (Nielsen 2005, Sherwin 2014). This avoids the unfamiliar and difficult to navigate horizontal timeline that is common in speech manipulation software, and also makes it natural to use the Page up/Page down keys. A horizontal timeline can both make even a relatively short recording seem unmanageably long and make the user often resort to zooming in and out of the material, switching between overview and detail. There are few spatial cues that facilitate navigation, and zooming changes what the timeline looks like, making it harder to associate a position with its content.

To be able to present the sound data in a vertically scrollable canvas, the sound file is automatically divided into chunks, approximately 5 seconds long, that are placed under each other. Tugg tries to divide the sound file at pauses in the speech, but if the user finds more suitable line breaks they can easily be inserted.

Each part of the sound file is represented by a spectrogram. A spectrogram is a visual representation of the sounds in the file, with white representing low intensities and black representing high intensities. The spectrogram shows not only the total intensity at a certain point in time but also how the intensity is distributed in the frequency spectrum. Intensities at the bottom of the spectrogram represent low frequencies, and intensities at the top represent high frequencies. Short periods of low intensity followed by a vertical line often do not indicate pauses in the speech, but are plosives, such as *p t k*. For examples of this, see Section 4.1, Figure 6.

The spectrograms used in Tugg are too small to be used for a detailed phonetic analysis, but help the user navigate and determine the start and end of utterances. Unless the user inserts a line break, the spectrograms will never change, which makes them a reliable spatial cue. The goal here is that even users that are not familiar with spectrograms will find the visual appearance useful as a navigation tool.

Tugg provides three transcription layers: phonetic/phonematic, gloss, and free translation. The purpose of the phonetic/phonematic layer is primarily as support for the glossing, but can be used to research intra- or inter-speaker variation if the researcher so chooses.

The gloss layer is for Leipzig glossing rules style morphosyntactic transcription and morphosyntactic glosses. Finally, the free translation layer is meant to run across the whole utterance.

The intended use of Tugg is that the user annotates one layer at a time by:

- marking a region (see definition in Section 4.1) in the spectrogram with the mouse,
- listening to the region with the apostrophe (') key, and
- entering the transcription.

Each layer behaves differently, according to the needs of that specific layer. Characters in the phonetic/phonematic layer are, for example, spread out over the whole region to maximise the probability that the phone/phoneme will appear close to the sound in the spectrogram. Text in the gloss layer is formatted according to Leipzig glossing rules, with each transcription and gloss token together.

Tugg can produce directly publishable LaTeX code formatted according to Leipzig glossing rules with automatic timestamps for each transcribed utterance.

3 *Setting up Tugg*

3.1 *Downloading*

Start by downloading the latest version of Tugg from https://codeberg.org/nakajimakoen/tugg and unpack it on your computer.

3.2 *Installing Python and necessary packages*

Install Python from your preferred package/software manager or download it from https://www.python.org/downloads/. The oldest Python version supported by Tugg is 3.9.

Tugg requires some Python packages:

- flask
- scipy
- pypng

Install them for instance using the package manager in your Python installation. If you use a unix system, you can probably install the packages you need by running the command:

```
pip install flask scipy pypng
```

3.3 *Configuring*

Place your audio files in the "`audio`" subdirectory of Tugg. All audio files need to be in the WAV audio format and end in "`.wav`". You can also later place additional audio files in this directory at any time.

3.4 *Running the server*

Now start the server "`server.py`". This starts a web server on your computer that only you can access.

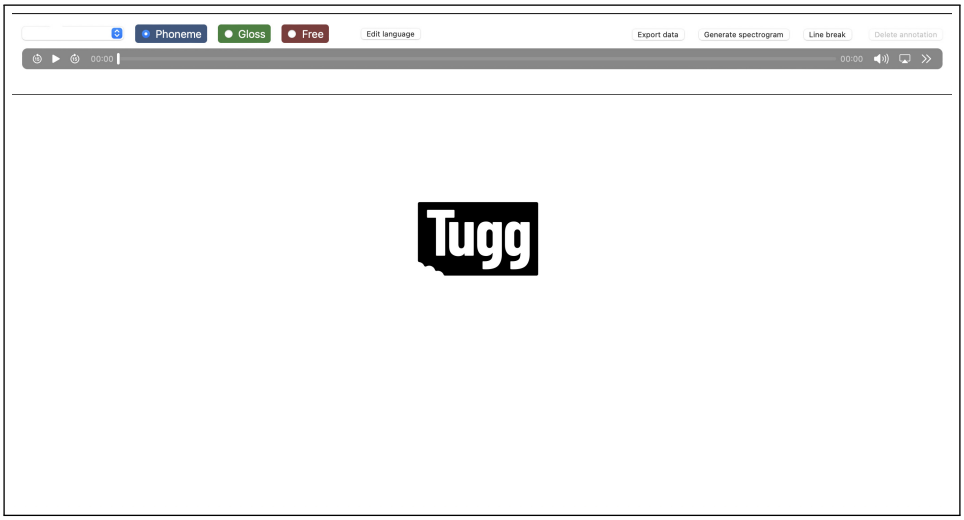


Figure 1: The interface when accessing the page the first time.



Figure 2: The interface after having selected a file to work with.

4 *Using Tugg*

Access the server using a web browser, using the address `http://127.0.0.1:5000`.

You should now see a screen similar to Figure 1. Select the audio file you want to work on in the upper left corner. You should now see the spectrogram of the speech data (Figure 2) with a sound player showing the whole sound

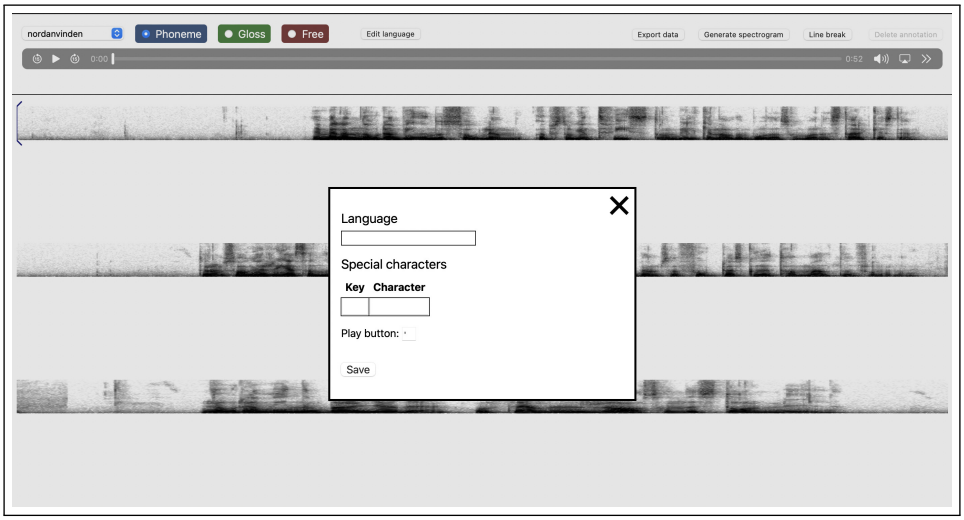


Figure 3: Settings dialog box for language and special characters.

file at the top of the screen. You can navigate either by using the sound player or clicking directly in the spectrogram. The recommended way of playing and pausing is by using the apostrophe (') key on your keyboard. This is to make it as easy as possible to play the contents of the region, without clicking any buttons with the mouse and at the same time leaving letter keys and the space bar available for typing.

If input of the apostrophe (') character is needed frequently enough in the annotation work that this becomes untenable, or the key is not placed conveniently on the keyboard, the key can be changed in the settings dialog box. Playing can also be started and stopped by using the sound player controls or the play/pause button on your keyboard (if you have one).

Tell Tugg what language you are working on by clicking the "Edit language" button at the top of the screen (Figure 3). Enter the language in the way you want it to look in your exported output. You can also assign special characters to keys in this dialog box. In the example in Figure 4, upper-case "D" has been assigned the IPA character for a voiced retroflex plosive, a sound needed to make the phonetic transcription we want. Saving the preferences makes the language name and the special characters appear at the top of the screen (Figure 5).

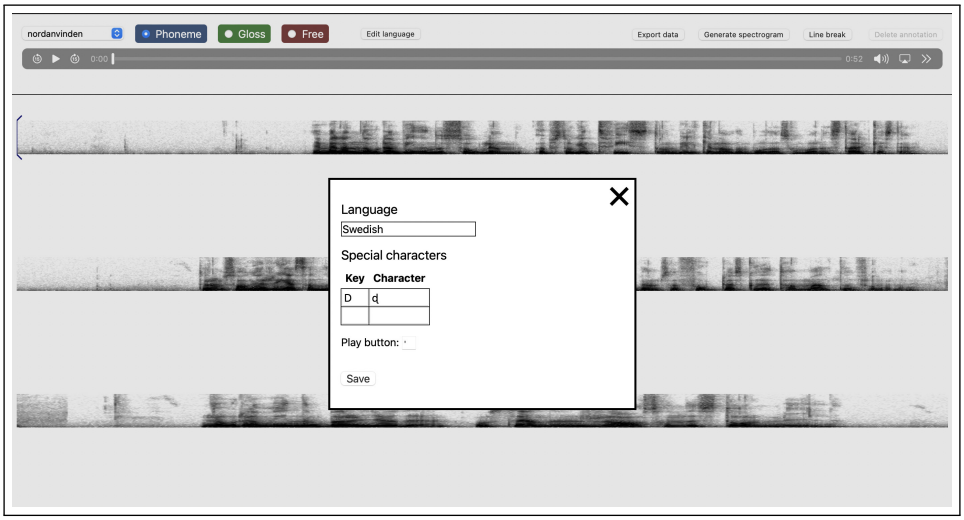


Figure 4: Entering language name and special characters.

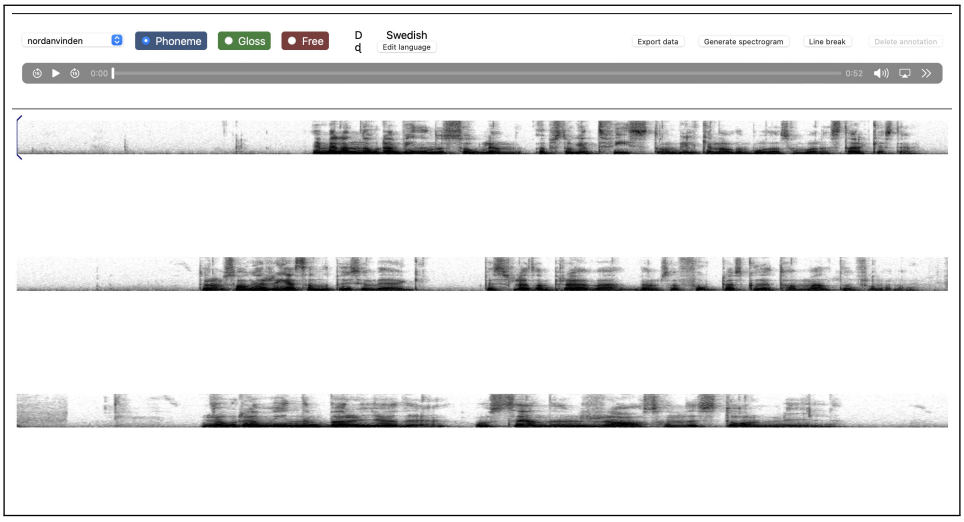


Figure 5: Interface after having saved language name and special characters.

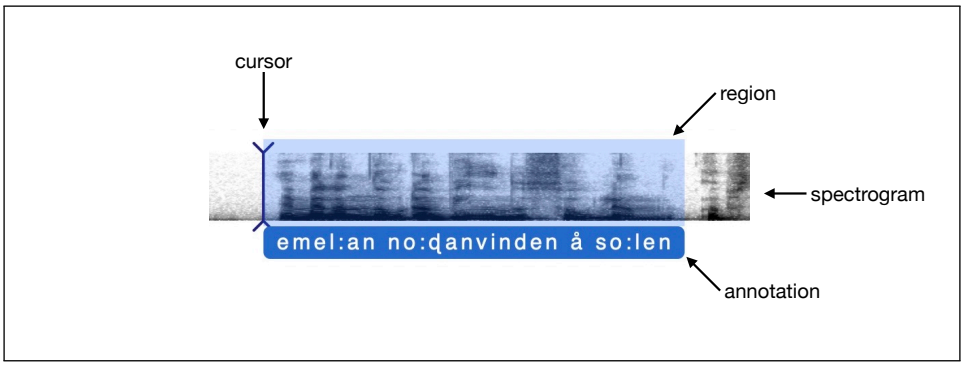


Figure 6: The main interface elements of Tugg.

4.1 *The interface elements*

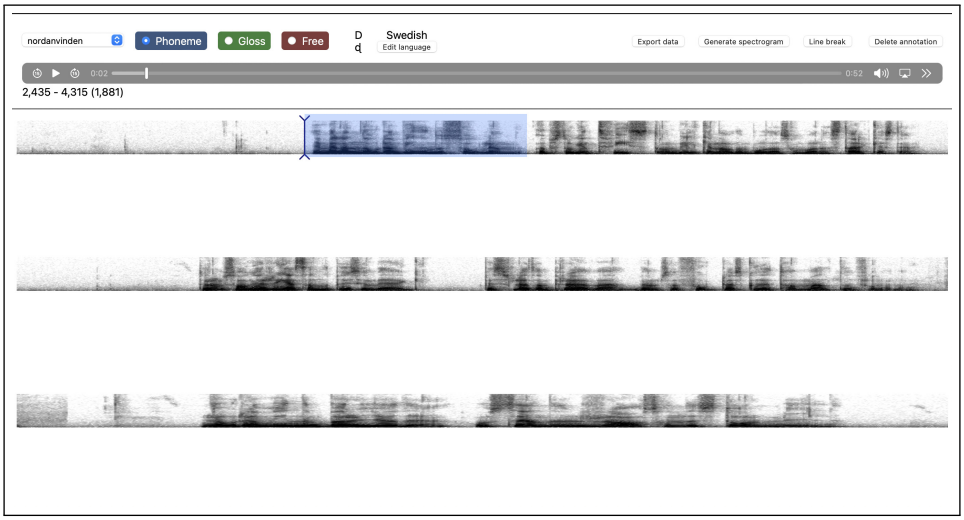
The main interface elements of Tugg are shown in Figure 6:

- the *cursor*: indicating where playback will start or (if playback is in progress) the current position of playback
- the *region*: a part of the recording
- the *spectrogram*: the time/frequency representation of the sound in the recording
- the *annotation*: the text corresponding to a region for a specific layer

4.2 *Marking a region and creating an annotation*

Now we are ready to start annotating. An annotation is created by holding down the mouse button while dragging the mouse cursor from the position in the spectrogram where we want the annotation to start to where we want the annotation to end (see Figure 7). As soon as a region is marked, we can start entering text by typing normally. The appearance of the text will be different depending on the layer type. In the lower part of Figure 7, we can see an example of text that has been entered in the phonetic/phonematic layer. If you want to select a region that already has an annotation, click the annotation.

In the current version of Tugg, a region cannot span two spectrogram rows.



a.



b.

Figure 7: Selecting a region of the spectrogram. In (a), the selected region is marked in light blue and the spectrogram cursor is placed at the start of the region. Screenshot (b) shows the region after entering text.

4.3 *The different layers*

There are three different layers that can be used for annotations. The layer types are shown at the top of the screen in three coloured boxes (Phoneme, Gloss and Free), see for example Figure 7. The layer types are:

- *The phonetic/phonematic layer.* This layer is suited for both broad and narrow transcriptions. There is, however, no specific support for easily adding IPA diacritics, such as voiced/voiceless, advanced/retracted, and raised/lowered which might be needed in certain narrow transcription situations.

  Annotations are limited to one row, and the entered characters will automatically fill the whole region. Text in this layer uses the "special characters" setting mentioned in Section 4.

  This layer is presented in blue, and is always the top-most layer.

- *The glossing layer.* This layer is suited for Leipzig glossing rules annotations. Annotations use two rows, and pressing the return key will change to the second row. Tokens are separated by space, and have to contain the same amount of hyphens (-) in the top and bottom rows.

  This layer is presented in green, and is always the middle layer.

- *The free translation layer.* This layer is suited to the idiomatic translation part of the Leipzig glossing rules. No special formatting is used in this layer.

  This layer is presented in red, and is always the bottom layer.

An example of entered text in all layers can be seen in Figure 8.

4.4 *Playing a selected region*

You play a selected region by pressing the apostrophe (') key on your keyboard. Tugg plays the region and then stops. If you press the key before the region is fully played, playing stops and the cursor is placed at the beginning of the region again. This differs from playing when a region is not selected, where pressing the apostrophe (') key pauses, and pressing it again resumes from that position.

4.5 *Editing annotations*

Annotations can be edited in two ways: changing the boundaries of the region and changing the text. First select the annotation by clicking on it.

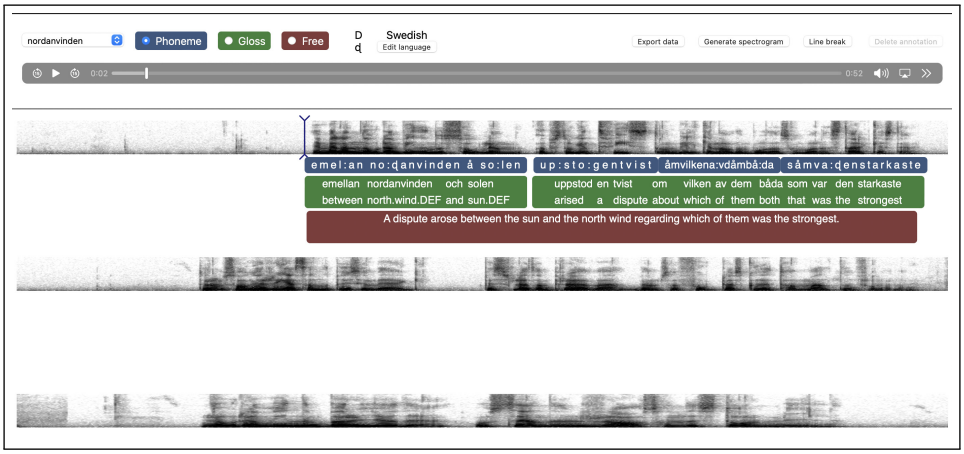


Figure 8: Example of entered text in all layers.

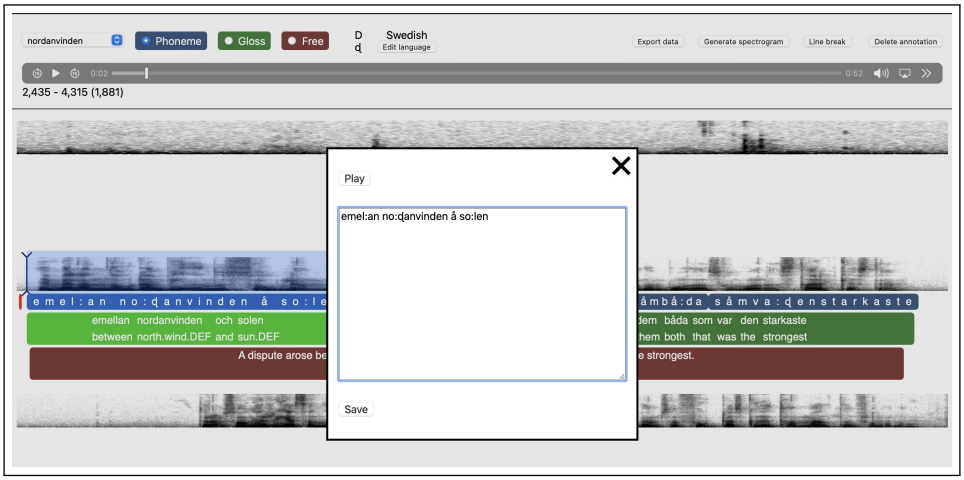


Figure 9: Editing an annotation in more detail. The user has here double-clicked an annotation, and a dialog box containing the text appears, allowing the user to edit the annotation as normal text.

Changing the boundaries is done by dragging the start or end of the region to the desired place.

The text is most commonly changed by just typing on the keyboard. This way, text can easily be entered without any extra steps. However, since changes can only be made to the end of the annotation, either by adding or removing characters, a more complicated but flexible way is also available.

By double-clicking an annotation, an edit dialog box appears (see Figure 9). In this dialog box, all standard edit tools can be used, such as editing in the middle of the annotation, selecting characters, and copy/paste.



Figure 10: Creating a manual line break. The little red line under the spectrogram cursor represents the new line break.

4.6 *Making line breaks*

Tugg tries to place automatic line breaks where there are pauses in the speech data. When this automatic method is not enough, the user can insert manual line breaks. This is done by placing the cursor at the desired position and clicking the button "Line break" at the top of the screen. The line breaks are indicated by a little red line under the spectrogram cursor (see Figure 10).

When the manual line breaks have been placed, click the "Generate spectrogram" button at the top of the screen. All line breaks are now recalculated (see Figure 11).

4.7 *Exporting data*

Data can be exported into different formats. At the moment, three formats are supported directly from the interface: the Tugg format, LaTeX gb4e or covington, and Praat TextGrid.

The Tugg format contains all the annotation and configuration data, and can be imported again into Tugg. Exporting to this format can be used for backup purposes, sending data to another person, or further processing.

Data exported as LaTeX gb4e/covington can be used directly in LaTeX documents using the common glossing packages `gb4e` and `covington`. The only necessary addition to the document is to include the correct package by using the `\usepackage{gb4e}` command for gb4e or `\usepackage{covington}`



Figure 11: Recalculated spectrogram after creating a manual line break. The red line break symbol is now near the beginning of the second row. The row that was previously the first row is now split between the first and second row.

for covington in the beginning of the document. An example of this output and the formatted result can be seen in Figure 12.

Exporting is done by clicking the "Export data" button at the top of the screen. This brings up the export dialog box, where a format can be chosen. After a format has been chosen, the exported data appears in a text field. This can either be copied directly, or downloaded by using the "Download as file" link below the text field (see Figure 13).

5 *Suggested workflows*

Tugg can be used in several different ways. Here we present two different approaches, but there are of course many other workflows possible. The workflow suggestions are meant to be possible to be read independently of each other, so some text will be repeated.

5.1 *Workflow 1: phonematic first*

One possible workflow is the "phonematic first" workflow. This is suitable when the annotator has little knowledge about the language and first has to decode the phonological structure of the language. When using this workflow, start by playing the recording from the beginning and simultaneously watch the spectrogram.

```
\digloss[ex,preamble={Swedish (nordanvinden.wav 2.4--7.9s)}]
    {emellan nordanvinden och solen uppstod en tvist om
     vilken av dem båda som var den starkaste}
    {between north.wind.\textsc{def} and sun.\textsc{def}
     arised a dispute about which of them both that was
     the strongest}
    {A dispute arose between the sun and the north wind
     regarding which of them was the strongest.}
```

a.

(1) Swedish (nordanvinden.wav 2.4–7.9s)

emellan	*nordanvinden*	*och*	*solen*	*uppstod*	*en*	*tvist*	*om*
between	north.wind.DEF	and	sun.DEF	arised	a	dispute	about
vilken	*av*	*dem*	*båda*	*som*	*var*	*den*	*starkaste*
which	of	them	both	that	was	the	strongest

'A dispute arose between the sun and the north wind regarding which of them was the strongest.'

b.

Figure 12: Example output from Tugg (a) and the same output formatted with the LaTeX `covington` package (b).

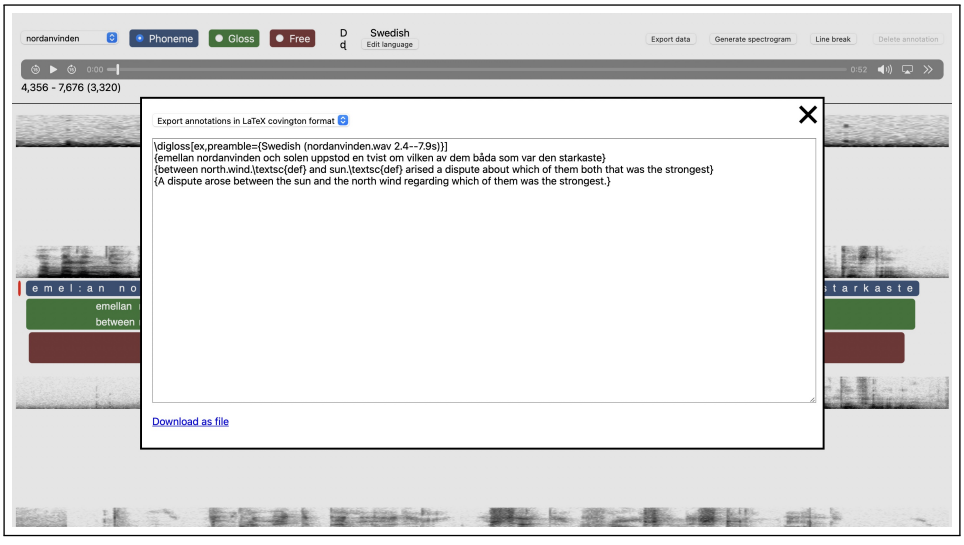


Figure 13: Exporting data from Tugg, in this case in the LaTeX covington format.

When you have located a place that you are interested in transcribing, select the approximate region by clicking and dragging the mouse from the beginning to the end. If the region would cross two lines, first place a line break before the relevant part and press "Generate spectrogram" to move it up to the same row. Try using quite short regions for the phonematic transcription, a couple of seconds.

Play the region by pressing the apostrophe (') key. The sound of the region will now play, and stop at the end of the region. Adjust the boundaries of the region by clicking and dragging the left or right edge and play again until the region matches what you want to transcribe.

Now start typing on the keyboard, alternating between typing and playing. If you make a mistake, you can use the backspace key to erase one character at a time.

If you later discover an error in the middle of the annotation, you can double click the annotation to bring up an edit dialog box where you can make full edits to your annotation.

Everything is saved immediately (except in the edit dialog box), so no additional action is necessary.

If you need to go back and edit a previous annotation, just click the annotation and change either the text or the boundaries.

Depending on how you want to approach the task, you can either annotate everything in the phonetic/phonematic layer first, before you proceed to the glossing stage, or gloss when you have completed one or a few rows of phonetic/phonematic annotations. If you choose the second approach you will have to switch between different layers often, so remember to switch layers before you select a region, since data entered into one layer cannot be moved to another layer without using the edit dialog box.

Annotations in the the glossing layer are meant to be longer than the annotations in the phonetic/phonematic layer, but in the end it depends on the nature of the task and what feels natural for the annotator. Since the annotator has the support of the phonetic/phonematic layer, it is often convenient to use a bit longer annotations. The regions can be played in the same way as before, by pressing the apostrophe (') key. One reason to not have very long annotations in the glossing layer is that edits are made on the whole annotation, starting with the transcription row and only afterwards on the gloss row.

Finally, the free translation layer (called the idiomatic translation in the Leipzig glossing rules) is meant to span a whole utterance. This is what is used as the basis for what counts as an utterance when exporting the glosses to a text document.

5.2 *Workflow 2: translation first*

If you already have easy access to a translation of the utterances, perhaps because you are a native speaker or because you have asked a native speaker for translations as the first step of your analysis, it can be useful to use the "translation first" workflow.

When using this workflow, start by playing the speech from the beginning and simultaneously watch the spectrogram in the same way you would in the "phonematic first" workflow, but listen for utterance boundaries. Before the first utterance, insert a line break by pressing the "Line break" button. Then insert a line break after each utterance, and if there is a long pause, also insert a line break just before the pause is over.

When you have inserted line breaks throughout the recording where needed, regenerate the spectrogram by pressing the "Generate spectrogram" button. Tugg will now take your line breaks into consideration when deciding what is displayed on each row of the spectrogram.

Continue by selecting the "Free" layer type and making a region that encompasses the entire utterance. The length of the free translation annotations is what will be used as the basis for what counts as an utterance when exporting the glosses to a text document.

Play the region by pressing the apostrophe (') key. The sound of the region will now play, and stop at the end of the region. Adjust the boundaries of the region by clicking and dragging the left or right edge and play again until the region matches what you want to transcribe.

Now start typing on the keyboard, filling in what the translation of the whole utterance is. If you want access to full editing tools for your text, simply double click the annotation to bring up an edit dialog box, otherwise the only editing available is pressing the backspace key to erase one character at a time. Everything is saved immediately (except in the edit dialog box, where you have to press the "Save" button), so no additional action is necessary.

If you need to go back and edit a previous annotation, just click the annotation and change either the text or the boundaries.

5.2.1 *Subworkflow 2a: glossing directly from speech*

If the annotator has direct access to the words of the language, for example by being a native speaker, or otherwise knows it well, glossing can start directly from the speech data. Beware that if the orthography you use is not phonological, or you want to investigate phonological processes, a better approach might be to use subworkflow 2b instead.

First switch to the glossing layer type by clicking the "Gloss" button at

the top of the screen. It is often practical to use smaller gloss regions than the whole utterance. This ultimately depends on the nature of the task and what feels natural for the annotator. One reason to avoid very long annotations in the glossing layer is that edits are made on the whole annotation, starting with the transcription row and only thereafter on the gloss row.

In this subworkflow, the phonetic/phonematic layer is only necessary if you need this data for other purposes than Leipzig style glossing. The data will not be used in formatted glossing output.

5.2.2 *Subworkflow 2b: glossing from phonetic/phonematic data*

If the annotator has little knowledge about the language and first has to decode the phonological structure of the language, creating a phonetic/-phonematic annotation before glossing might be a good idea.

Follow the instructions in Section 5.1 except the paragraph about the translation layer.

6 *Practical use case*

One of the authors initially created Tugg as part of a course in field linguistics, and this section will briefly walk through the process of using Tugg in this course.

6.1 *Background*

The language studied was the Indo-Aryan language of Khowar, spoken primarily in northern Pakistan. The language consultant involved in this project grew up in the district of upper Chitral in Pakistan and is a native Khowar speaker. The consultation sessions took place over a teleconferencing platform, with the consultant in Islamabad (Pakistan), recording their own voice and sending the audio files after each session, which ensured good audio quality that did not depend on the quality of the teleconferencing platform or any network problems.

The aim of the study was to investigate the construction of polar (yes/no) and content questions in Khowar. An additional goal was to provide enough clearly sourced transcriptions showing examples of the language data.

6.2 *Session structure*

Interactions during the sessions consisted mainly of elicitations, but also explanation parts, where the documenter asked the consultant for clarifica-

tion regarding certain previous utterances, sometimes by playing back these utterances using Tugg and sometimes by asking follow-up questions.

The elicitations ranged from how to greet people in different situations and common daily interactions to more specific questions on how to reformulate a certain type of utterance that had already been encountered.

6.3 *Transcription*

Some of the transcriptions were made offline, between sessions. This was primarily the phonematic transcriptions, which required a lot of analysis to construct a hypothesis of the phoneme inventory of Khowar. This has been described before, but the described study was purposefully made without any prior knowledge of Khowar.

The glossing, and especially the free translations, were in large parts entered directly during the explanation parts of the sessions. Since all sessions were recorded, there was always the possibility to go back and fill in translations afterwards, but this was made somewhat more difficult by Tugg not having any system for linking audio files. When working on the translation of one file, the spoken translation from the consultant was recorded in another session and thus stored in another file. This necessitated the use of auxiliary text files and manually keeping track of what position in which file contained the explanations/translations.

A mix of workflow 1 and 2 was used. When the language consultant had provided translations before or during the annotation work, workflow 2 was used, otherwise workflow 1 was used. As the amount of recorded speech data was quite large, only a small portion of the data was phonematically annotated. Data that was neither phonematically annotated nor translated was left unannotated. In total, 228 annotations were made over 65 utterances (about 3.5 annotations per utterance, or about 1.2 annotations per utterance and layer). Most utterances were selected for their relevance to the study aim of polar/content question construction. Especially very early in the study, some were instead randomly selected because of the need to investigate the phoneme inventory.

The speed of making annotations with Tugg was the most obvious benefit. In Praat, annotations are made by splitting up the timeline, while in Tugg regions are created with one dragging action with the mouse. Also, a lot of zooming in and out is needed in Praat to switch between overview and detail, something that is not needed in Tugg thanks to the vertical organisation of the annotations.

Using Praat, the examples would have had to be formatted manually to present in the final report. Since all the formatting work was done inside

Tugg, going back and making changes to the analysis did not feel as much of a burden as if every modified example would have had to be reformatted manually.

The vertical organisation also feels more like a traditional document, making it easier to show to the language consultant.

7 *Conclusions and future improvements*

Tugg has now reached the point where it can begin to be useful in real-world contexts. Audio files can be transcribed using three specialised layers, a phonetic/phonematic layer, a glossing layer, and a free translation layer. It uses the more familiar vertical scrolling method for navigation, combining the automatic splitting of lines with the possibility of manual partition. Tugg can also export the transcriptions to a few widely used formats: code for the two LaTeX glossing packages `gb4e` and `covington`, making the transcriptions directly publishable, as well as the Praat TextGrid format for further analysis and processing.

Even though Tugg has been tested in a small user study and used for a few transcription tasks, leading to several improvements, it has not yet been tested on non-linguists. Going forward, testing Tugg on non-linguists should be one of the main priorities, since non-expert usability is one of the goals of the tool.

At the moment, regions and annotations cannot span multiple spectrogram rows. Since this limits the length of utterances that can be annotated, implementing multirow annotations would enable the annotation of longer utterances. It is important, however, to do this while keeping much of the simplicity of the current interface, since that is a key advantage of Tugg.

More export and import options would be useful for users wanting to annotate the same speech data in different applications. Tugg already supports exporting to the Praat TextGrid format, but handling other formats and being able to import the TextGrid format would improve interoperability.

The current version of the edit dialog box shows the raw text without any formatting. A future version could align the words in the upper and lower row, in the same way the normal editing does. Combined with features for controlled vocabulary of abbreviations and lexicon support would greatly improve the glossing process.

For the phonetic/phonematic annotations, a way to align the annotations to the sounds in the spectrogram could be useful for certain applications. A more detailed spectrogram view, in for example the edit dialog box, could enable certain simple analyses without resorting to a more complicated application.

Adding a linking feature, where timestamps in one file can be linked to timestamps in another would make recordings mentioning other recordings much easier to keep track of. Also, adding a notes feature, where the annotator can scribble non-structured metadata about a certain section of the recording, would be very helpful.

Finally, all annotation types would benefit from the ability to merge and split annotations.

We plan to use and evaluate Tugg in other transcription projects. This will lead to more features being incrementally added, but Tugg is already usable in its current state, supporting the basic needs of many transcription tasks.

Acknowledgements

The authors would like to thank Natalia Kalyva and Henrik Liljegren for their valuable input.

The development of Tugg was partially funded by the Swedish Research Council: Swe-CLARIN/The National Language Bank of Sweden (2017-00626).

References

Boersma, Paul & David Weenink. 2022. *Praat: doing phonetics by computer*. http://www.praat.org/.

Cameron, Deborah, Elizabeth Frazer, Penelope Harvey, M. B. H. Rampton & Kay Richardson. 1992. *Researching language*. Routledge.

Dwyer, Arienne M. 2006. Ethics and practicalities of cooperative fieldwork and analysis. In Jost Gippert, Nikolaus P. Himmelmann & Ulrike Mosel (eds.), *Essentials of language documentation*, chap. 2, 31–66. De Gruyter Mouton.

Max Planck Institute for Evolutionary Anthropology. 2015. *Leipzig glossing rules*. https://www.eva.mpg.de/lingua/pdf/Glossing-Rules.pdf.

Max Planck Institute for Psycholinguistics, The Language Archive. 2024. *Elan* (*version 6.8*). Nijmegen. https://archive.mpi.nl/tla/elan.

Mosel, Ulrike. 2006. Fieldwork and community language work. In Jost Gippert, Nikolaus P. Himmelmann & Ulrike Mosel (eds.), *Essentials of language documentation*, chap. 3, 67–86. De Gruyter Mouton.

Nielsen, Jakob. 2005. *Scrolling and scrollbars*. https://www.nngroup.com/articles/scrolling-and-scrollbars/.

Sherwin, Katie. 2014. *Beware horizontal scrolling and mimicking swipe on desktop*. https://www.nngroup.com/articles/horizontal-scrolling/.

Corresponding author

Magnus Ahltorp
Nakajima Koen Research Institute
magnus@nakajimakoen.org

14 Exploring parallel corpora with STUnD: A Search Tool for Universal Dependencies

Arianna Masciolini
University of Gothenburg

Herbert Lange
University of Gothenburg

Márton András Tóth
University of Gothenburg

We introduce STUnD (Search Tool for Universal Dependencies), a corpus search tool designed to facilitate working with parallel data. STUnD employs a query language that allows describing syntactic structures and specifying divergence patterns, which in turn make it possible to look for systematic differences between texts. Furthermore, the tool can automatically detect the differences between two similar documents. To achieve all this, STUnD leverages Universal Dependencies (UD), a cross-lingually consistent standard for morphosyntactic annotation. Input can consist of pre-annotated UD treebanks or raw text, which the tool automatically processes through a third-party parser. As demonstrated in the case study included in the present chapter, STUnD is especially well-suited for comparing syntactic structures across languages, with applications in the context of typology and translation studies. Other use cases include retrieving grammatical errors from parallel learner corpora and comparing different analyses of the same text.

1 *Introduction*

This chapter introduces STUnD (Search Tool for Universal Dependencies), a web-based corpus search tool.[1] Unlike most other such services, STUnD is designed with parallel texts in mind.[2] Its interface makes it possible to inspect and compare two sentence-aligned corpora side by side. Upon running a query, these are processed in parallel, retrieving pairs of semantically equivalent sentence fragments matching the given query. STUnD's

1 A live demo of the tool is available at harisont.github.io/STUnD.
2 The acronym "STUnD" initally stood for *Sökverktyg för Tvåspråkiga Universal Dependencies trädbanker*, Swedish for 'Search Tool for Bilingual Universal Dependencies treebanks'.

HUM INFRA Arianna Masciolini, Herbert Lange & Márton András Tóth. 2025. Exploring parallel corpora with STUnD: A Search Tool for Universal Dependencies. In Gerlof Bouma, Dana Dannélls, Dimitrios Kokkinakis & Elena Volodina (eds.), *Huminfra handbook: Empowering digital and experimental humanities* (NEALT Proceedings Series 59), 455–503. University of Tartu Library. DOI: 10.58009/aere-perennius0183 ©The authors, CC BY 4.0

pattern matching language also supports so-called *divergence patterns*, which let users provide additional corpus-specific details to look for systematic differences between the two input texts. When STUnD is used on a bilingual Swedish-English corpus, for instance, a simple query allows retrieving all Swedish present-tense verb constructions and their English translations, whereas a divergence pattern enables looking for the specific instances where a Swedish *enkelt presens* 'simple present' is rendered as a *present continuous* (*be* + *-ing*) in the English translation.[3] To achieve this, STUnD relies on rich morphosyntactic annotation following the cross-lingually consistent Universal Dependencies (UD) standard (de Marneffe et al. 2021). Input corpora are either pre-annotated UD *treebanks*, i.e. collections of dependency-syntactic trees representing individual sentences, or raw text files that the tool automatically processes through a third-party parser.

Cross-lingual studies, common in typology and translation studies, are STUnD's main area of application. The first prototype of the tool (Masciolini & Tóth 2024) was in fact developed to facilitate the task of investigating the ways in which Swedish verb constructions are rendered in other more or less closely related languages. UD makes it possible to – at least to some extent – identify cross-lingual semantic correspondences between sub-sentence units on the basis of syntactic information alone (Masciolini & Ranta 2021) and allows queries to be largely language-agnostic. Parallel treebanks, however, need not necessarily be multilingual. We therefore propose two more use cases for STUnD: learner corpus research and quality control for linguistic annotation. In the former scenario, we suggest that the tool be applied to parallel *L1-L2 treebanks* (Lee, Li, et al. 2017) – a format increasingly popular within the UD community – for tasks such as grammatical error retrieval and analysis. As for the latter case, we show how the tool can be used to locate and examine discrepancies between different analyses of the same text. Together with STUnD's built-in editing functionalities, this can prove useful for resolving inter-annotator conflicts in the context of a treebanking project, but also play a role in parser evaluation, where system output often needs be compared against gold standard data or alternative automatic analyses.

This chapter is structured as follows. In Section 2, we provide a short introduction to the UD framework, followed by an overview of the treebank tools that most closely relate to STUnD (Section 3). Section 4 consists in a technical description of the tool itself, mostly intended as a practical introduction for new users but also detailing its inner workings. In Section 5, we present a multilingual case study which employs STUnD to explore the

3 Unlike in English, verbs in Swedish only have one present tense form. Without any further context, in fact, it is impossible to tell whether the Swedish sentence *jag lyssnar på musik* means 'I (usually) listen to music' or 'I am (right now) listening to music'.

tense-aspect forms used to render the Swedish *presens perfekt* tense into English and Japanese, while two additional use cases for STUnD are presented in Section 6. Finally, Section 7 is dedicated to some considerations about the further development of the tool and its application in future projects.

2 *Universal Dependencies*

Universal Dependencies (UD) (de Marneffe et al. 2021) is a well-established framework for dependency-based morphosyntactic annotation and a community effort which has resulted in a large collection of treebanks in over 150 languages, with applications in both linguistics and Natural Language Processing (NLP). As the term "Universal" suggests, the UD annotation scheme aims to maximize cross-linguistic consistency without sacrificing accuracy for the analysis of individual languages. This is done by providing a universal inventory of categories (part-of-speech tags, dependency relations and morphological features) and annotation guidelines while also allowing language-specific extensions. An exhaustive description of the UD standard is beyond the scope of this chapter.[4] We will, however, cover some of the basic principles of UD and introduce the text-based data format used to represent treebanks, both of which are necessary prerequisites for understanding the way STUnD works and the examples provided in the rest of this chapter.

2.1 *Basic principles*

UD annotation can be seen as a three-step process consisting of (1) sentence and word *segmentation*, (2) *word-level annotation* and (3) *syntactic annotation*. This section gives an example-based overview of each of these steps and introduces some commonplace UD terminology.

2.1.1 *Segmentation*

UD analysis is carried out at the level of individual sentences, which are further *tokenized*, i.e. segmented into *syntactic words*. Tokenization can be a nontrivial problem, especially for languages with no orthographic word boundaries such as Japanese (see Section 5.2). For the sake of simplicity, however, we will momentarily assume an identity between orthographic (i.e. space-separated) and syntactic words. This largely holds for English and

4 For up-to-date annotation guidelines, we encourage the reader to visit the official UD website, universaldependencies.org.

Swedish, which will be used for all examples in this section, as well as for a wide variety of UD languages. It must be noted, however, that punctuation marks are considered as separate "words". Segmenting the sentence *We are annotating our first sentence.*, for example, results in the following seven tokens:

(1) We are annotating our first sentence .

2.1.2 *Word-level annotation*

Once a sentence has been tokenized, each word is *lemmatized* and assigned a coarse-grained or *Universal* Part-Of-Speech (UPOS) tag.[5] The words that make up our example sentence (1), for instance, can be analyzed as

	PRON	AUX	VERB	PRON	ADJ	NOUN	PUNCT
(2)	We	are	annotating	our	first	sentence	.
	we	be	annotate	we	first	sentence	.

In addition, each word can optionally be assigned a finer-grained, language-specific POS tag and a set of morphological features. A more complete annotation of the word *We*, for instance, would include the fine-grained POS tag `PRP` (personal pronoun) and the features `Case=Nom|Number=Plur|Person=1|PronType=Prs`, describing it as the nominative form of a first-person plural personal pronoun.[6]

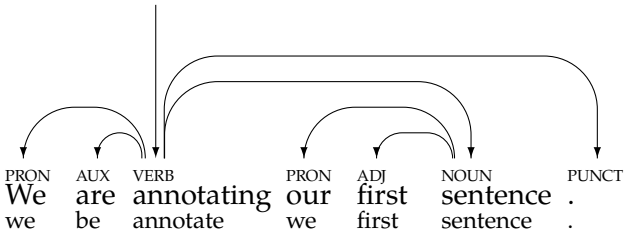
2.1.3 *Syntactic annotation – dependencies*

UD belongs to a class of grammatical theories known as *dependency grammar*, whose core idea is that structural information about a sentence can be captured by directed links, referred to as *dependencies*, between its parts.

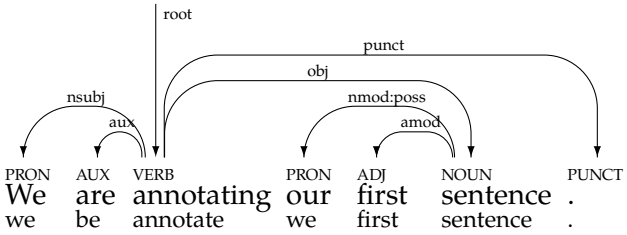
As in several other dependency-based frameworks, syntactic annotation in UD consists in establishing a set of dependency relations between individual words. These are one-to-one correspondences that assign each token a syntactic *head*. Nominals, for instance, are often the syntactic heads of the adjectives that modify them. Conversely, verbs typically have a subject, an object and possibly other complements as *dependents*. As illustrated below, these head-dependent links, represented by arrows, result in a tree structure rooted in one of the words (typically, but not always, the main verb):

5 The complete list of UPOS tags is available at universaldependencies.org/u/pos.

6 A list of all morphological features and their possible values is available at universaldependencies.org/u/feat/all.

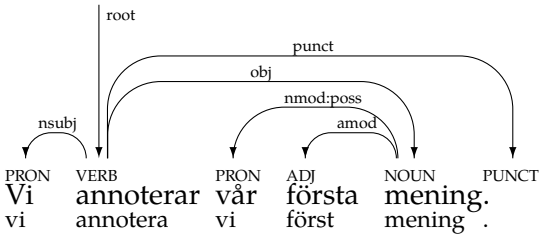
(3)


In UD, each link is also assigned a label that specifies the syntactic role of the dependent with respect to its head, i.e. the *relation type*:

(4)


In example (4) above, for instance, the `nsubj` label between *We* and the root of the tree, the verb *annotating*, specifies that *We* is the nominal subject of the sentence, whereas its object is the subtree labeled as `obj`, rooted in the noun *sentence*. The latter, in turn, has two dependents: the pronoun *our*, acting as a possessive nominal modifier – hence the label `nmod:poss`, where `nmod` is the main *relation type* and `poss` its *subtype* – and the adjectival modifier (`amod`) *first*.[7]

This first example also illustrates one of the core principles behind the UD annotation guidelines, that of *primacy of content words*, which states that dependency relations hold primarily between content words, while function words are usually attached as direct dependents of the most closely related content word. Unlike in most other dependency frameworks, in fact, the `root` of the sentence is not the auxiliary *are*, but the lexical verb *annotating*. The rationale behind this is to maximize cross-lingual parallelism. A straightforward Swedish translation of our example sentence, *Vi annoterar vår första mening.*, in fact, makes no use of auxiliaries but presents the exact same high-level structure:

(5)


7 For a complete taxonomy of UD relation labels, see universaldependencies.org/u/dep.

2.2 *Data format*

UD treebanks are distributed in a machine-readable text-based format called CoNLL-U.[8] A CoNLL-U file consists of a series of blocks, each representing a sentence. Sentence blocks are separated from each other by a blank line and can be preceded by metadata information, stored in lines starting with a hash sign (`#`). Metadata usually includes a sentence identifier (sent_id) and the unannotated text (`text`).

The words that make up a sentence are stored *vertically*, i.e. with one token and the annotations associated with it per line. Each line consists of a series of tab-separated values:

1. ID: progressive token index;
2. FORM: word form or punctuation symbol;
3. LEMMA: lemma or stem of the word form;
4. UPOS: universal POS tag;
5. XPOS: language-specific POS/morphological tag (optional);
6. FEATS: morphological features, stored as a list of key-value pairs in the format `Key1=Value1|Key2=Value2...` (optional);
7. HEAD: ID of the syntactic head of the current word (`0` for the root of the sentence);
8. DEPREL: relation type;
9. DEPS: enhanced dependency graph, stored as a list of head-deprel pairs (optional);
10. MISC: any other annotation, stored as a list of key-value pairs (optional).

Underscores (_) are used as placeholders for missing values. The CoNNL-U representation of the example sentence in (1–4) is given in Figure 1. For the sake of readability, we also provide a tabular representation of the same sentence (Figure 2).

3 *Treebank tools*

The success of UD as an annotation standard has led to a growing demand for tools and query languages for treebanks. In this context, a variety of

8 The CoNLL-U format is based on the format that was used in the context of the shared task on multilingual dependency parsing at the Tenth Conference on Computational Natural Language Learning (CoNLL-X).

```
# sent_id = ex1_en
# text = We are annotating our first sentence.
1     We          we          PRON   PRP    Case=Nom..   3     nsubj _     _
2     are         be          AUX    VBP    Mood=Ind...  3     aux   _     _
3     annotating  annotate     VERB   VBG    Tense=Pres... 0    root  _     _
4     our         we          PRON   PRP$   Case=Gen...   6     nmod:poss   _     _
5     first       first       ADJ    JJ     Degree=Pos... 6     amod  _     _
6     sentence    sentence    NOUN   NN     Number=Sing  3     obj   _     _
7     .           .           PUNCT  .      _            3     punct _     _
```

Figure 1: Textual representation of the example sentence in (1–4) in CoNLL-U format. For the sake of compactness, we omit some of the morphological features, as well as the DEPS and MISC fields.

ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL	DEPS	MISC
1	We	we	PRON	PRP	Case=Nom...	3	nsubj	_	_
2	are	be	AUX	VBP	Mood=Ind...	3	aux	_	_
3	annotating	annotate	VERB	VBG	Tense=Pres...	0	root	_	_
4	our	we	PRON	PRP$	Case=Gen...	6	nmod:poss	_	_
5	first	first	ADJ	JJ	Degree=Pos...	6	amod	_	_
6	sentence	sentence	NOUN	NN	Number=Sing	3	obj	_	_
7	.	.	PUNCT	.	_	3	punct	_	_

Figure 2: Tabular representation of the CoNNL-U fragment shown in Figure 1, representing the example sentence in (1–4). For the sake of compactness, we omit some of the morphological features, as well as the DEPS and MISC fields.

services such as PML-TQ (Prague Markup Language – Tree Query, Pajas & Štěpánek 2009), SETS (Scalable and Efficient Tree Search, Luotolahti et al. 2015) and Grew-match (Guillaume 2021) were extended and/or developed from scratch to work with CoNLL-U files.

What many of these tools have in common is that, as opposed to classic string- or sequence-based corpus search tools, they allow searching for syntactic constructions based on the structural information present in CoNLL-U files (cf. Section 2.1). In Grew syntax, for instance, the query

```
pattern {
    X [VerbForm = Part];
    Y [lemma = "have", Tense = Pres];
    X -> Y;
}
```

can be used to search for occurrences of the English *present perfect* (*have* + past participle). To do that, it specifies that results must contain at least two tokens `X` and `Y`, respectively a participle and a present form of the verb *to have*, and that the former must be a dependent of the latter.[9] STUnD uses a

9 This query can be tested on official UD treebanks at `universal.grew.fr`.

different query language, largely based on the one used in the gf-ud toolset for dependency trees and interlingual syntax (Kolachina & Ranta 2016, Ranta & Kolachina 2017), but is more or less equivalent to Grew-match in terms of expressivity of the queries, at least when used on a single treebank.

When it comes to parallel data, however, the number of available options is much more limited. Some tools, such as Grew-match itself, allow displaying hits alongside the corresponding sentence in a parallel treebank based on sentence-level metadata. However, none of the existing query engines aside from STUnD's command-line predecessor, L2-UD (Masciolini 2023), allow performing parallel searches such as those discussed in the introduction.

This is not to say that STUnD is the only tool with a focus on treebank comparison. The SETS-based treebank analysis tool STARK (Krsnik et al. 2019), for example, allows extracting comparative statistics and dependency patterns from pairs of UD treebanks (not necessarily parallel). The comparisons enabled by STARK, however, are treebank-level rather than sentence-level. Some annotation platforms such as Arborator-Grew (Guibon et al. 2020), on the other hand, offer the possibility to compare alternative analyses of the same text by highlighting annotation differences between individual words. To the best of our knowledge, however, none of the existing tools allow sentence-level comparisons between a text and its translation or two otherwise different versions of a given text.

4 *Tool description*

In the following, we provide a technical description of STUnD, aimed for both new users and programmers interested in contributing to the tool's further development or in reusing some of its components. Implementation details intended for the latter group are enclosed in grey text boxes that can be safely skipped by readers whose main focus is learning how to use the finished product.[10] We start with a presentation of the tool's user interface (Section 4.1) and input formats (4.2), followed by a discussion of the underlying query engine and language (4.3). Section 4.4 introduces some additional features.

4.1 *User interface*

STUnD is a web-based treebank search application consisting of a Haskell backend and a simple HTML+JavaScript frontend. The software is open

10 A step-by-step tutorial is also available at harisont.github.io/STUnD/tutorial.

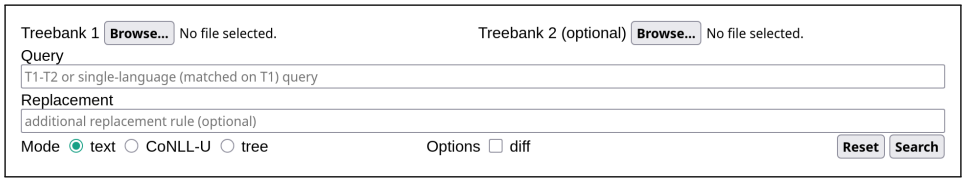


Figure 3: STUnD's input fields.

source[11] and easy to run locally.[12] In addition, a live demo of the tool is available at harisont.github.io/STUnD.

Upon launching STUnD, the user is presented with a series of input fields (cf. Figure 3). Two file selectors ("Treebank 1" and "Treebank 2") allow choosing one or two input files, henceforth T_1 and T_2. These can be UD treebanks in CoNLL-U format or plain text files (cf. Section 4.2). Through the two text input fields, "Query" and "Replacement", the user can specify a search and a replacement pattern, both discussed in depth in Section 4.3.[13] Two buttons, "Search" and "Reset" are used to run the query and clear the input fields respectively.

Before running a query, the user can choose in what "Mode" to get the search results. In "text" mode (default, cf. Figure 4), hits are presented in a format similar to KeyWords In Context (KWIC) or concordances, i.e. highlighted in bold in the context of the full sentences in which they appear. Results can be downloaded as plain text or, when a parallel treebank is searched, TSV (Tab-Separated Values) by clicking on "parallel file". This is intended for further processing, for instance in a spreadsheet program.

In "CoNLL-U" mode (cf. Figure 5), results can be inspected in their underlying structured format. Unlike text mode, CoNLL-U mode isolates the sentence fragments matching the query. Results consist of well-formed UD trees , which can be downloaded by clicking on "T1 file" and "T2 file" and fed back into STUnD for finer-grained corpus analysis. Alternatively, they can be stored as a single treebank – where T_1 and T_2 sentences are interleaved – by clicking on "parallel file".

Finally, query-matching segments can be rendered as graphical UD trees (cf. Figure 6). Exporting is similar to CoNLL-U mode, but the resulting files consist of a series of SVG (Scalable Vector Graphics) fragments, easy to embed in a webpage or directly visualize in a web browser.

11 The source code can be found at github.com/harisont/STUnD.
12 See installation instructions at harisont.github.io/STUnD/installation.
13 Note that these fields are both optional: by leaving them empty, the user can inspect the full treebank.

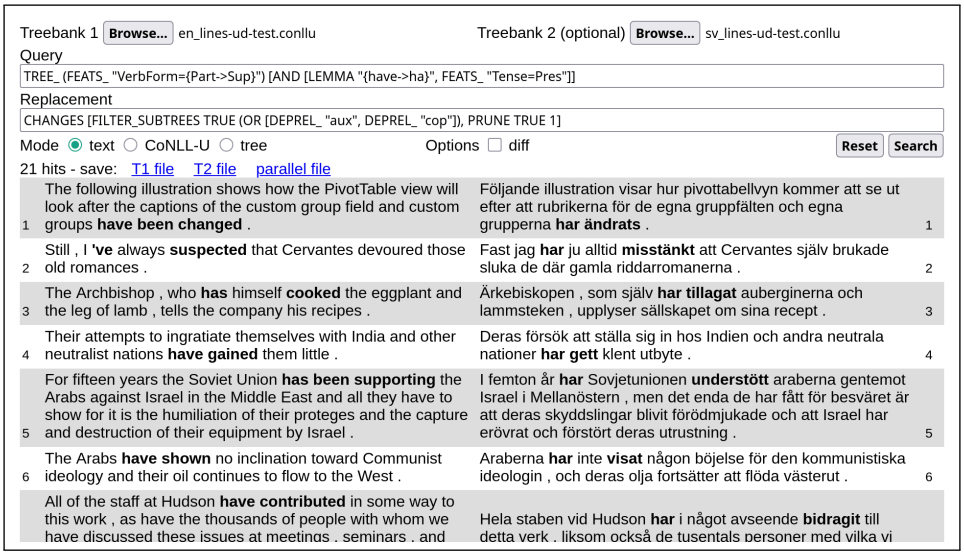


Figure 4: Using STUnD to look for sentences containing occurrences of literally translated present perfect in the English-Swedish parallel treebank LinES (Ahrenberg 2007, 2015) in text mode. A replacement pattern is used to isolate the relevant verb constructions.

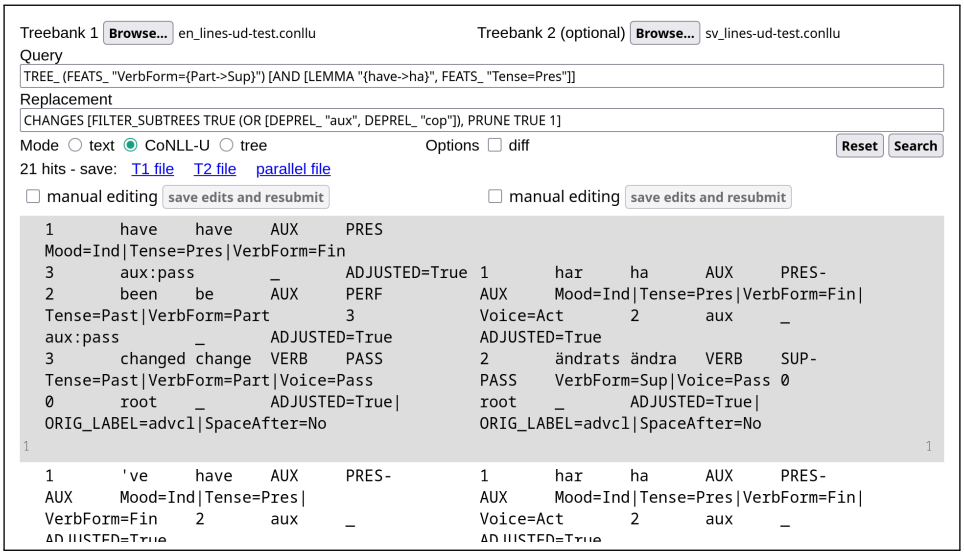


Figure 5: Search results for the same query as in Figure 4 in CoNLL-U mode.

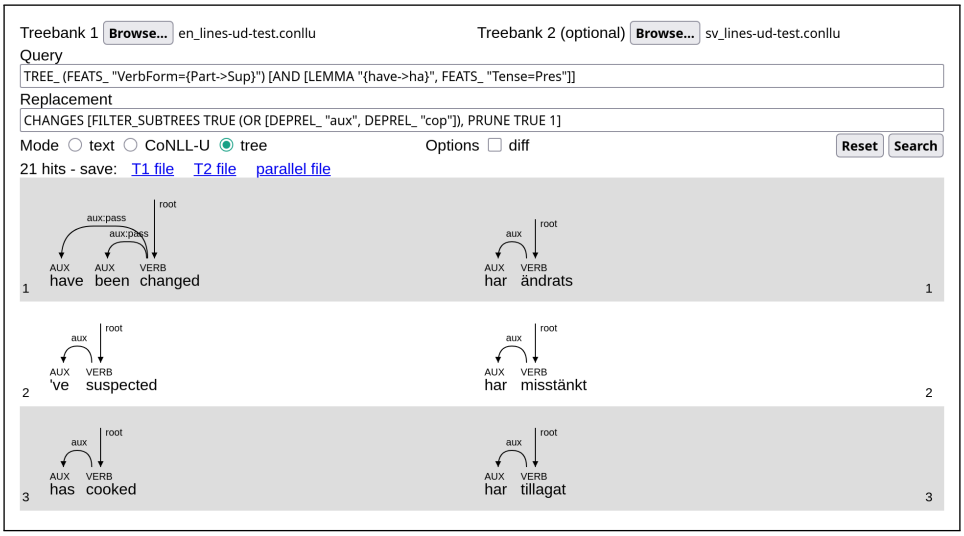


Figure 6: Search results for the same query as in Figure 4 in tree mode.

4.2 *Input formats*

STUnD's first prototype only accepted CoNLL-U input. The number of available pre-annotated parallel UD treebanks is, however, rather limited. Therefore, users often need to automatically process other texts with a dependency parser. To make our tool more accessible, we opted for integrating parsing into STUnD itself, which now also supports raw text input.

If, upon clicking on "Search", one or both input files consist of raw text, the first step is determining their language. If the file name starts with a two letter ISO 639-1 language code[14] followed by an underscore (e.g. `sv_` for Swedish), the text is assumed to be in the corresponding language. Otherwise, the language of the text is automatically inferred.

Once the language has been determined, after asking the user for consent, the text is sent to the UDPipe API,[15] where it is parsed with the default model for the language in question. If parsing is successful, the given query is run directly on the output. Afterwards, however, any parse errors can be manually corrected with STUnD's manual editing functionality (cf. Section 4.4.3).

14 See `loc.gov/standards/iso639-2/php/English_list.php` for the full list of codes.
15 `lindat.mff.cuni.cz/services/udpipe/api-reference.php`

Under the hood: Language identification Language identification is based on `langid.js`,[a] a JavaScript implementation of `langid.py` (Lui & Baldwin 2012) which uses a multinomial Naïve Bayes classifier over byte n-grams (with $1 \leq n \leq 4$). The model has been trained on data for 97 languages and 5 domains: government documents, software documentation, newswire, Wikipedia, and internet crawls. It has been shown that this method achieves high accuracy across domains (Lui & Baldwin 2011).

a github.com/saffsd/langid.js

4.3 *STUnD queries*

This section delves into the core of STUnD: the query engine. We start by discussing how the tool can be used to explore individual treebanks, presenting the basic query language and pattern matching algorithm (Section 4.3.1). We then move on to querying parallel treebanks, discussed in Section 4.3.2. Finally, Section 4.3.3 introduces replacement patterns, which allow applying systematic changes to the search results so as to refine them.

4.3.1 *Simple queries*

STUnD queries are based on the pattern matching functionalities of the gf-ud toolset for dependency trees and interlingual syntax (Kolachina & Ranta 2016, Ranta & Kolachina 2017),[16] whose UD-related utilities are now maintained as a separate software library, deptreehs.[17] The latter provides a Haskell-embedded pattern matching language with a concise syntax especially well-suited for describing syntactic (tree) structures.

In the following, the aim is to give the reader a good understanding of the fundamental constructs of the query language so as to be able to follow the examples presented in this chapter and get started with writing their own queries. Certain details are therefore omitted. The full specification of the query language is available as part of the deptreehs documentation.[18]

Basic patterns The most basic STUnD patterns describe the contents of a single CoNLL-U field. They consist of the name of the field itself followed by a string of text, delimited by double quotes, that specifies its desired value. The currently supported fields are FORM, LEMMA, UPOS, XPOS, FEATS, DEPREL and MISC. When a query consists of a single basic pattern,

16 Source code available at github.com/GrammaticalFramework/gf-ud.
17 Source code available at github.com/harisont/deptreehs.
18 github.com/harisont/deptreehs/tree/main/docs

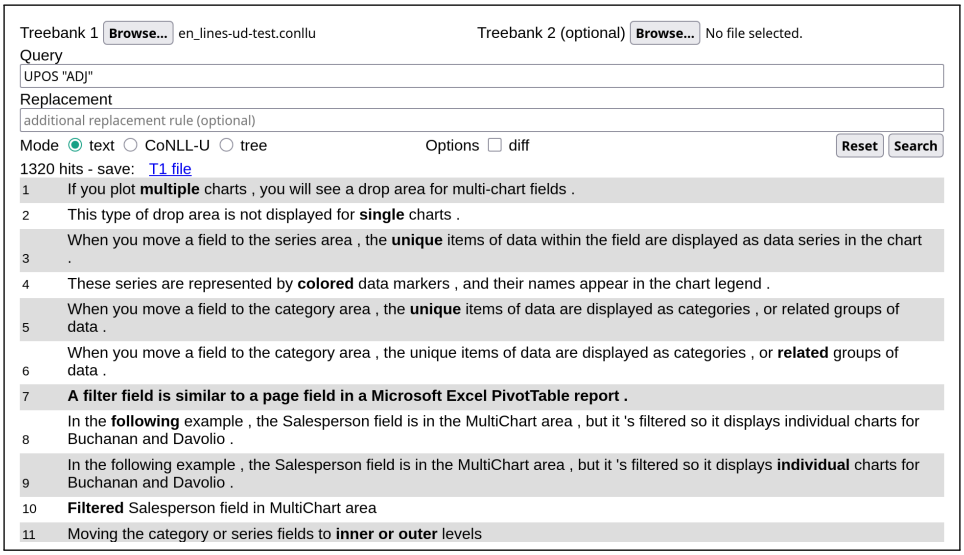


Figure 7: Search results for the query `UPOS "ADJ"` on the English half of the LinES treebank. Most hits consist of single tokens, but when adjectives have a modifier or a conjunct, such as in hit number 11, the entire subtree is retrieved. In one case, the adjective *similar*, used as a copula complement, is the root of the sentence, which is therefore considered a match in its entirety (cf. hit number 7).

such as `UPOS "ADJ"`, STUnD will retrieve all dependency subtrees rooted in an adjective encountered in the corpus, regardless of their "depth" in their parent tree (cf. Figure 7).

Strings used in a basic pattern can start or end with the wildcard `*`, matching any substring. The pattern `DEPREL "*comp"`, for example, matches both classes of UD clausal complements (`ccomp` and `xcomp`). In addition, some basic patterns have a "relaxed" version, characterized by an underscore following the CoNLL-U field name. The pattern `DEPREL_ "aux"`, for instance, ignores relation subtypes, matching subtrees labelled both `aux` and `aux:pass`, whereas its stricter counterpart `DEPREL "aux"` only matches auxiliaries in non-passive constructions.

Compound patterns Basic patterns can be combined by means of *logical operators*, into *sequences* and/or into *tree patterns*.

The query language provides three logical operators: `AND`, `OR` and `NOT`. For example, the pattern

```
AND [OR [UPOS "VERB", UPOS "AUX"], NOT (DEPREL "cop")]
```

matches all (sub)trees rooted in a verb or auxiliary, except those used as a

copula. As displayed in the example above, the `AND` and `OR` operators are followed by a comma-separated list of patterns delimited by square brackets, while the `NOT` operator is followed by a single pattern, enclosed in parentheses.[19] Additionally, the pattern `TRUE` can be used to match any subtree.

Similar to `AND` and `OR` expressions, sequences of patterns can be built by prepending the keyword `SEQUENCE` to a list of patterns. The pattern

```
SEQUENCE [DEPREL "nmod:poss", UPOS "NOUN"]
```

for instance, matches all possessives immediately followed by a noun. If the sequence need not be contiguous, a `SEQUENCE_` pattern, where this constraint is relaxed, can be used instead.

Last but not least, tree patterns, used to search for dependency-syntactic structures, are the real strength of the query language. The form of a tree pattern is `TREE p [p1, p2, p3...]`, where `p` is a pattern to be matched by the head of the subtree, while `p1`, `p2`, `p3`... are to be matched by its dependents in the order in which they appear in the query, without any other intervening subtrees. For example, the pattern

```
TREE (UPOS "NOUN") [DEPREL "nmod:poss"]
```

appears similar to the previous example, but there are some important differences between the two. First of all, the tree pattern specifies that there has to be a head-dependent relation between the noun and its modifier, i.e. a link that connects the former to the latter. Therefore, search results with this query will include nouns that have a possessive modifier attached to them, independent of the order in which the two elements appear and even if there are other intervening words. This pattern, however, excludes all trees that have both a determiner and some other token among their dependents.

A more relaxed – and more commonly used – version of this pattern, `TREE_`, allows for additional dependents at any point in the sequence. Replacing `TREE` with `TREE_` in the pattern above therefore produces a query that matches all noun-possessive pairs in a head-dependent relation, including those with other modifiers. The running example from Section 2.1.3, *We are annotating **our** first **sentence**.* – whose dependency analysis is given in (4) – would only be matched by this second version of the tree pattern, since the word *first* occurs between the possessive pronoun and the noun and is also a dependent of the latter.

Finally, the slightly more complex pattern

```
TREE_ (FEATS_ "VerbForm=Part") [AND [LEMMA "have", FEATS_
    "Tense=Pres"]]
```

19 Note that square brackets are used to delimit *lists* of patterns, whereas round brackets surround single patterns whenever they occur as part of larger compound ones, similar to parentheses in a mathematical expression.

matches all participles with a present-tense inflection of the verb *to have* among its dependents, i.e. all English present perfect constructions.[20] Its non-underscored version, on the other hand, would only retrieve instances where *have*/*has* is the only dependent of the participle.[21]

Under the hood: Basic query matching From a technical standpoint, when a query is run on a treebank T, STUnD goes through all sentences in the treebank and visits each tree t and all of its subtrees recursively, looking for any subtree t' matching the given pattern. Formally, search results always consist of a set of subtrees (possibly several per sentence), with two edge cases: (1) single tokens, i.e. "leaves" and (2) complete sentences, i.e. full dependency trees. This means that, in all cases excepts the latter, results may not be well-formed CoNLL-U trees with an explicit root and progressive token IDs starting from 1. STUnD, however, rescales the ID and HEAD fields and overwrites the dependency label of the head of each extracted subtree with `root`, storing its original value in the MISC field. This is to ensure that search results can be further processed without any post-editing.

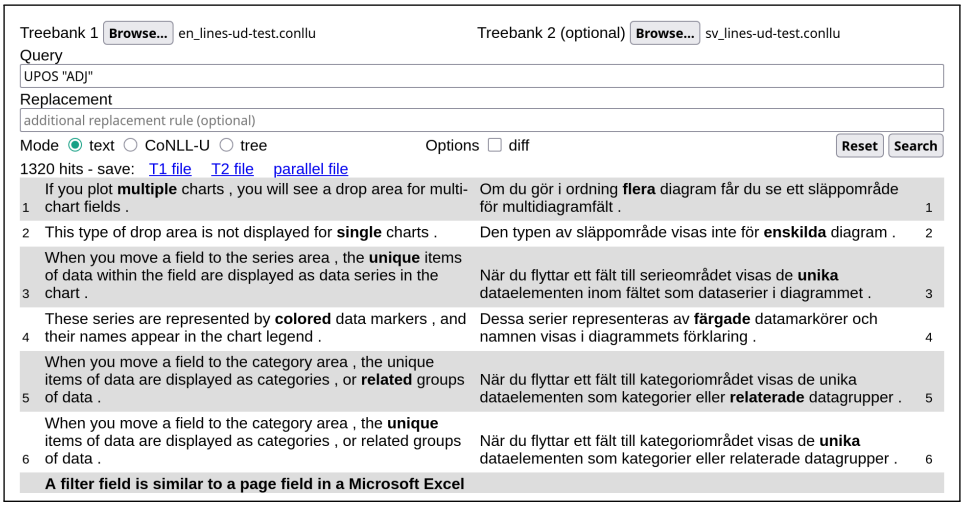
4.3.2 *Parallel queries*

Up to this point, we have discussed how simple queries can be applied to a single UD treebank. This is similar to what can be achieved with all treebank search tools mentioned in Section 3. STUnD, however, was designed with the goal of facilitating parallel treebank exploration, which will be the focus of this section. We start by discussing what happens when a query is run on two sentence-aligned treebanks and then move on to query language extensions specifically meant for parallel search.

When a query is run on a parallel corpus, STUnD matches it on the left-hand-side treebank (T_1) as described in Section 4.3.1, but also attempts to find a semantic correspondence in the right-hand-side treebank (T_2). If, as in Figure 8, the user looks for adjectives, hits will consist of subtree pairs consisting of an element of T_1 that matches the query directly (i.e. is rooted in an adjective) and its counterpart in T_2. The latter does not necessarily need to match the query: in hit number 7, for instance, the entire sentence is rooted in the adjective *similar*, used as a copula complement for the auxiliary *is*, whereas the root of the Swedish translation is the verb *liknar* 'resemble'.

20 This pattern is equivalent to the Grew-match query shown in Section 3.

21 This is never the case in standard English because the obligatory subject of a present perfect-tense clause is also always a dependent of its head, which is the participle itself.



Treebank 1 Browse... en_lines-ud-test.conllu Treebank 2 (optional) Browse... sv_lines-ud-test.conllu

Query

UPOS "ADJ"

Replacement

additional replacement rule (optional)

Mode ◉ text ○ CoNLL-U ○ tree Options ☐ diff Reset Search

1320 hits - save: T1 file T2 file parallel file

1	If you plot **multiple** charts , you will see a drop area for multi-chart fields .	Om du gör i ordning **flera** diagram får du se ett släppområde för multidiagramfält .	1
2	This type of drop area is not displayed for **single** charts .	Den typen av släppområde visas inte för **enskilda** diagram .	2
3	When you move a field to the series area , the **unique** items of data within the field are displayed as data series in the chart .	När du flyttar ett fält till serieområdet visas de **unika** dataelementen inom fältet som dataserier i diagrammet .	3
4	These series are represented by **colored** data markers , and their names appear in the chart legend .	Dessa serier representeras av **färgade** datamarkörer och namnen visas i diagrammets förklaring .	4
5	When you move a field to the category area , the unique items of data are displayed as categories , or **related** groups of data .	När du flyttar ett fält till kategoriområdet visas de unika dataelementen som kategorier eller **relaterade** datagrupper .	5
6	When you move a field to the category area , the **unique** items of data are displayed as categories , or related groups of data .	När du flyttar ett fält till kategoriområdet visas de **unika** dataelementen som kategorier eller relaterade datagrupper .	6
	A filter field is similar to a page field in a Microsoft Excel		

Figure 8: Search results for the query `UPOS "ADJ"` on the parallel English-Swedish treebank LinES.

For finding such semantic correspondences, STUnD relies on the concept-alignment Haskell library[22] (Masciolini & Ranta 2021), which uses a set of customizable rules to determine which sentence fragments correspond to each other in a parallel treebank. This software package was originally designed for automatically bootstrapping translation lexica from multilingual corpora, but can also be applied, sometimes with even better results, to other types of parallel data (cf. Section 6).

Under the hood: Subtree alignment Technically, what concept-alignment does is comparing two sentence-aligned UD treebanks and recursively extracting pairs of semantically corresponding subtrees, i.e. words or phrases, based on a set of customizable *alignment criteria*. In the case of STUnD, these are heuristics based on UD's cross-lingual abstractions, such as comparing dependency labels, UPOS tags of content words and – for same-language parallel treebanks – lemmas. It follows that the quality of the alignments varies depending on the syntactic similarity between the two treebanks and on the quality of the UD annotation. For a manually annotated corpus of Swedish to Norwegian translations, for instance, concept-alignment is quite reliable. At the opposite extreme, results obtained from an automatically parsed corpus of texts in two typologically unrelated languages can be less satisfactory, especially if tokenization conventions differ. For this reason, it is important for matches to also be available in the context of the full sentences they appear in.

22 github.com/harisont/concept-alignment.

Under the hood: Matching queries on parallel treebanks The parallel query matching algorithm, first introduced in Masciolini (2023) is as follows:

1. the alignment module extracts phrase- and word-level alignments from the input parallel treebank $\langle T_1, T_2 \rangle$. The result of this step can be seen as a larger, intermediate parallel treebank $\langle T'_1, T'_2 \rangle$ where each element is a pair of aligned UD subtrees retaining information about the "parent sentence" of each subtree
2. T_1 is recursively scanned for subtrees matching the query, exactly as discussed in Section 4.3.1
3. each matching subtree t_1 is looked up in T'_1. At this point, there are two possibilities:
   - if it is found, the hit will consist of the subtree alignment $\langle t'_1, t'_2 \rangle$, where $t_1 = t'_1$
   - if there is no such subtree pair in $\langle T'_1, T'_2 \rangle$, it means that the alignment step has failed to find a counterpart for t_1 in T_2[a] and t_1 is returned alongside a "dummy" empty UD tree. STUnD's UI, however, still makes it possible – at least in text mode – to compare the $\langle t_1, t_2 \rangle$ full sentence pair the match occurs in.

a This is sometimes due to the limitations of the alignment module, but also common and expected when the query looks for something eminently language-specific, such as certain function words. Looking for definite determiners in an English-Swedish treebank, for instance, typically results in a lot of partial matches, since definiteness is mostly expressed morphologically in Swedish.

Divergence patterns We have seen how basic queries alone can be a useful tool for searching parallel treebanks. In a number of cases, however, users might want to specify what exactly to look for in each of the two halves of a parallel treebank. When studying translation divergences, for instance, it may be useful to retrieve all instances where a direct object in the source text is rendered as an oblique in its translation (cf. Figure 9) rather than looking for all English objects and their Swedish counterparts only to filter the results manually.

In STUnD, this is possible through so-called *divergence patterns*, i.e. queries containing a T_1 and a T_2-specific portion. A divergence pattern essentially consists of two basic queries to be matched in parallel, one on T_1 and one on T_2. Looking for objects translated as obliques, for instance, means looking for subtree alignments where t'_1 is labelled as `obj` and t'_2 as `obl`. For the sake of conciseness, however, the basic query language has been extended with an arrow syntax which allows specifying the treebank-specific portions of

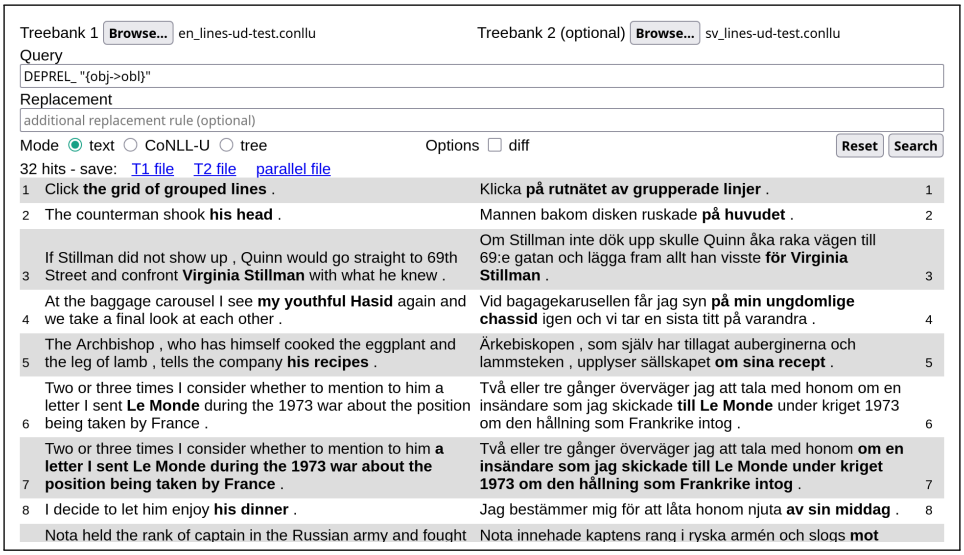


Figure 9: STUnD used to search for translation divergences in the parallel English-Swedish treebank LinES.

the query without rewriting the entire pattern twice. The query

```
DEPREL_ "{obj -> obl}"
```

for example, reads as "find all T_1 subtrees whose head is a direct object and whose counterpart in T_2 is a subtree rooted in an oblique".

Under the hood: Matching divergence patterns The matching algorithm for divergence patterns is exactly the same used when running a basic query on a parallel treebank, except that the arrow notation is expanded into two separate queries $\langle q_1, q_2 \rangle$ (in this simple example, q_1: `DEPREL_ "obj"` and q_1: `DEPREL_ "obl"` and that a sentence pair $\langle t_1, t_2 \rangle$ alignment is only considered a match if, given one of its $\langle t'_1, t'_2 \rangle$ alignments, t'_1 matches q_1 and t'_2 matches q_2.

4.3.3 *Replacement patterns*

As mentioned previously, STUnD offers replacement patterns as a way to refine search results. More broadly, replacement patterns can be used to apply arbitrary changes to a treebank. This can be useful in settings other than treebank querying, e.g. for fixing systematic annotation errors or adapting an existing treebank to updated annotation guidelines that affect specific linguistic phenomena.



Figure 10: STUnD used to search for occurrences of the Swedish *presens perfekt*, with and without using a suitable replacement pattern. On the right hand side, the replacement is used to isolate the verb construction.

The replacement patterns most widely used in STUnD, however, are the ones that allow filtering or pruning the trees that the user obtains as a result of running a query, thus isolating their most salient parts.

```
FILTER_SUBTREES (UPOS "VERB") (UPOS "AUX")
```

for instance, can be used to remove all dependents of verbs other than their auxiliaries. The query

```
PRUNE TRUE 1
```

on the other hand, decreases the depth of all subtrees to 1, thus reducing the search results to the subtrees' heads and their immediate dependents.

It is also possible to combine several changes into a single replacement pattern. There are two ways to do this. Prepending the keyword `CHANGES` to a list of replacements `[X,Y,Z]` goes through all subtrees looking for possibilities to perform each operation and applies the first applicable one. With `COMPOSE [X,Y,Z]`, on the other hand, the edit operation `X` is performed first wherever possible, `Y` is applied to the resulting treebank, `Z` on the results of `Y` and so on. For example,

```
COMPOSE [FILTER_SUBTREES (UPOS "VERB") (UPOS "AUX"), PRUNE TRUE 1]
```

combines the two patterns discussed above making it so that pruning happens only after irrelevant subtrees have been filtered out (for a real-world example of application of this pattern, see Figure 10).

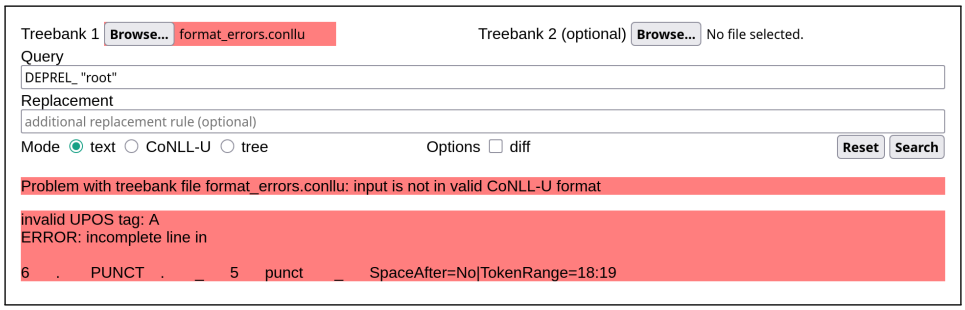


Figure 11: Trying to load an invalid CoNNL-U file into STUnD. The results are two error messages, one signaling that `A` is not a valid Universal POS tag, one indicating that one of the lines does not contain all expected values.

4.4 *Other functionalities*

Aside from treebank querying, STUnD has a set of features geared towards treebank annotators: treebank validation, diff mode and manual editing.

4.4.1 *Validation*

In the examples above, STUnD was used to explore officially released treebanks, the gold standard for UD annotation. Treebanks under development, however, often present format errors that can hinder or completely prevent processing with STUnD, leading to issues such as failures to render dependency trees and misleading query results. For this reason, STUnD performs some basic validation of the input data, checking that all word lines contain 10 tab-separated fields, as well as that all token IDs, UPOS tags and DEPRELs are valid as per the UD universal annotation guidelines. If the treebank contains one or more format errors, these are listed in the user interface (cf. Figure 11). This feature has proved to be useful when grading UD annotation assignments and can help new treebanks in the initial phases of their development, but it should not replace the official UD validator.[23]

4.4.2 *Diff mode*

Diff mode helps users identify discrepancies between two similar treebanks. This feature can be useful when exploring parallel L1-L2 treebanks, as well as when comparing alternative analyses of the same text. Both use cases will be

23 The official UD validator can be downloaded at github.com/UniversalDependencies/tools/blob/master/validate.py and is automatically run whenever an official UD repository is added or updated.

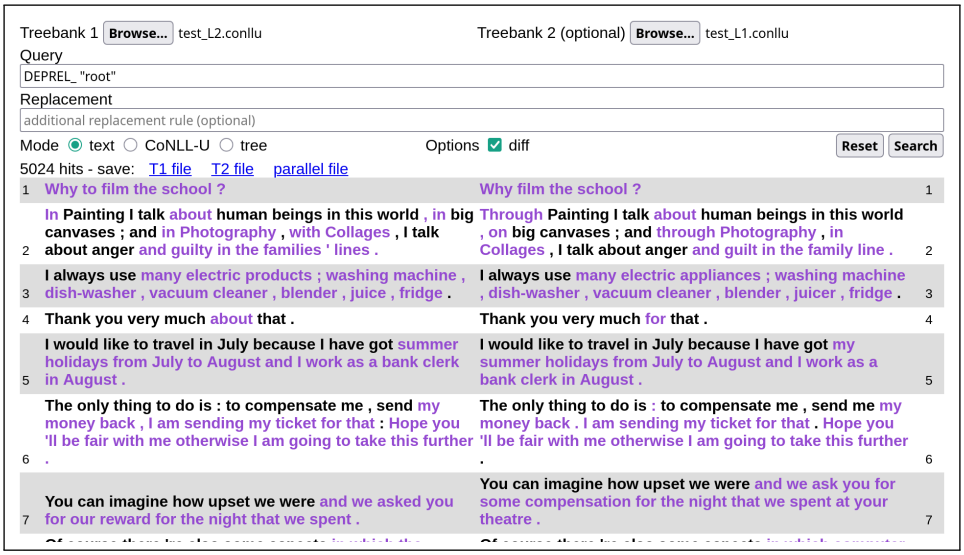


Figure 12: STUnD to browse the test set of the English as a Second Language (ESL) treebank (Berzak et al. 2016) with diff mode enabled. Treebank 1 consists of original learner productions, while Treebank 2 contains the corresponding corrections. Therefore, the highlighting helps identifying grammatical errors.

discussed in Section 6. With diff mode enabled, pairs of aligned subtrees that diverge in terms of surface word forms and/or annotation are highlighted in purple, as illustrated in Figure 12. Comparing subtrees rather than individual tokens yields visualizations that are sometimes coarser-grained compared to those available in text-based diff tools and annotation platforms such as Arborator-Grew (Guibon et al. 2020), but more flexible: it is in fact possible to use STUnD's diff mode on any two variants of a given text, potentially even a text and its translation to a different language, provided that the definition of "divergence" is adequately relaxed.[24]

4.4.3 *Manual editing*

In Section 4.3.3, we discussed how replacement patterns can be used not only to refine search results, but also to apply arbitrary changes to the treebank(s). This, however, is often not the most practical way to edit individual trees. The user can therefore also edit the treebank manually. This is currently supported in CoNLL-U mode only, with plans to add a table view for less

24 STUnD's UI does not currently make it possible to configure the tool to obtain meaningful diffs when comparing different languages, but the underlying code is easy to modify in this sense.

error-prone editing. Compared to other CoNLL-U editors such as Arborator-Grew (Guibon et al. 2020), in any case, STUnD's editing functionalities are basic. Editing in STUnD should therefore be seen as a quick way to apply the final touches to a treebank that is undergoing annotation rather than as the main annotation tool for a treebanking project in its early stages.

5 *Case study: cross-linguistic comparison*

In the following, we demonstrate the use of STUnD in a case study. The case study explores the usage of the Swedish and English present perfect tense-aspect forms and its Japanese equivalents. While Japanese has no tense-aspect form directly corresponding to the present perfect form of Swedish and English, we chose the form *-teiru* in Japanese as a nearly corresponding form, as *-teiru* has reportedly similar uses to the present perfect of English (Shirai 2000, Tóth 2022). For the sake of simplicity, we will use the abbreviation *PresPf* for the present perfect when referring to both the present perfect in English and its counterpart *presens perfekt* in Swedish. We use language-specific namings for the categories of Swedish and Japanese, in order to avoid confusion with English categories. The overarching purpose of the case study is to show how the STUnD tool can be used to make cross-linguistic observations on grammatical structures, more specifically on tense-aspect systems.

Languages have various grammatical forms to express time and by investigating the distribution of the forms cross-linguistically we can draw conclusions on the similarities and differences between tense-aspect systems of different languages. The PresPf is interesting in this sense since it shows great cross-linguistic variation (De Swart 2016). For instance, the use of the PresPf in Swedish interrogative clauses such as *Har du sovit gott?* is not unusual, while the preterite *Did you sleep well?* would be preferable in English in a similar situation (Larsson & Lyngfelt 2011). Thus, the questions are to what extent the use of the PresPf agrees between two or more languages, and what forms a language which has no designated PresPf form – such as Japanese – uses in the same situations.

5.1 *The tense-aspect systems of English, Swedish and Japanese*

Swedish and English have a similar division of tense forms, shown in Table 1, where we exemplify the tense forms with the verb *to run* in English and *springa* 'run' in Swedish. The Swedish tense forms correspond approximately to those of English, although these tense forms are used slightly differently in the two languages.[25]

Table 1: Tense forms in English and Swedish.

English tense forms	Swedish tense forms
Simple present: *runs*	Enkelt presens: *springer*
Preterite: *ran*	Preteritum: *sprang*
Present perfect: *has run*	Presens perfekt: *har sprungit*
Pluperfect: *had run*	Pluskvamperfekt: *hade sprungit*
will + infinitival verb: *will run*	ska + infinitival verb: *ska springa*
would + infinitival verb: *would run*	skulle + infinitival verb: *skulle springa*
(is) going to + infintival verb: *is going to run*	kommer att + infinitival verb: *kommer att springa*
(was) going to + infinitival verb: *was going to run*	kom att + infinitival verb: *kom att springa*

One point of difference between the two languages is that English has a form for progressive aspect, namely the *be* + *ing*-construction, which can be combined with the PresPf, as in *I have been eating* (cf. *I have eaten*; both forms can be translated to Swedish with the PresPf *jag har ätit*). Swedish has a repertoire of constructions for expressing progressive aspect (e.g. *hålla på att äta*, lit. 'hold on to eat') but it has no inflectional form for progressive aspect. The PresPf of Swedish consists of the auxiliary *ha* in present tense 'have' and the supine form[26] of the main verb, e.g. *har ätit* 'has eaten', and refers to a past event with some sort of relevance at present time (Teleman et al. 1999). A sentence such as *Jag har ätit min macka* 'I have eaten my sandwich' indicates that as a result of eating the sandwich in the past, the sandwich 'is' eaten at present. In Swedish, the auxiliary *ha* can be omitted in subordinate clauses e.g. *Jag ser att du redan* (*har*) *ätit* 'I see that you (have) already eaten', meaning that the supine form appears without an auxiliary. The Swedish

25 For instance, in cases where English would use the future tenses *will run* or *going to run*, the preferred counterpart in Swedish is often *enkelt presens* (the simple present): *springer* 'run(s)'.

26 The supine form (e.g. *brutit* 'broken') of Swedish, used in the present perfect, is distinct from the past participle (e.g. *brutet* 'broken').

Table 2: Tense-aspect forms in Japanese.

	Non-past	Past
Non-continuous	*-ru*	*-ta*
Continuous	*-tei-ru*	*-tei-ta*

PresPf corresponds to the English present perfect, which consists of the auxiliary *have* in present tense and the past participle, e.g. *have eaten*.

Japanese has four basic tense-aspects forms (Kudou 1995), as summarized in Table 2. The four forms are inflections and thus directly attach to verbs. All four forms convey both temporal (past/non-past) and aspectual (continuous/non-continuous) information. Non-past forms are generally used for events occurring at the present moment or in the future, while past forms are generally used for events occurring in the past. The continuous forms often express progressive aspect – however, see (7) below. The four forms are exemplified below.[27]

(6) a. *Ashita wa sushi wo tabe-ru*
Tomorrow TOP sushi ACC eat-NONPAST
'I will eat sushi tomorrow'

b. *Kinou wa sushi wo tabe-ta*
Yesterday TOP sushi ACC eat-PAST
'I ate sushi yesterday'

c. *Ima, sushi wo tabe-tei-ru*
Now sushi TOP eat-ASP-NONPAST
'I'm eating sushi right now'

d. *Sono toki, sushi wo tabe-tei-ta*
That time, sushi ACC eat-ASP-PAST
'At that time, I was eating sushi'

However, all four forms interact with lexical aspect and the aspectual interpretation of a form might depend on the verb used. For instance, the *-teiru* form tends to indicate stativity with telic verbs (verbs with an inner boundary), as in (7) below.

(7) *Terebi ga tsui-te-iru*
TV NOM turn.on-ASP-NONPAST
'The TV is turned on'

27 The list of abbreviations used in the glosses can be found in at the end of this chapter.

The default interpretation of (7) is that as a result of the TV being turned on, it 'is' turned on at the present moment. Furthermore, *-teiru* can also indicate that a past event has some sort of connection to the present, as in (8 a–b).[28]

(8) a. *Satou wa, 10 nen mae kara kono kaisha de hatarai-tei-ru*
Satou TOP 10 year before since this company LOC work-ASP-NONPAST
'Satou has been working at this company since ten years back'

b. *Mado, ake-tei-ru yo*
window open-ASP-NONPAST PART
'I've opened the window'

In (8 a), the event of 'working' has held from ten years back up to present time. In (8 b), the event of opening the window has a result at present, namely that the window is open.

Japanese has forms which are used for coupling sentences, similarly to conjunctions such as *and* in English. These forms are the verb stem (e.g. the verb *hashi-ru* 'to run' becomes *hashi-ri*) and the *-te*-form (e.g. *hashi-ru* → *hashi-tte*), as in *Kawa made hashi-tte/hashi-ri, ie ni kaetta* '(I) ran to the river, and returned home'. A written language variety of *-teiru*, namely *-teoru*, can also be inflected for the verb stem *-teori* when coupling sentences.

Morphologically speaking, the continuous forms in (6 c–d) consist of the *-te*-form followed by the auxiliary *i-ru*, which can be conjugated for past tense, i.e. *i-ta*. However, since *-teiru* and *-teita* function holistically as a verb conjugation, we will not divide the glossing of Japanese examples into *-te* and *i-ru*/*i-ta*, but will refer to these forms as *-teiru* and *-teita*.

Japanese has constructions that express tense-aspect, such as the aforementioned *-te*-form combined with the verb *kuru* 'to come' and *iku* 'to go', which can also be conjugated in the *-ta*-form, i.e. *ki-ta* 'came' and *it-ta* 'went'. For instance, *-te kuru* as in *fue-te kuru* means that the event ('increase' in this case) has been progressing 'up to a certain point in time'.

The *-ta koto ga aru*-construction is a sort of experiential construction and expresses that the person or thing referred to has the prior experience of participating in the event in question. Thus, *Igirisu ni i-tta koto ga a-ru* means that '(I) have been to the U.K. (before)'.

Japanese tense-aspect forms can be combined with other suffixes e.g. for voice, politeness and negation, as in (9 a–d).

28 The examples in (8 a–b) are based on Nitta et al. (2007).

(9) a. *Tabe-mashi-ta*
eat-POL-PAST
'(I) ate'

b. *Tabe-rare-ta*
eat-PASS-PAST
'(It) got eaten'

c. *Tabe-naka-tta*
eat-NEG-PAST
'(I) didn't eat'

d. *Tabe-rare-mas-en* *deshi-ta*
eat-PASS-POL-NEG COP-PAST
'(It) didn't get eaten'

5.2 *Material and method*

The material used was the parallel corpus Parallel Universal Dependencies (PUD) version 2.12 (Zeman et al. 2017), which consists of 1000 sentences gathered from Wikipedia texts and newspaper articles. 750 of these sentences have English as source language. The remaining 250 sentences were gathered from German, French, Italian and Spanish texts, but these have in turn been translated to the other languages via English. The material has been translated to 17 languages, among them Swedish and Japanese, and the UD annotation has been carried out – or has at least been reviewed – manually.

Since we are interested in both exact correspondences and translation divergences, we decided to perform four separate parallel searches using simple queries. First, we used the query

```
TREE_ (FEATS_ "VerbForm=Part") [AND [LEMMA "have", FEATS_
    "Tense=Pres"]]
```

on English-Swedish data to look for occurrences of the English PresPf and their Swedish equivalents.[29] A similar query, capturing the Swedish PresPf construction, was then run on both Swedish-English and Swedish-Japanese data:

```
TREE_ (FEATS_ "VerbForm=Sup") [AND [LEMMA "ha", FEATS_
    "Tense=Pres"]]
```

Searching for the Japanese form *-teiru* proved harder than expected. First of all, PUD tokenizes *-teiru* into *-te* + *-iru* (cf. Section 5.1 for the rationale behind this). Furthermore, the two tokens have several equivalent spellings,

29 This query was already discussed in Section 4.3.1.

which made it impractical to use FORM-based queries. Last but not least, *-teiru* can be conjugated in *-ta* – a form outside the scope of this study – or appear in several negated forms, which are, on the contrary, all relevant for our purposes. Therefore, the *-teiru* query underwent several revisions:

1. Our first attempt was the following:

```
SEQUENCE [XPOS "助詞-接続助詞", XPOS
   "動詞-非自立可能-上一段-ア行"]
```

   This query is based on language-specific POS tags as these are rather fine-grained when it comes to Japanese: *-te* is tagged as 助詞-接続助詞 (*joshi-setsuzokushi*, meaning 'particle-conjunction'),[30] whereas the XPOS tag for *-iru* is 動詞-非自立可能-上一段-ア行 (*doushi-hijiritsukanou jouichidan-agyou*, 'verb – non-independent jouichidangyou paradigm').[31] The `SEQUENCE` keyword combines these two forms and was expected to give us instances of *-teiru*. However, it turned out that results also included instances of the form *-teita* (*-teiru* conjugated in *-ta*, see above). Manual checking also revealed that this query failed to retrieve all instances of *-teiru*: whenever multiple instances of *-teiru* were present in the same sentence, only one was retrieved. This is because `SEQUENCE` queries return the smallest subtree covering the entire given sequence, which often "includes" several otherwise easily distinguishable matches. Negated *-teiru* forms, such as *-teinai*, were also excluded, as the negations imply changes in the second token, *-iru*. For this reason, we decided to base our second attempt on information from the UniDic morphological lexicon (Den et al. 2008), which the treebank stores in the MISC field under the key `UniDicInfo`.

2. This yielded a second, more compact query, which was ultimately the one we used:

```
TREE (MISC "UniDicInfo" "*ている") []
```

   ている is read as *teiru* and refers to the *-teiru* form.[32] This query correctly identified all instances of *-teiru*, including negated forms and

30 Alone, `XPOS "助詞-接続助詞"` also retrieves other forms which work as conjunctions, such as the particle *ga* and the verb form *nagara*.

31 動詞-非自立可能-上一段-ア行 actually refers to verbs with a certain inflectional paradigm and not just the auxiliary *iru*.

32 Here, using a `TREE` query with an empty list of dependents ensures that the token matching the subpattern `MISC "UniDicInfo" "*ている"` is a leaf. This helps better isolating the relevant sentence segments and not getting duplicate hits, since otherwise the whole subtree rooted in *-te* – potentially very extensive – would also be a match (`UniDicInfo` has the same value for both *-te* and *-iru*).

multiple occurrences in the same sentence. However, results also included *-teita-* forms.

3. We tried getting rid of these with a the third query, combining UniDic and XPOS information:

   ```
   SEQUENCE [TREE (MISC "UniDicInfo" "*ている") [], (NOT (XPOS
       "助動詞-助動詞-タ"))]
   ```

   助動詞-助動詞-タ is read as *jodoushi-jodoushi-ta* 'auxiliary-auxiliary-ta' and refers to the -ta form. This query failed however once again in finding negated *-teiru* forms and multiple occurrences in the same sentence. For this reason, we settled upon the second query and manually filtered out *-teita* forms.

5.3 *Results*

We present our results pairwise, based on the searches conducted. We start out with the search on English, yielding Swedish correspondences, and move next to Swedish, yielding English correspondences. We then do the same for Swedish and Japanese, i.e. search on Swedish to Japanese, and finally Japanese to Swedish.

We will only focus on *tense-aspect forms* in the Japanese categorization, and disregard other suffixes, e.g. for politeness, voice and polarity as these are not relevant for our purposes. Thus, a verb form such as *tabe-rare-tei-masen deshita* 'It had not been eaten (polite)' will be categorized as *teiru*-form. In the Japanese glossing, we will use the labels -TA and -TEIRU instead of PAST and ASP-NONPAST in order to highlight places where the *-ta* and *-teiru* forms appear (cf. examples 6 b and 6 c). Also, note that Japanese has a different writing system than Swedish and English. To make the results accessible to a broader audience, Japanese examples are romanized.

5.3.1 *Swedish equivalents to the English present perfect*

The search on the English PresPf gave 87 hits. The tense forms in the Swedish translation are given in Table 3. *Presens perfekt* occurred in the majority of cases (e.g. *har föreslagit* 'has proposed'), but in six cases *supinum utan hjälpverb* 'supine without auxiliary' appeared instead (e.g. *tidslinjer som flutit runt i medierna*, where *flutit* is the supine form). In one case *preteritum* (*visade* 'showed') appeared and in three cases *enkelt presens* (e.g. the main verb *har* 'have', where English used the present perfect (*e.g. have made*).

Table 3: Swedish equivalents to English present perfect.

Form	Frequency
Presens perfekt	77 (88,5%)
Supinum utan hjälpverb	6 (6,9%)
Enkelt presens	3 (3,4%)
Preteritum	1 (1,1%)
Sum	87 (100%)

Table 4: English equivalents to Swedish presens perfekt.

Form	Frequency
Present perfect non-progressive	75 (94,9%)
Present perfect progressive	3 (3,8%)
Simple present	1 (1,3%)
Sum	79 (100%)

5.3.2 *English equivalents to Swedish presens perfekt*

The search on the Swedish *presens perfekt* gave 79 hits. The tense-aspect forms in the English text that corresponded to the Swedish *presens perfekt* are given in Table 4. 'Non-progressive' in the table refers to cases where the progressive form was not used. The present perfect non-progressive (e.g. *have suggested*) occurred in most cases, but the present perfect progressive (e.g. *has been rising*) also occurred three times. No progressive constructions were used in Swedish in these three cases, but simply *presens perfekt* appeared, e.g. *har ökat* (which can be interpreted both as 'has been rising/has risen'). The only occurrence of the simple present in English was in a passive construction (*the Emperors are often referred to*), where Swedish used the PresPf in a passive construction (*har kejsarna ofta omtalats med*). No cases of the preterite or the pluperfect were found.

5.3.3 *Japanese equivalents to Swedish presens perfekt*

The query on the Swedish *presens perfekt* gave 79 hits. The tense-aspect forms in the Japanese text that corresponded the Swedish *presens perfekt* are given in Table 5. As the table shows, a great variety of tense-aspect forms appeared in the Japanese data. The most common form was *-ta* (e.g. *yosoku shi-**ta***

Table 5: Japanese equivalents to the Swedish present perfect.

Form	Frequency
-ru	6 (7,5%)
-ta	27 (34,1%)
-teiru	24 (30,3%)
-teita	3 (3,7%)
-ta koto ga aru	1 (1,2%)
-te	1 (1,2%)
-te kita	3 (3,7%)
Verb stem	7 (8,8%)
-teori	3 (3,7%)
Other	4 (5%)
Sum	79 (100%)

'foresaw'), closely followed by *-teiru* (e.g. *yousei shi-**teiru*** 'have requested').[33] *Presens perfekt* corresponded to *-ta* when the event was set in the past, usually with some sort of result as in (10) below.

(10) a. Swedish
*Röstning **har**, i terrorns jargong, **blivit** det nya mjuka målet*
voting has in the.terror's jargon become the new soft target

b. Japanese
*Touhyou wa, tero no kanten kara suru to, atarashii sofuto taagetto to nat-**ta**.*
voting TOP terror GEN perspective from do if new soft target PART become-TA
'Voting has, in the vernacular of terror, become the new soft target'

In (10), *har blivit* and *nat-ta* express that as a result of 'voting becoming the new soft target', it 'is' the new soft target at present. *Presens perfekt* corresponded to *-teiru* mainly in two cases: (1) with completed past events,

33 Note that the tense-aspect forms of the English translations do not fully mirror the meaning of the Japanese tense-aspect forms entirely, as these are rough translations. For instance, *yousei shi-teiru* in Japanese could be translated as 'requested/have requested/had requested/are requesting' depending on the context. We chose those forms in the English translation which we considered most natural in the case in question.

often with a result at present, as in (11), (2) an event that holds from the past up to present time, as in (12) below.

(11) a. Swedish
Mer än 5,7 miljoner floridabor ***har*** *redan*
more than 5,7 million Florida.citizens have already
gått *och* ***röstat*** *efter cirka två veckor med personligt*
gone and voted after about two weeks with personal
förtidsröstande
early.voting

b. Japanese
Yaku 2 shuukan no kijitsumaetouhyou go ni 570
about 2 weeks GEN early.voting after PART 570
man ijou no Florida juumin ga sude.ni
ten.thousand more GEN Florida citizen GEN already
*touhyou shi-****teiru*** *koto ga wakatta*
vote do-TEIRU thing NOM understood
'More than 5.7 million Floridians have already hit the polls after about two weeks of in-person early voting.'

(12) a. Swedish
Mycket av debatten, från Demokraternas sida i år,
much of the.debate from the.Democrats' side this year
har handlat *om vit manlig identitet*
has been about white male identity

b. Japanese
Kotoshi, minshutou gawa kara wa, hakujin dansei
this.year Democrats side from TOP white.people male
*no aidentiti ni.tsuite ooku no giron ga nas-are-****teiru***
GEN identity about much GEN debate NOM do-PASS-TEIRU
'Much of the debate, from the Democratic side this year, has been about white male identity.'

In (11) *har gått och röstat* and *touhyou shi-teiru* express that the event of 'voting' has already been completed for many participants. In (12), *har handlat om* and *giron ga nasare-teiru* express that the event of 'debating white male identity' has been occurring up to now.

The *-ru*-form appeared six times. An example of this is *tsune ni seijitekina sokumen ga a-****ru*** 'always has had a political dimension', corresponding to Swedish ***har*** [...] *alltid* ***haft*** *en politisk dimension* 'has always had a political dimension'. *-teita* occurred three times, as in where Japanese *rokuten tot-****teita***

Table 6: Swedish correspondences to Japanese *-teiru*.

Form	Frequency	
Enkelt presens	65	(65%)
Preteritum	11	(11%)
Presens perfekt	11	(11%)
Skulle + infinitival verb	2	(2%)
Kommer + infinitival verb	1	(1%)
Infinitiv verbfras	1	(1%)
Perfekt infinitiv	1	(1%)
Other	8	(8%)
Sum	100	(100%)

'(someone) took six points' corresponded to Swedish *vi* ***har tappat*** *sex poäng* 'we have lost six points'.

There were also cases where *presens perfekt* corresponded to the forms *verb stem* (e.g. *rikai shi* - 'having understood'), *-teori* (e.g. *zouka shi-te ori* 'having increased') and *-te* (*at-te* 'having met').[34]

The only occurrence of *-ta koto ga aru* (*atta koto mo nai* 'there has not been') was conjugated for negation and the *ga*-particle was replaced with *mo* 'also'. The construction *-te kita* occurred three times, as in *chousa shi****-te kita*** 'has investigated up to now', which corresponded to Swedish *har undersökt* 'has investigated'.

The Other category consists of cases where *presens perfekt* did not correspond to a verb, e.g. Japanese *aru Sueeden no kenkyuu ni yoru to* 'according to Swedish research, ...' corresponded to Swedish *en svensk studie har visat att [...]* 'a Swedish study has shown that...'.

5.3.4 *Swedish equivalents to Japanese* -teiru

The search for Japanese *-teiru* to Swedish gave 356 hits. Some of the hits were excluded, since the query also included *-teita* forms which are not in the scope of this study. Due to time restrictions we extracted only the first 100 occurrences of *-teiru* that appeared in the concordance. Table 6 shows the Swedish forms that corresponded to Japanese *-teiru*.

The most frequent form in Swedish was *enkelt presens*, which made up the majority of the occurrences. In many of these cases the Swedish *säger* 'says'

34 These are not tense-aspect forms per se, as they do not carry tense-aspect information - the exception being *-te ori*, which carries information for progressive aspect - but since they occurred multiple times we included them in our results.

corresponded to the Japanese verb *nobe-teiru* '(someone) mentions'. *Preteritum* and *presens perfekt* occurred to the same extent. *Preteritum* appeared for instance as *stödde* 'supported', corresponding to Japanese *shiji shi-teiru* 'have supported'. and *presens perfekt* appeared as for instance *har bett* 'has requested', corresponding to Japanese *yousei shi-teiru* 'has requested'.

Skulle 'would' + infintival verb appeared twice. In both cases, the auxiliary *ska* was conjugated in the past tense *skulle*, which is why we indicate these occurrences as *skulle* and not *ska*.

Kommer 'is going to' + infintival verb occurred only once, which was the occurrence *kommer öka* 'is going to increase', where the auxiliary was conjugated in present tense *kommer* 'come' and not *kom* 'came'. *Infinitiv verbfras* 'infinitive verb phrase' appeared once, where Swedish (*för*) ***att prata*** (*om...*) '(to) talk (about)' corresponded to Japanese [...] *ni tsuite hanashite-ki-**teiru*** 'has talked about'. *Perfekt infinitiv* 'perfect infinitive' also appeared once, where Swedish *att **ha varit** inblandade* 'to have been involved' corresponded to Japanese *juuji shi-**teiru*** 'have been involved in'.

The Other category consists of cases where the *-teiru* form either did not correspond to a verb or it did not correspond to anything at all in the Swedish text. For instance, Swedish *enligt ...* 'according to...' corresponded to Japanese *to houjirare-teiru* 'was reported that...', which is not a verb in Swedish. An example of no correspondence is *kai-teiru* 'writes' in (13) below – which does not have any correspondence in the Swedish text.

(13) a. Japanese
"Saishin no ibento" wa Reddit, Twitter, Spotify ya
recent GEN event TOP Reddit Twitter Spotify PART
sono.ta ooku no ninki websaito ga kanketsu.ni
other many GEN popular website NOM briefly
*kai-**teiru**, 10 gatsu 21 nichi no tero jiken dearu*
write-TEIRU October 21st GEN terror incident COP
kanousei ga takai
probability NOM high

b. Swedish
De "nyliga händelserna" är troligen attackerna den 21
the recent events are likely the.attacks the 21
oktober som tillfälligt slog ut populära
October which temporarily knocked out popular
webbsidor som Reddit, Twitter och Spotify såväl som
websites like Reddit Twitter and Spotify as.well as
många andra
many others
'The "recent events" are likely to be the attacks of 21 October that briefly took down popular websites such as Reddit, Twitter and Spotify as well as many others.'

No instances of other tenses, such as *pluskvamperfekt* 'pluperfect' were found.

5.4 *Discussion*

In this section, we briefly discuss what tendencies in the use of tense-aspect forms could be seen in our data. Due to space limitations, we restrict our observations to general patterns.

The case study showed similarities and differences in distribution between Swedish and English, and considerable variation between Swedish and Japanese. We refrain from making any in-depth observations on the semantics of tense-aspect forms, but our ambition is to address semantic issues in future studies.

5.4.1 *Swedish and English*

Our results showed a strong correspondence between the Swedish and English PresPf. However, the English present perfect was used to a greater extent than the Swedish *presens perfekt*.

The difference seems to depend on both language-specific conventions and translation-induced variation. For instance, the English present perfect in [...] *I* ***have made*** *a full recovery* was translated as *enkelt presens*: [...] ***är*** *jag fullt återhämtad* in Swedish. This ought to be due to the fact that the expression *make a full recovery* is idiomatic and does not have an immediate counterpart in Swedish. Therefore, the translator might have preferred the expression *vara fullt återhämtad* which in the present context sounds more natural in the *enkelt presens* than in *presens perfekt*. Another example is where the English present perfect *But a scan* ***has shown*** *the tumour in Ms Pugh's right lung is growing* [...] was translated to *preteritum* in Swedish: *Men en*

undersökning ***visade*** *att tumören i Mr Pughs högra lunga växer* [...]. It is unclear why *preteritum* is used here, but one explanation might be that *preteritum* is preferred in Swedish in narrative texts (Holm 2010).

Furthermore, our results might be due to genre-specific differences. A pilot study we carried out on the LinES treebank, which consists to a great extent of fiction texts, showed that the Swedish *presens perfekt* was used more frequently than the English present perfect. In the future, it would therefore be advantageous to use bigger datasets encompassing various genres to investigate the distribution of tense-aspect forms in the two languages.

5.4.2 *Swedish and Japanese*

The Swedish and Japanese results gave a diverse picture of the tense-aspect systems of the two languages. Judging from our results, the Swedish *presens perfekt* mostly overlaps in use with Japanese *-ta* and *-teiru*. In our data, both *-ta* and *-teiru* tended to refer to past events, often with a result holding after the occurrence of the event, and it was in these cases that these forms corresponded to *presens perfekt* (examples 10 and 11). However, one use that *presens perfekt* had in common with *-teiru* but not with *-ta* was to refer to continuative events starting in the past and persisting up to present time (example 12).

Based on our data, Japanese *-teiru* mostly overlaps in use with Swedish *enkelt presens*, chiefly when the two forms refer to an event holding at present. *-teiru* also overlapped in use with *preteritum* and *presens perfekt* when referring to past events, often with a result holding after the occurrence of the event. Since both *presens perfekt* and *preteritum* can often be used in similar contexts (Teleman et al. 1999), the occurrence of *preteritum* was not entirely unexpected. However, it was surprising that *preteritum* and *presens perfekt* occurred to the same extent, since we expected *-teiru* to correspond mainly to Swedish tenses with a present time reference, i.e. *presens* and *presens perfekt*. A possible reason for the unexpected low frequency of *presens perfekt* might be that *-ta* also overlaps in use with *presens perfekt* (as shown in 5.3.3, cf. also Tóth 2020) when referring to past events with a connection to the present, and *-ta* might correspond to *presens perfekt* more frequently than *-teiru* in the corpus. A future investigation on the Swedish correspondences with *-ta* would likely shed light on whether this is the case and if this might depend on semantics, genre, translation issues or other factors.

The unexpected high frequency of *preteritum* might, on the other hand, be due to restrictions on how tense is used in a greater context in Swedish and Japanese. There is a tendency in Swedish to either use so-called presentical tenses, i.e. *presens* and *presens perfekt* or preterital tenses, i.e. *preteritum* and

pluskvamperfekt throughout a sequence (Ekerot 2011). In case a text sequence mainly refers to past events, this could facilitate the use of *preteritum* throughout the text, even in cases where *presens perfekt* could be used. Japanese, on the other hand, does not have such a tendency, as past and non-past forms can freely be used throughout a text. Since we do not have access to contextual information in the PUD corpus we unfortunately cannot check whether the use of *preteritum* contra *presens perfekt* is due to textual factors.

By comparing Swedish, English and Japanese we could demonstrate how STUnD can be used to investigate the distribution of tense-aspect forms in languages with fairly similar tense-aspect systems (i.e. Swedish and English) and languages with fairly different tense-aspect systems (i.e. Swedish and Japanese). It needs to be kept in mind that both the Swedish and Japanese texts are translations from English, which likely influences the tense-aspect forms used. Swedish has a morphosyntax similar to that of English, while Japanese morphosyntax differs to a greater extent, which might influence the use of tense-aspect forms depending on how the clauses in the sentence are structured. Even so, we judged that the texts reflected natural language use and the use of tense-aspect in our material should reflect the general use of tense-aspect forms in all three languages.

5.5 *Considerations on using STUnD*

The tool was useful for finding corresponding tense-aspect forms between the three languages. However, we encountered issues when Japanese was included in our search.

One strength turned out to be that we could easily identify all instances of PresPf in Swedish and English, since the tool can find dependencies between the auxiliary and supine form in Swedish or the perfect participle in English. Through `TREE` queries, the tool could do this even in cases where several words appeared between the auxiliary and the main verb, as in *har kejsarna ofta omtalats* (approximately 'have the.emperors often been.referred'). The English query allowed us to find occurrences in Swedish where the supine form appeared without an auxiliary. As Figure 4 shows, the tool marks the matching segments and – when possible – their counterpart in the other treebank. This made the search results easily comprehensible. Thus, based on our study, the alignment between Swedish and English, as well as form-based TREE-queries work well in the tool.

While searching for the PresPf in Swedish and English proved easy and gave us the information we wished to retrieve, as we showed in 5.2, the search for *-teiru* in Japanese turned out to be more complicated. As mentioned in the same section, we had to revise the query for Japanese several times in

order to get favourable search results. Some difficulties in this sense, such as homographs making it laborious to look for specific forms through FORM-based queries, are intrinsic to the language itself. Some others, however, are related to the exceptionality of the UD annotation guidelines for Japanese, especially when it comes to tokenization (Guillaume et al. 2024).

One point that pertains to all three languages is that the highlighting was not always precise: the tool sometimes highlighted whole segments in bold and not only the corresponding forms, which still required manual checking. This is due to the very nature of the queries, which always return (unpruned) subtrees as results, and can be mitigated through the application of replacement patterns, which were however challenging to formulate and therefore not used extensively for Japanese. Through the replacement pattern

```
COMPOSE [FILTER_SUBTREES (UPOS "VERB") (UPOS "AUX"), PRUNE TRUE 1]
```

the tool highlighted the Swedish *presens perfekt* and English present perfect forms correctly in most cases. However, in 8 out of 79 cases from Swedish to English and 9 out of 87 cases in the opposite direction, the tool did not highlight any segments in the right-hand-side treebank, even when we used the aforementioned replacement pattern. The same replacement pattern was less successful for Japanese. One-lemma forms such as *-ta* and *-ru* were highlighted correctly, while in the case of forms that UD analysis decomposes as two separate tokens, such as *-teiru*, the tool also highlighted tokens other than the verb form corresponding to the Swedish *presens perfekt*.

In this study, we did not use divergence patterns. In the future, however, these could be useful in searching for translation divergences found in other corpora, such as LinES (Ahrenberg 2007, 2015).

6 *Other use cases*

In Section 5, we have shown how STUnD can be used in the context of a cross-lingual study. As mentioned earlier, however, the tool can be used in a variety of contexts, including cases where the parallel corpus of interest is not multilingual. With the hope to provide inspiration for new STUnD users, we present two such applications: grammatical error retrieval and qualitative parser evaluation.

6.1 *Grammatical error retrieval*

UD is gaining popularity as an annotation standard for learner corpora, with treebanks already available for second-language (L2) Chinese (Lee, Leung, et al. 2017), English (Berzak et al. 2016, Kyle et al. 2022), Italian (Di Nuovo et al.

2022), Korean (Sung & Shin 2024) and Russian (Rozovskaya 2024) and one under development for Swedish. The main advantage of UD in this setting is that it allows studying the differences and similarities between a learner's first and second language, between different L2s or between standard and learner language. Moreover, the richness of UD annotation is usually sufficient for grammatical error retrieval, without the need for any explicit annotation layer in this sense. This is especially true for *L1-L2 treebanks*, a type of parallel UD corpus where (L2) learner productions are contrasted with expert corrections, assumed to be native-like (L1).

In Section 4.4.2 (cf. Figure 12 in particular), we showed how to use STUnD's diff mode to identify errors in one such corpus, the ESL treebank (Berzak et al. 2016). To complete the picture of how STUnD can be put into use in learner corpus research, this section will demonstrate how parallel queries can help retrieve instances of particular error patterns. Since the ESL treebank lacks morphological tagging, however, we will use VALICO-UD (Di Nuovo et al. 2022), a parallel treebank of L2 Italian.

First of all, many grammatical errors are easily retrieved via simple queries, without needing to compare learner productions with their corrected versions. Agreement errors are a good example of this. The query

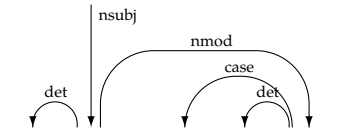
```
TREE_
    (FEATS_ "Gender=Fem")
    [AND [OR [DEPREL_ "det", DEPREL_ "amod"], FEATS_
        "Gender=Masc"]]
```

for instance, matches all subtrees where a feminine token (typically a noun) is introduced by a masculine determiner or modified by the masculine form of an adjective (see Figure 13 for an example of application).

In this case, results are no different from what could be obtained by using any other treebank search tool mentioned in Section 3. Querying parallel L1-L2 treebanks, however, allows users to write simpler, more general queries, whilst limiting search results to grammatical errors. If we were to simplify the query to

```
TREE_ (FEATS_ "Gender=Fem") [FEATS_ "Gender=Masc"]
```

with the goal to look for head-dependent gender agreement errors without specifying the syntactic function of either, the results would include a vast amount of "false positives", such as the sentence fragment

(14)


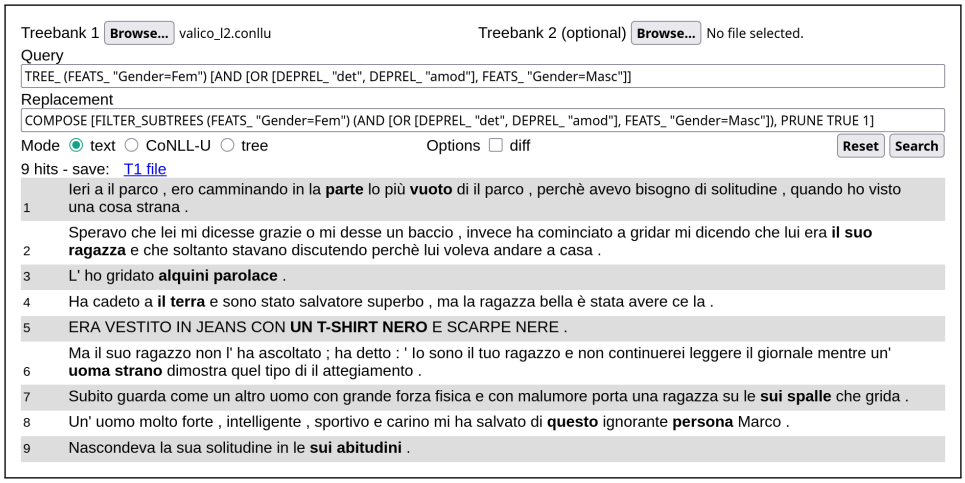


Figure 13: Using STUnD to retrieve agreement errors from L2 productions from the VALICO-UD treebank. In this case, the search is limited to cases where a masculine determiner or adjective is attached to a feminine nominal.

which matches the query not because of an agreement error, but simply because the nominal modifier of the feminine noun *natura* 'nature' is a masculine (*parco*, 'park'). The divergence pattern

```
TREE_ (FEATS_ "Gender=Fem") [FEATS_ "Gender={Fem->Masc}"]
```

on the other hand, limits the results to the cases in which the gender of the dependent is masculine in the original learner sentence but has been corrected to agree with that of its syntactic head in the L1 half of the treebank.

Furthermore, divergence patterns make it possible to identify errors that would be impossible to locate in a monolingual L2 treebanks. The even simpler query

```
FEATS_ "Gender={Fem->Masc}"
```

for example, finds all instances where a feminine token in the L1 half of the treebank corresponds to a masculine one in its L2 counterpart, including cases where the correction involves replacing a correctly inflected but lexically inappropriate masculine noun with a more fitting feminine one, as well as sentences where an entire noun phrase is incorrectly inflected for gender (cf. Figure 14).

6.2 *Qualitative parser evaluation and inter-annotator conflict resolution*

So far, we have seen how STUnD can be used on parallel multilingual and L1-L2 treebanks. A third use case is comparing different analyses of the same

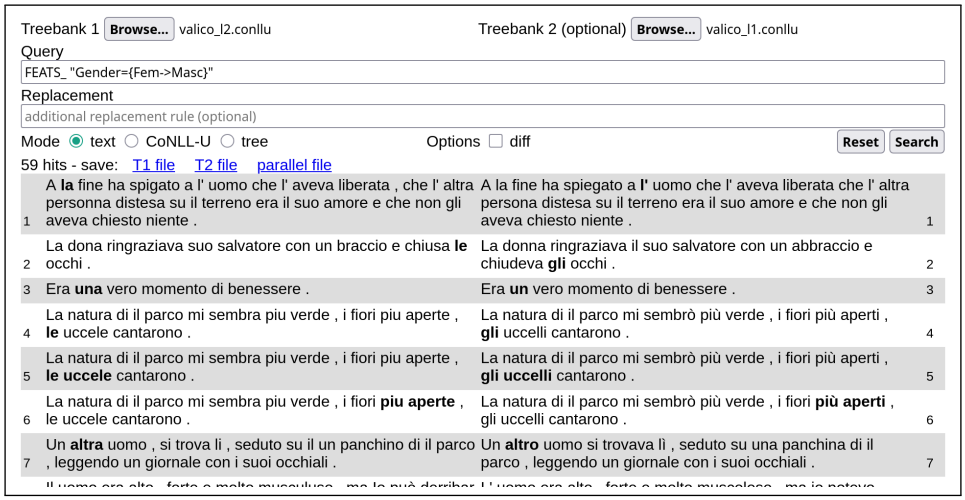


Figure 14: The divergence pattern `FEATS_ "Gender={Fem->Masc}"` run on the VALICO-UD treebank. In many cases, matches are still representative of agreement errors. Hit number 5, however, shows a case in which the entire noun phrase *le uccele* is incorrectly inflected for gender. In the first hit, the two highlighted segments do not correspond to each other, probably due to an annotation error.

surface text. This can be useful both when evaluating automatic parsers on a sentence-by-sentence basis, e.g. by comparing two different models, and when resolving conflicts between human annotators.

In Figures 15–16, an example text in English has been automatically annotated using two different UDPipe2 models.[35] Using STUnD's diff functionality (cf. Section 4.4.2) in text mode gives us a general overview of how similar the analyses are (cf. Figure 15). To gain any insight into how the analyses are different, however, the user needs to switch to either tree or CoNNL-U mode and look beneath the surface. In tree mode, we can see that *Search*, *Tool*, and *Dependencies* are assigned the POS tag `NOUN` on the left side and the tag `PROPN`, i.e. proper noun, on the right (cf. hit number 1 in Figure 16). Some other discrepancies can only be identified by inspecting the CoNNL-U source. The GUM model, for instance, assigns the word *specifically* the feature `Degree=Pos`, which is lacking in the EWT-based analysis. If these were two competing human annotator analyses in the context of a treebanking project, conflicts could be resolved by editing one of the files directly in STUnD.

35 The example text is the abstract of this chapter and the models used are `english-gum-ud-2.15-241121` and `english-ewt-ud-2.15-241121`.

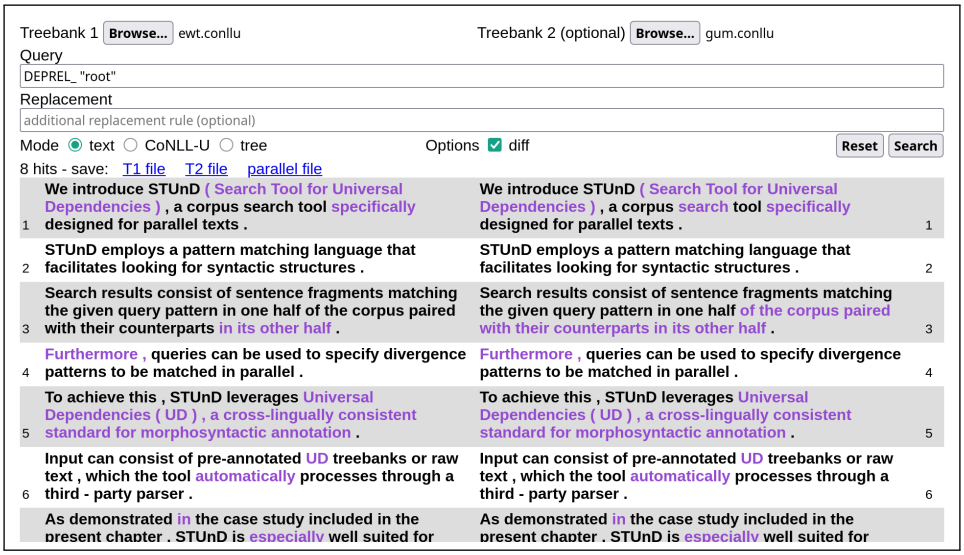


Figure 15: Using diff mode to locate annotation inconsistencies in text mode.

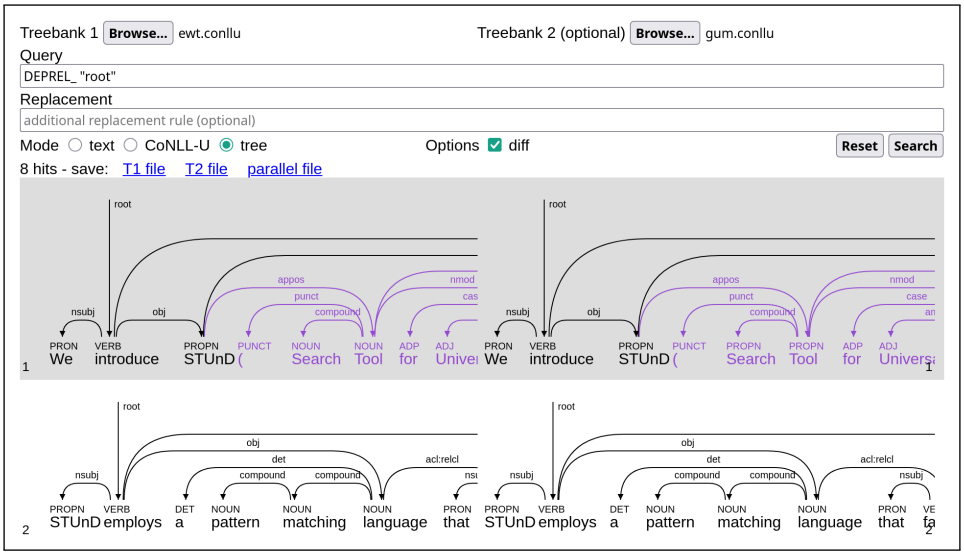


Figure 16: Using diff mode to locate annotation inconsistencies in tree mode.

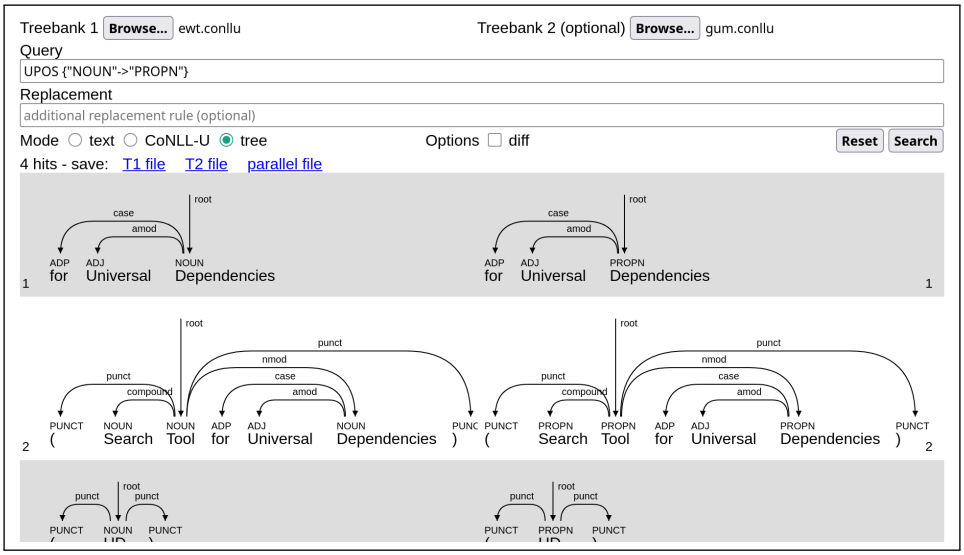


Figure 17: Using divergence patterns to search for specific types of annotation inconsistencies.

Finally, we can use divergence patterns to check whether the differences we have identified are systematic. For example, we can use the following query to look for all cases where a token is classified as `NOUN` by one model and `PROPN` by the other:

```
UPOS "{NOUN->PROPN}"
```

As we can see in Figure 17, there are several occurrences of this phenomenon in this short text, so we can expect this to be due to inconsistencies in the way in which the tags are used in the treebanks on which the two models have been trained.

7 *Concluding remarks*

We presented STUnD, a search tool that leverages Universal Dependencies to facilitate the exploration of parallel corpora. At the core of STUnD are two unique features (1) a query language that allows searching for (diverging) syntactic structures in a largely language-agnostic fashion and (2) a subtree alignment algorithm that enables parallel query matching and fine-grained comparisons between two versions of the same text. In addition, STUnD distinguishes itself from most other corpus search tools in that it integrates NLP modules that make it possible to also work on raw text, as well as allow

for the user to interact with the search results by editing them directly in its interface (cf. Sections 4.3.3 and 4.4.3) and exported to a variety of formats, including CoNNL-U itself.

We identified cross-lingual studies as the prime area of application for STUnD. We therefore put the tool to the test in a trilingual case study (cf. Section 5), which shed light on its strengths, but also revealed its current limitations from a user perspective. In particular, the study revealed that query-matching segments extracted from the tool were often larger than one would have hoped and that suitable replacement patterns – which are intended to mitigate this – were challenging to formulate. This, together with the consideration that only a small subset of the available replacement patterns are actually necessary to improve the search results (cf. 4.3.3), prompts us to experiment with automating the refinement step. Queries themselves were sometimes difficult to express, which encourages us to resume experiments already initiated on example-based querying (also referred to as *similar example retrieval*, cf. Masciolini et al. 2023).

In relation to the use of STUnD on learner corpora (Section 6.1), we plan to integrate the experimental (error) pattern extraction functionalities presented in Masciolini et al. (2023) into STUnD's user interface. Finally, STUnD is being tested as a conflict resolution tool in the context of an ongoing treebanking project and will be further developed to meet the annotators' needs. More generally, users and potential contributors are encouraged to follow the development of the tool, report bugs and propose new features.[36]

Acknowledgements

This work is supported and partly funded by Språkbanken – jointly funded by its 10 partner institutions and the Swedish Research Council (2025–2028; project id 2023-00161) as well as (2018–2024; dnr 2017-00626). We are grateful to the Huminfra handbook editors and reviewers for their support throughout the writing process. We would also like to thank our colleagues at Språkbanken Text and the Department for Swedish, Multilingualism, Language Technology at the University of Gothenburg, in particular Niklas Zechner, for their valuable feedback.

36 For an overview of the development process, visit github.com/users/harisont/projects/8. For bug reports and feature request, open an issue at github.com/harisont/STUnD/issues or email arianna@fripost.org.

References

Ahrenberg, Lars. 2007. LinES: An English-Swedish parallel treebank. In Joakim Nivre, Heiki-Jaan Kaalep, Kadri Muischnek & Mare Koit (eds.), *Proceedings of the 16th Nordic Conference of Computational Linguistics* (*NODALIDA 2007*), 270–273. Tartu, Estonia: University of Tartu, Estonia. https://aclanthology.org/W07-2441.

Ahrenberg, Lars. 2015. Converting an English-Swedish parallel treebank to Universal Dependencies. In Joakim Nivre & Eva Hajičová (eds.), *Proceedings of the Third International Conference on Dependency Linguistics* (*Depling 2015*), 10–19. Uppsala, Sweden: Uppsala University. https://aclanthology.org/W15-2103.

Berzak, Yevgeni, Jessica Kenney, Carolyn Spadine, Jing Xian Wang, Lucia Lam, Keiko Sophie Mori, Sebastian Garza & Boris Katz. 2016. Universal Dependencies for learner English. In Katrin Erk & Noah A. Smith (eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (*ACL*) (*Volume 1: Long Papers*), 737–746. Berlin, Germany: Association for Computational Linguistics. DOI: 10.18653/v1/P16-1070.

de Marneffe, Marie-Catherine, Christopher D. Manning, Joakim Nivre & Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics* 47(2). 255–308. DOI: 10.1162/coli_a_00402.

Den, Yasuharu, Junpei Nakamura, Toshinobu Ogiso & Hideki Ogura. 2008. A proper approach to Japanese morphological analysis: Dictionary, model, and evaluation. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis & Daniel Tapias (eds.), *Proceedings of the Sixth International Conference on Language Resources and Evaluation* (*LREC'08*). Marrakech, Morocco: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/pdf/258_paper.pdf.

Di Nuovo, Elisa, Manuela Sanguinetti, Alessandro Mazzei, Elisa Corino & Cristina Bosco. 2022. VALICO-UD: Treebanking an Italian learner corpus in Universal Dependencies. *Italian Journal of Computational Linguistics* 8(1). DOI: 10.4000/ijcol.1007. (4 October, 2024).

Ekerot, Lars-Johan. 2011. *Ordföljd, tempus, bestämdhet: Föreläsningar om svenska som andraspråk* [Word order, tempus and definiteness: Lectures on Swedish as a Second Language]. 2nd edition. Malmö: Gleerups.

Guibon, Gaël, Marine Courtin, Kim Gerdes & Bruno Guillaume. 2020. When collaborative treebank curation meets graph grammars. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk & Stelios

Piperidis (eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC 2020)*, 5291–5300. Marseille, France: European Language Resources Association. https://aclanthology.org/2020.lrec-1.651/.

Guillaume, Bruno. 2021. Graph matching and graph rewriting: GREW tools for corpus exploration, maintenance and conversion. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL): System Demonstrations*, 168–175. Online: Association for Computational Linguistics. DOI: 10.18653/v1/2021.eacl-demos.21.

Guillaume, Bruno, Kim Gerdes, Kirian Guiller, Sylvain Kahane & Yixuan Li. 2024. Joint annotation of morphology and syntax in dependency treebanks. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti & Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 9568–9577. Torino, Italia: ELRA & ICCL. https://aclanthology.org/2024.lrec-main.836.

Holm, Lisa. 2010. Presens och preteritum i skönlitterär prosa. In Gunilla Byrman, Anna Gustafsson & Henrik Rahm (eds.), *Svensson och svenskan. Med sinnen känsliga för språk. Festskrift till Jan Svensson den 24 januari 2010*, 49–59. Lund University Publications.

Kolachina, Prasanth & Aarnte Ranta. 2016. From abstract syntax to Universal Dependencies. *Linguistic Issues in Language Technology* 13. DOI: 10.33011/lilt.v13i.1391.

Krsnik, Luka, Kaja Dobrovoljc & Marko Robnik-Šikonja. 2019. *Dependency tree extraction tool STARK 1.0*. Software. http://hdl.handle.net/11356/1284.

Kudou, Mayumi. 1995. *Asupekuto-tensu taikei to tekusuto* [Aspect-tense: System and text]. Tōkyō, Japan: Hitsuji Shobō. DOI: 10.11501/3155604.

Kyle, Kristopher, Masaki Eguchi, Aaron Miller & Theodore Sither. 2022. A dependency treebank of spoken second language English. In Ekaterina Kochmar, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Nitin Madnani, Anaïs Tack, Victoria Yaneva, Zheng Yuan & Torsten Zesch (eds.), *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, 39–45. Seattle, Washington: Association for Computational Linguistics. DOI: 10.18653/v1/2022.bea-1.7.

Larsson, Ida & Benjamin Lyngfelt. 2011. Tempus i svenskan [Tense in Swedish]. In *Tid och tidsförhållanden i olika språk* (Studia Interdisciplinaria, Linguistica et Litteraria (SILL) 2), 66–88. Gothenburg, Sweden: Institutionen för Språk och Litteraturer, Göteborgs universitet.

Lee, John, Herman Leung & Keying Li. 2017. Towards Universal Dependencies for learner Chinese. In Marie-Catherine de Marneffe, Joakim Nivre & Sebastian Schuster (eds.), *Proceedings of the NoDaLiDa 2017 Workshop on*

Universal Dependencies (*UDW 2017*), 67–71. Gothenburg, Sweden: Association for Computational Linguistics. https://aclanthology.org/W17-0408.

Lee, John, Keying Li & Herman Leung. 2017. L1-L2 parallel dependency treebank as learner corpus. In *Proceedings of the 15th International Conference on Parsing Technologies*, 44–49. Pisa, Italy: Association for Computational Linguistics. https://aclanthology.org/W17-6306.

Lui, Marco & Timothy Baldwin. 2011. Cross-domain feature selection for language identification. In Haifeng Wang & David Yarowsky (eds.), *Proceedings of 5th International Joint Conference on Natural Language Processing*, 553–561. Chiang Mai, Thailand: Asian Federation of Natural Language Processing. https://aclanthology.org/I11-1062.

Lui, Marco & Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In Min Zhang (ed.), *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics* (*ACL*): *System Demonstrations*, 25–30. Jeju Island, Korea: Association for Computational Linguistics. https://aclanthology.org/P12-3005.

Luotolahti, Juhani, Jenna Kanerva, Sampo Pyysalo & Filip Ginter. 2015. SETS: Scalable and Efficient Tree Search in dependency graphs. In Matt Gerber, Catherine Havasi & Finley Lacatusu (eds.), *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics* (*NAACL*): *Demonstrations*, 51–55. Denver, Colorado: Association for Computational Linguistics. DOI: 10.3115/v1/N15-3011.

Masciolini, Arianna. 2023. A query engine for L1-L2 parallel dependency treebanks. In *Proceedings of the 24th Nordic Conference on Computational Linguistics* (*NoDaLiDa*), 574–587. Tórshavn, Faroe Islands: University of Tartu Library. https://aclanthology.org/2023.nodalida-1.57.

Masciolini, Arianna & Aarne Ranta. 2021. Grammar-based concept alignment for domain-specific machine translation. In *Proceedings of the Seventh International Workshop on Controlled Natural Language* (*CNL 2020/21*). Amsterdam, Netherlands: Special Interest Group on Controlled Natural Language. https://aclanthology.org/2021.cnl-1.2.

Masciolini, Arianna & Márton András Tóth. 2024. STUnD: Ett Sökverktyg för Tvåspråkiga Universal Dependencies-trädbanker [STUnD: A search tool for bilingual Universal Dependencies treebanks]. In *Proceedings of the Huminfra Conference* (*HiC 2024*), 95–109. Gothenburg, Sweden: Linköping University Press. DOI: 10.3384/ecp205013.

Masciolini, Arianna, Elena Volodina & Dana Dannélls. 2023. Towards automatically extracting morphosyntactical error patterns from L1-L2 parallel dependency treebanks. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications* (*BEA 2023*), 585–

597. Toronto, Canada: Association for Computational Linguistics. DOI: 10.18653/v1/2023.bea-1.50.

Nitta, Yoshio et al. 2007. *Gendai nihongo bunpō* [Modern Japanese grammar]. Vol. 3: *dai 5-bu. Asupekuto ; dai 6-bu. Tensu ; dai 7-bu. Kōhi* [Part 5: Aspect; Part 6: Tense; Part 7: Polarity]. Tōkyō, Japan: Kuroshio Shuppan.

Pajas, Petr & Jan Štěpánek. 2009. System for querying syntactically annotated corpora. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL-IJCNLP 2009): Software Demonstrations*, 33–36. Suntec, Singapore: Association for Computational Linguistics. https://aclanthology.org/P09-4009.

Ranta, Aarne & Prasanth Kolachina. 2017. From Universal Dependencies to abstract syntax. In *Proceedings of the NoDaLiDa 2017 workshop on universal dependencies (UDW 2017)*, 107–116. Gothenburg, Sweden: Association for Computational Linguistics. https://aclanthology.org/W17-0414.

Rozovskaya, Alla. 2024. Universal Dependencies for learner Russian. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti & Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 17112–17119. Torino, Italia: ELRA & ICCL. https://aclanthology.org/2024.lrec-main.1486.

Shirai, Yasuhiro. 2000. The semantics of the Japanese imperfective -teiru: An integrative approach. *Journal of Pragmatics* 32(3). 327–361. DOI: 10.1016/S0378-2166(99)00051-X.

Sung, Hakyung & Gyu-Ho Shin. 2024. Constructing a dependency treebank for second language learners of Korean. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti & Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 3747–3758. Torino, Italia: ELRA & ICCL. https://aclanthology.org/2024.lrec-main.332.

de Swart, Henriëtte. 2016. Perfect usage across languages. *Questions and answers in linguistics* 3(2). 57–62.

Teleman, Ulf, Staffan Hellberg & Erik Andersson. 1999. *Svenska akademiens grammatik* (*'Swedish Academy grammar'*), vol. 4 – Satser och meningar. Svenska Akademien. https://svenska.se/grammatik/.

Tóth, Márton András. 2020. *The present perfect in Swedish, Japanese and Hungarian*. Osaka: Osaka university. (MA thesis).

Tóth, Márton András. 2022. Kontinuativt presens perfekt i svenskan [Continuative present perfect in Swedish]. *IDUN – Journal of Nordic Studies* 24. 1–19. DOI: 10.18910/87434. (30 September, 2024).

Zeman, Daniel, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria de Paiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonça, Tatiana Lando, Rattima Nitisaroj & Josie Li. 2017. CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies. In Jan Hajič & Dan Zeman (eds.), *Proceedings of the CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies*, 1–19. Vancouver, Canada: Association for Computational Linguistics. DOI: 10.18653/v1/K17-3001. https://aclanthology.org/K17-3001/.

List of abbreviations

Abbreviations used in glosses

ACC	accusative particle
ASP	aspectual inflection
COP	copula
GEN	genitive particle
NOM	nominative particle
PART	particle with various functions
TOP	topicalizing particle

Other

NLP	Natural Language Processing
POS	Part of Speech
UD	Universal Dependencies
PUD	Parallel Universal Dependencies

CoNNL	Conference on Computational Natural Language Learning
L1	first language
L2	second or foreign language
TSV	Tab-Separated Values
SVG	Scalable Vector Graphics

8 *Corresponding authors*

Arianna Masciolini
Språkbanken Text
Department of Swedish, Multilingualism, Language Technology
University of Gothenburg
arianna.masciolini@gu.se

Herbert Lange
Språkbanken Text
Department of Swedish, Multilingualism, Language Technology
University of Gothenburg
herbert.lange@gu.se

Márton András Tóth
Department of Swedish, Multilingualism, Language Technology
University of Gothenburg
marton.toth@gu.se

15 Low-code web scraping and text analysis with Octoparse and KNIME: An example from the CICuW project

Daniel Ihrmark
Linnæus University

Hanna Carlsson
Linnæus University

Fredrik Hanell
Linnæus University

Low-code tools play an important role in making data analysis and visualization accessible to researchers and students with limited experience, or interest, in programming. While low-code tools do introduce closed-box issues, they can still be considered important stepping stones toward computational approaches. This chapter draws on two such tools, Octoparse and KNIME (Konstanz Information Miner), to present a workflow from data collection from online sources, through text pre-processing, toward text classification in the context of the ongoing project Cultural Institutions and the Culture War (CICuW) that investigates the democratic implications of the pervasiveness of far-right digital discourse.

This chapter will introduce web scraping, topic modeling, and sentiment analysis in an accessible way, while also showcasing state-of-the-art approaches to the analysis components through the use of BERT (Bidirectional Encoder Representations from Transformers) models and zero-shot classification. The chapter will take a critical perspective on the described methods by discussing how they contribute to creating methodological closed-boxes and how quantitative techniques can be fruitfully combined with qualitative approaches.

1 *Introduction*

In Sweden, the growing influence of the far-right has turned cultural institutions into political symbols in an emerging "culture war" (Harding 2022). Carlsson et al. (2022) show how digital forums and social media play a significant part in orchestrating the ideologically laden conflicts and

HUM INFRA Daniel Ihrmark, Hanna Carlsson & Fredrik Hanell. 2025. Low-code web scraping and text analysis with Octoparse and KNIME: An example from the CICuW project. In Gerlof Bouma, Dana Dannélls, Dimitrios Kokkinakis & Elena Volodina (eds.), *Huminfra handbook: Empowering digital and experimental humanities* (NEALT Proceedings Series 59), 505–540. University of Tartu Library. DOI: 10.58009/aere-perennius0184 © The authors, CC BY 4.0

confrontations that public libraries are currently facing. Digitally mediated threats from the far-right may obstruct the statutory mission of the institution to promote democracy, but knowledge about how such threats develop and unfold, as well as the relation between online interactions and offline events (Scrivens et al. 2020), is lacking. These insights form the starting point for the ongoing research project Cultural Institutions and the Culture War (CICuW) that focuses the democratic implications of the pervasiveness of far-right digital discourse through an investigation of the digital environments where supporters of far-right ideas interact, share thoughts, construct collective identities, and coordinate actions toward libraries.

This chapter details the methods involved in conducting a pilot project of such an investigation, using online interactions as a sample explored through sentiment analysis and topic modeling. The sample in the original project was collected via web scraping, which is also described in the chapter. However, as the intended audience of this handbook is wide, both in terms of field of study and in languages spoken, the use case provided in the chapter will not make use of the Swedish language data explored in the pilot study and instead focus on creating a more general English language dataset.

The chapter begins with a short discussion regarding the use cases for the methods being introduced and some of the ethical considerations at play when working with online data collection, such as web scraping. The authors intend that readers should be able to follow along with their own project while reading the chapter, regardless of whether they make use of the same data sources as we do or decide to follow along with a data source of their own. However, the introductory sections must be read, understood, and considered before the work begins. Sections 2 to 5 make up what could be considered the pre-reading before the active participation begins in Section 6.

2 *Research problem and relevance to Digital Humanities*

Text analysis is by no means a new addition to the field of Digital Humanities, and could be considered the starting point for the field through the work of Roberto Busa and IBM during the early days (Jones 2016). The techniques included in text analysis are also often found at the intersection Digital Humanities share with Corpus Linguistics, although the aims of the two fields can often be seen as quite different. While linguists, as the name would suggest, are often concerned with language features and the development of language over time, digital humanists who take an interest in text analysis are instead more often concerned with the content of the language or what is being expressed. While these two groups often overlap, it is important to

note that the techniques discussed in this chapter fall into the latter category, and are aimed at producing content overviews of a collection of texts for further exploration.

This aim brings us to a second important thing to note about the contents of the chapter, namely that the project upon which it is based as well as the project introduced in the walkthrough components would fall under what is considered mixed-methods research. Mixed-methods research refers to an approach that mixes quantitative and qualitative research in order to arrive at its conclusions. For Digital Humanities projects, this often becomes the mode of combining perspectives from more traditional humanities with results derived from computational methodologies. For instance, in the CICuW project, quantitative overview methods such as sentiment analysis and topic modeling are used to create narrower focal points for qualitative methods, which in greater detail explore the discourse surrounding the selected topics and expressions of sentiment. While this chapter emphasizes the computational methods found at the first stages of exploration, the actual results of the pilot are intended to be formulated based on the latter qualitative component. As a whole, CICuW aims to offer a timely investigation into the democratic consequences of the pervasiveness of far-right discourse, focusing online information activities, on-site experiences, and their interconnectedness. In this way, digital methods commonly used withing the Digital Humanties are used in novel ways and combined with qualitative procedures to explore the nature of far-right digital discourse targeting cultural institutions and the relations between online far-right discourse and on-site hostilities.

In addition to text analysis, this chapter also includes web scraping as a mode of data collection. Briefly described, web scraping is a method for collecting materials from webpages and can be used for the collection of text, images, or any other component of a website. Many different technologies are available for web scraping, but they often rely on the structural element of the webpage in order to find the correct data to collect. This means that in order for web scraping to be efficiently applied, some basic know-how regarding webpages and web technologies is needed.

Furthermore, as we start to interact with the world wide web, privacy concerns and ethics become increasingly important. The online world provides us access to immense amounts of information of different types and a lot of that information is provided by the users of the web pages we are collecting information from. While data collection involving humans would normally involve informed consent and ethical approval, web scraping of online data is a bit more tricky. In this chapter, we will discuss the ethics of web scraping and online data collection based on the Association of Internet

Researchers (AoIR) ethics framework (franzke et al. 2020), but it is important to also consider the national context and institutional policy that might be relevant to your project.

3 *Research/use scenarios*

3.1 *Web scraping*

As mentioned in the introduction, web scraping is all about collecting information from the world wide web. This use case is often accomplished by interacting with a web browser through automation and then collecting the relevant components of whatever website or set of websites that one might be interested in. Automating the web browser interaction is the key to the usefulness of web scraping as a technique, as it allows for collection of large amounts of data quite quickly. While this chapter is focused on the use of web scraping for collecting text data, webpages also contain many other modalities which could be collected using a similar approach. In order to successfully perform data collection through the use of web scraping in a research context, the main components to be conscious of would be (1) the design of the target websites, (2) the type of automation involved, and (3) the ethical considerations at play in the given context. These will be briefly explained in regards to their relationship to web scraping in the following sections.

3.1.1 *Web sites*

At their core, web pages could be seen as explicitly structured documents containing different kinds of information and materials. The primary structure of a web page consists of HyperText Markup Language (HTML), which organizes all of the page's different types of content and provides the skeleton structure of where the pieces on when displayed in a browser. So, while we might see a visual representation of a web page projected for us in the browser, the actual web page consists of structured information. To have a quick look at how this might look, one can use the "Inspector" feature included in most web browsers and navigate to the "Elements" tab. An example of the Huminfra main landing page viewed through the "Inspector" in a Google Chrome browser is seen in Figure 1.

This structured information is key for the browser to be able to correctly display a page according to the designers instruction. Furthermore, it is this very structure that we hijack in order to successfully navigate the web page to collect relevant data when web scraping.

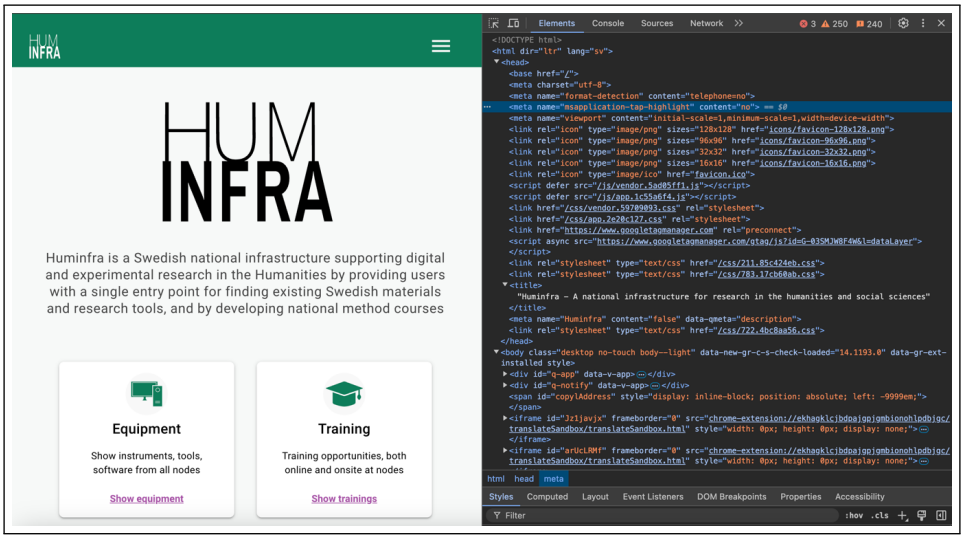


Figure 1: The Huminfra landing page Elements tab seen through the Inspector view of a Google Chrome browser.

The HTML structure that makes up the web site consists of elements, each enclosed by tags that describe whatever content is found inside. Returning to the example in Figure 1, we can see that the tag "`<title>`" encloses the text "Huminfra – a national infrastructure for research in the humanities and social sciences", which makes up the title element of the web site being displayed, for instance, in the browser tab where the page is open. We also see quite a few `<div>` tags, which are used to group elements into divisions. The different tags provide the hierarchy of the structured information, and make up what is known as the Document Object Model (DOM).

In order for our web scraper to be able to select the appropriate component of the web site to collect, we will need to specify a specific part of the DOM either by use of tags or named divisions. However, not all web sites are structured the same and the more diverse our intended target sites are the more complicated this selection process is. For the example in this chapter, we are going to collect information from one single web site. This means that we can assume that the structure will be the same for each web page will be largely the same in terms of its structure and element naming scheme.

3.1.2 *Automation*

Now that we have a better understanding of the way that web sites are structured, we can move on to discussing how we avoid collecting material manually from one place at the time. This is where automation shines, as it enables the collection of large amounts of data with minimal human intervention (although some is always needed). In addition to allowing us to avoid the manual portion of collecting the materials, automation also allows research projects to perform continous data collection or to set up a schedule of web scraping at intervalls. This allows research projects to scale up and to start engaging with the large amounts of materials often preffered for digital methods.

In the typical web scraping scenario, automation is applied through the use of a programming script drawing on the diverse multitude of resources available through programming languages such as Python or through the use of specialized software which is already configured to perform web scraping tasks. This chapter will make use of one such software tool, Octoparse (2024), which will be introduced in a later section. The automation needs to be able to perform the main tasks of identifying and collecting the appropriate data, and many solutions will also include the functionality to clean the data to a certain extent.

However, automation will often have to extend further than just these basic tasks. Due to the interactive nature of many web pages, web scraping tools will also often need to be able to interact with a site the way that a user would, using the browser interface to navigate the site, open up sub-pages, and occasionally entering information in order to search for materials or perform actions requiring a user to be logged in. This is especially true when interacting with web pages containing JavaScript, which creates dynamic web content which might not be loaded onto the page until certain interactions are performed. In programming, this is accomplished through the use of resources such as Selenium (Selenium 2024), but for software tools it is something that will need to be taken into account when selecting the right tool for the job.

While automation is highly beneficial and provides some of the main benefits of picking up web scraping as a skill, it also comes with a heightened level of responsibilty when compared to manual data collection. Web sites will often implement security measures such as login prompts or CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) in order to counter automated attacks or limit the amount of interactions made with their content at a given time. While performing automated interactions with a web site, it is our responsibility to ensure that we do so

without creating an issue for the owner of the web site or the host. This includes limiting the number of interactions done in line with what would be considered reasonable, making sure that we are not clogging up the access to the website while collecting our materials, and making sure that we are not collecting things that we should not collect. This is where the ethical considerations come into play.

3.1.3 *Ethical Considerations*

As web scrapers rely heavily on automation and since their purpose is essentially the autonomous collection of data and performance of tasks, they are sometimes discussed as robots. However, regardless of how automated and autonomous the collection procedure is, the ethical considerations must still be thouroughly engaged with by the researcher. Depending on the size and topic of the project the ethics involved can vary in complexity, but it is an area in which one should always prefer to have done too much rather than too little.

The Internet Research Ethics (IRE) is a document containing ethics guidelines for research involving materials and data taken from the web, and is continually being worked on by the Association of Internet Researchers. At the time of writing this chapter the IRE is in its third version, but readers after the year of publication should consult the current version for the most up-to-date information. While this chapter will rely on the IRE 3.0 (franzke et al. 2020) as its main document of ethical consideration, it is very important that any researcher engaging in web scraping also consult policies or regulations in place at their university or otherwise relevant to their immediate context.

Returning to the possible harm caused by web scraping, the considerations are often due to the nature or information available online. As the IRE 3.0 guidelines highlight, internet research must balance the benefits of scraping large datasets with the potential for harm to individuals whose data is being collected. Even in cases where data appears publicly accessible, such as on social media or forums, researchers should remain critical about the expectations of the users whose data they scrape. Users may not always be aware that their data is being collected for research, leading to a potential violation of their privacy. In addition, considerations must also be made regarding the potential use of the data being collected and the possibility of linking said data to the account from which it originated, and further on from there to an actual person. This is especially important when collecting potentially sensitive data from web pages on which interactions amongst vulnerable groups take place, such as Reddit sub-forums connected to sexual

orientation or religion. These examples would also be representative of the types of information that might be unethical collect according to local or national policy and regulation in certain cases.

In addition to the general guidelines of ethical internet research, web sites can also supply information about which types of information are allowed to be collected. These limitations are expressed in a robots.txt file, which should be available by simply adding "/robots.txt" after the main address of the site. The robots.txt is essentially a web standard that allows website administrators to communicate which parts of a site should not be accessed by automated robots. Respecting the instructions in a robots.txt file is a baseline ethical obligation for researchers as it represents the website owner's consent—or lack thereof—for their content to be collected. Ignoring robots.txt can lead to potential legal and ethical violations through accessing data that the site owner has explicitly marked as not to be collected. Researchers could see this as a form of informed consent in the digital realm, similar to the expectations outlined in the IRE 3.0 guidelines. However, it is very important to note that the ethical considerations that go into collecting online data for research go beyond the robots.txt, as research projects are expected to engage in a more comprehensive view of the ethics involved in their methods.

The ethical considerations do not end once the data has been collected, but must also extend to the way that the data is saved and stored. Depending on the sensitivity of the data, ethical policies and guidelines might dictate that certain precautions be taken in terms of anonymizaton and storage in order for confidentiality to be maintained. For these reasons projects collecting data through web scraping might also need to include use of a data mangement plan as a part of their design, especially so when submitted for formal ethics review. Norms underpinning principles for ethical research practices include "respect for persons, beneficence, and justice" (franzke et al. 2020: 4). These norms are important both when doing quantitative and qualitative studies online. In the context of qualitative studies, additional emphasis is often placed on reflexivity, the situatedness of human practices, and how to balance risks for participants against the potential value of the planned research (cf. Kozinets 2020). Such insights can also enrich method-driven studies often associated with the Digital Humanities (Hanell & Severson 2023).

3.2 *Text analysis*

Once we have completed web scraping and have a collection of texts to work with, we can start applying text analysis to explore the new dataset. As mentioned previously, the text analysis techniques discussed in this chapter

are aimed at producing content overviews which can serve as a starting point for further work. The end result of applying text analysis, in this specific use case, is therefore not an analysis of the materials but rather serve to make it possible to identify components of the materials for further qualitative assessment. Due to the overarching research design of the project which the chapter is based on, the output of the methods used are exploratory visualizations that can be used to explore the data.

Text analysis is an umbrella term which encompasses a myriad of different techniques, often with the purposes of identifying patterns in textual data, creating overview of different features in the text, or to annotate texts according to whatever classification might be relevant for a given project. In this chapter, we focus on topic modelling, which identifies patterns of topics in a text collection, and sentiment analysis, which annotates texts based on approximations of expressed sentiment.

3.2.1 *Zero-shot sentiment analysis with BERT*

Sentiment analysis, as mentioned in the previous paragraph, is all about figuring out the sentiment expressed in a text. Sentiment in this context is understood along the lines of emotional tone, and is often categorized as being positive (I love this!), negative (I hate this...), or neutral. Do note the missing example for the last category, as it often ends up being the bin category for texts in which neither positive or negative sentiments are found. While this chapter is going to focus on zero-shot sentiment analysis, as used in the CICuW project, we are going to start by discussing two other approaches to the concept, lexicon-based and supervised machine learning-based, that will serve to highlight the specific method choice made for the project.

Lexicon-based sentiment analysis, as the name suggests, relies on a lexicon comprised of sentiment-carrying terms which are accumulatively identified in the text. The number of positive terms and negative terms are added up, and the total is considered the sentiment of the text. The technique does not only count the words, of course, but instead relies on sentiment weights being attached to each term in the lexicon, identifying them as more or less negative or positive. In addition, more advanced lexicon-based methods such as VADER (Hutto & Gilbert 2014) also take the context of the term into consideraton and accounts for expressions of intensity as well as negation. Considerations when applying a lexicon-based approach should include how well the lexicon fits with the context to which it is applied. For instance, VADER (Valence Aware Dictionary and sEntiment Reasoner) was formulated specifically for microblog-like contexts, such as social media.

Supervised machine learning-based sentiment analysis may sound intimidating, but it is actually quite easy to apply once the pre-requisite components are assembled. The approach relies on using a machine-learning model to identify sentiments based on pre-labelled examples, the so-called training data, and then apply the same labels to previously unseen texts. Unlike lexicon-based methods, which rely on pre-defined word lists, machine learning approaches can capture more nuanced relationships in the training data by using algorithms such as support vector machines, logistic regression, or neural networks. The key consideration when using supervised machine learning for sentiment analysis is the quality and size of the training data, as the model's performance relies on having a representative dataset to learn from. However, representative is not a general quality in this case, and must be understood contextually. Training a model on emails and the applying it to YouTube comments is unlikely to produce viable results.

Zero-shot sentiment analysis makes use of large models, such as BERT (Bidirectional Encoder Representations from Transformers, Devlin et al. 2019), to identify expressed sentiment without the use of a specific lexicon or a training dataset. This relies in part on the massive amounts of data that have already been used to train the model, which allows it to infer the sentiment from an unseen text based on the labels provided even when explicit examples have not been provided. The model is prompted with a set of labels, for instance Positive, Negative or Neutral, and then evaluates the new text according to what the model already has available for, or "knows" about, each of the provided labels.

When compared to traditional supervised models, zero-shot sentiment analysis stands out by removing the dependence on labeled datasets, leading to a higher potential scalability and adaptability to new contexts. This can be an especially important feature of the technique when one is conducting a project where the available data is limited, or where there are no resources for manually labelling a large enough quantity of training data. However, the performance of zero-shot models may not always match that of supervised methods fine-tuned for a specific task if large amounts of labeled data are available, as supervised models can optimize their predictions based on that dataset.

3.2.2 *Topic modelling with LDA*

While there are many different techniques available for applying topic modelling, they all have in common that the end goal is to identify latent topics or themes. What is meant by latent here is that the topics or themes do not have to be explicitly defined before the technique is applied, but are rather

created by the method by applying some kind of conceptualization of what a topic should be to the data. Oftentime, we consider a topic to be made up of content words that co-occur with one another frequently within the collected texts.

One of the more popular methods relying on this rationale is Latent Dirichlet Allocation (LDA), introduced by Blei et al. (2003). The method represents a perspective on topics in texts which assumes that every text is in fact made up of multiple topics which are distributed across the texts in a collection. Since topics are created dynamically based on the content of the texts in the collection and since LDA produces the distribution of topics amongst the texts, the results are very useful for identifying texts that are mainly on one topic but might include other topics to a lesser extent. This is especially useful when applying topic modelling as a mode of exploration, as we are doing here.

In order to apply LDA we must first specify certain variables, such as the number of topics to be extracted and the number of terms to be included within each topic. Here we have two separate options: Qualitative assessment of the results or quantitative optimization. The project this chapter is based on decided on qualtitative assessment due to the texts in the collection being short in length and due to the specific interests of the the later qualitative analysis. In practice, this meant iterating over the collection with different settings and deciding on the number of topics and terms that ended up producing the most relevant results in terms of the content described in each topic.

For quantitative optimization of LDA, the "elbow method" is often applied. The elbow method involves running the LDA model with different numbers of topics and plotting a metric representing the quality of the topic, such as a perplexity or coherence score, against the number of topics. As the number of topics increases, the model improves, but after a certain point, the improvement slows down, forming the eponymous "elbow" in the plot. This would mean that the number of topics used just before the bend is the most optimal.

However, before we start topic modelling our text collection we must spend some time on pre-processing the texts. Since the LDA method relies on frequent co-occurrence of terms, we want to make sure that the terms included in each topic are relevant ones, previously discussed as "content words". From an extremely reductive perspective, we can split language into two categories: content words which convey content, and function words which perform a syntactical function within the given grammatical system. Examples of the latter category would be words like conjunctions and articles. The pre-processing involved in successfully applying co-occurrence-based

topic modelling will often involve stopword removal, which removes function words from the text. This is based on a stopword list, which contains all of the words to be removed. It is important to be aware of what is being removed from the text, as different stopword lists contain different words and vary greatly in length.

In addition to stopword removal, pre-processing can also include techniques to condense the text. The purpose here is to remove different variants of the same word resulting in more correct frequencies for the concepts represented by the words and less noise in the results. This could for instance be done through lemmatization, which replaces every variant of a word with its dictionary form. Finally, punctuation is often also removed. Here it is important to be aware of all the changes made, as the pre-processed text to which the method is finally applied is quite different in its form than the original text first collected. It is also important to keep a keen eye on whether a method is being fed the pre-processed text or the original, as it can greatly influence the outcome.

3.3 *Exploratory information visualization*

While seen here from a narrow view in its role as the output of the text analysis components of the project, it is important to be aware of the fact that information visualization is a field of research in its own right. The functionality of different types of information visualization of course depends on use case and purpose, but for our approach we are mainly interested in interactive visualizations that allow us to overview the large dataset and find items that are relevant for our manual analysis. Shneiderman (1996) is often quoted as describing the design rationale for such interactive visualizations as: "Overview first, zoom and filter, and then details on-demand", which provides an idea of both the needed functionality as well as guidelines for intended usage. However, we also have to consider the types of information we want to visualize.

In the case of the project on which this chapter is based, we are mainly interested in visualizing numerical values derived from sentiment analysis (negative to positive) and topic modelling (distribution of topics across different items). However, the latter must also include a representation of the terms included in each topic for visualization legibility. The individual texts make up the relevant granularity of our visualizations, as the main purpose of the visualizations produced will be to identify relevant texts within the sample. Considering this with Shneiderman's mantra in mind, we will need exploratory visualizations that allow for an overview of the results from the sentiment analysis and topic modelling, zoom and filtering

including the selection of individual texts, and details that will allow for the selection of individual texts for further study.

4 *Resource / tool description*

4.1 *Low-Code and No-Code Tools*

The broad group of tools referred to as low-code or no-code should in the context discussed here be understood as digital environments which allow for the user to create workflows and applications with, as the terms suggest, little or no actual programming needing to take place. The environments often utilize a drag-and-drop interface which allows the user to combine pre-built components and functionalities in a visual workflow. These tools are intended to maintain a balance between transparency and ease of use. While not involving programming, the use of low-code and no-code tools still involves a level of computational thinking and problem decomposition which makes them a good point of entry to computational text analysis methods (Tyrkkö & Ihrmark 2024).

Computational thinking, as defined by (Wing 2006), is a problem-solving process that involves various cognitive skills, such as abstraction, algorithmic thinking, decomposition, and pattern recognition. Essentially, it is intended as a way of tackling a problem by breaking it down into its components and then designing a workflow that answers to the identified components. This is also the approach needed for work with low-code and no-code tools, as the conceptual and abstract work of designing the process remains just as important as it would when programming, just with different building blocks.

These types of tools also introduce concerns regarding the use of closed-box approaches within research. The term closed-box refers here to the way that fairly complicated methods can now be employed with relative ease, without requiring the researcher to necessarily comprehend exactly how the methods work or arrive at their results. While low-code and no-code tools allow us to experiment with and apply new and interesting methods, the responsibility of understanding what is happening is still on the researchers. This also includes understanding the biases, assumptions, and limitations of every method as applied in whatever context is being engaged with.

In this chapter, the context for application is both data collection and analysis. For data collection, considerations need to be made regarding what is collected and how it is stored, as discussed in 3.1.3. For the text analysis workflow, the focus is more on understanding the components of the

workflow and the rationale of the methods in order to produce reliable and comprehensible results. The first tool introduced is Octoparse, which will be used for web scraping. This is followed by KNIME (Konstanz Information Miner, Berthold et al. 2009), which will be used for the text analysis.

4.2 *Octoparse 8*

Octoparse 8 is intended as a no-code tool that allows for automated data collection from online sources. In other words, it is a web scraper with a point-and-click interface that enables data collection without a need for programming. The tool allows for collection from both static web pages and dynamic pages with JavaScript elements. It also offers functionality for storing login information and automating access that might be behind a password barrier, as well as support for browser interactions that can allow for navigating the site or closing pop-ups. The tool shows the provided URL in a browser window, and allows the user to either select elements for collection manually or through the auto-detect feature.

The auto-detect feature, which maps the page elements and generates the parsing rules needed to identify them within a group of pages with similar structures, is one of the main functionalities that makes Octoparse a great entry point into web scraping. This skips some of the more engaged steps in conducting manual scraping, such as reading the page's HTML and figuring out the connection between its tags and the relevant content. However, the feature is not flawless and some fine-tuning is to be expected in most cases, meaning that some understanding of the web page structure is still needed. Figure 2 showcases the tool browser after having applied the auto-detect feature to the IMDb (Internet Movie Database) page for the 1999 classic "Office Space", resulting in a dataset consisting of the actors in the film alongside their roles. The resulting dataset is showcased at the bottom of the tool window.

Like many no-code tools, Octoparse works based on a workflow. The workflow dictates the steps of navigating the site, entering information, locating relevant information, and collecting it. The workflow also allows for the creation of loops as a part of the process, for instance enabling collection to be possible iteratively based on a list of URLs on the site. In Figure 3, the workflow for a scraping task based on a list of URLs is shown. For common tasks, such as data collection from social media, the tool provides task templates.

Whether one wants to compile a list of specific URLs for scraping or collect data based on a list of links already available on a website depends on goals of the project. For the CICuW project use-case, the target website's

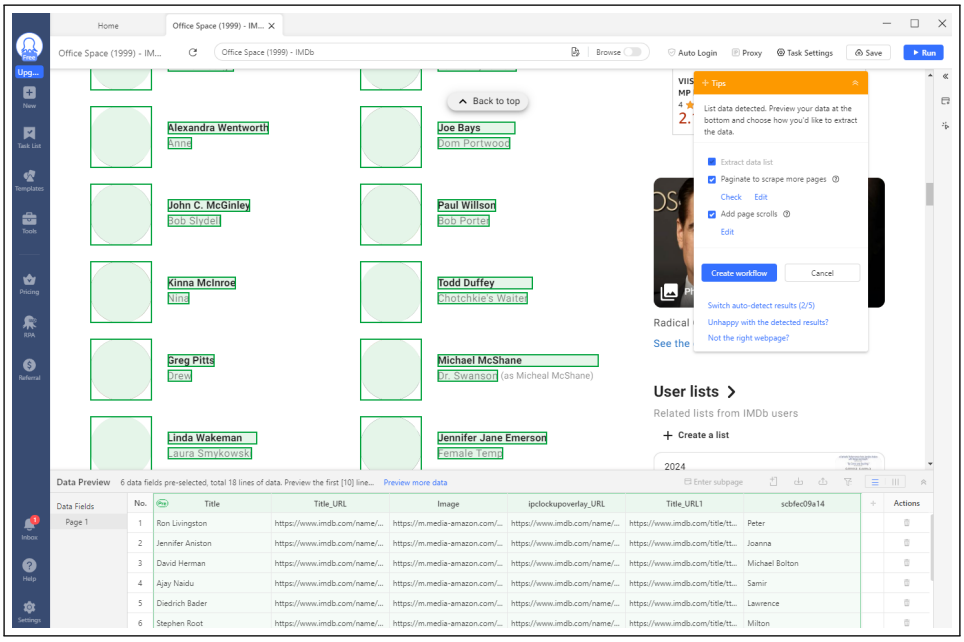


Figure 2: Octoparse interface showcasing extracted elements from IMDb page for "Office Space" (1999).

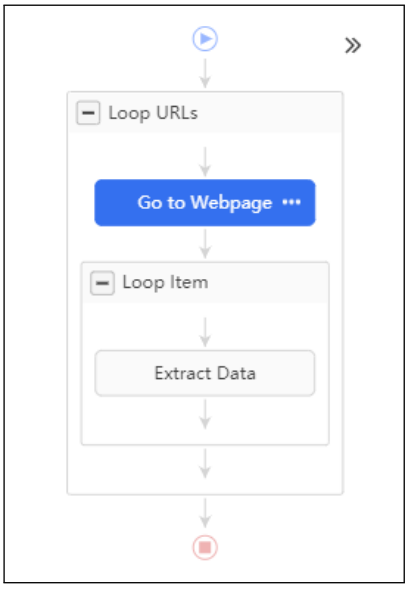


Figure 3: Octoparse workflow.

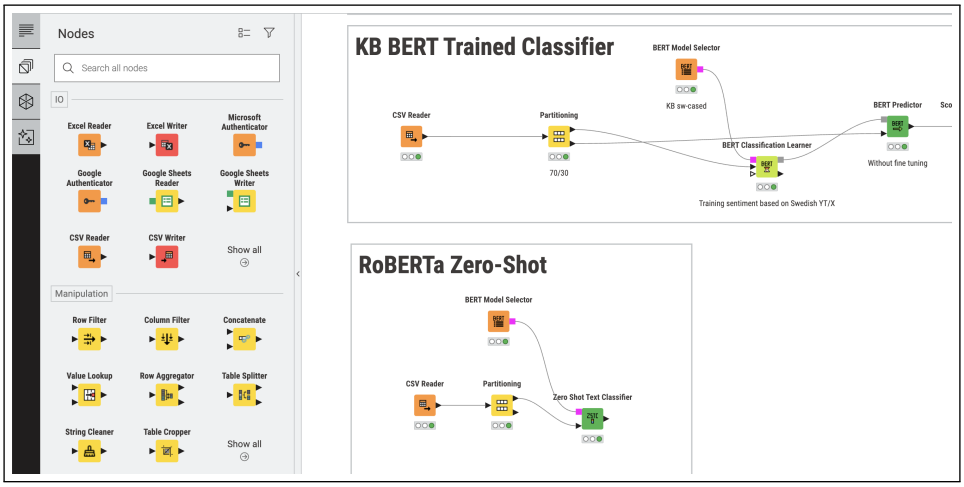


Figure 4: Workflow with node sidebar shown on the left.

search function was used to find the relevant materials, which created a list of links on the page to work based on. However, for other collection targets, one might prefer to manually curate a list of URLs for greater control over what is included in the resulting dataset.

4.3 *KNIME*

KNIME is a powerful, open-source, low-code tool that offers many different types of analysis. In addition, KNIME is also capable of producing interactive visualizations as a part of the workflow, meaning that it is a good fit for cases where the visualizations are used to filter data for further work. It is worth noting that KNIME is defined as a low-code tool, as it offers the opportunity to include programming scripts in the workflow to further extend its functionalities. However, no programming will be necessary for the uses showcased in this chapter.

The core of KNIME is its visual workflow editor, which is used to string prebuilt functions together for process being conducted. The functions are represented by nodes, which visually indicate the placement of the function in the workflow. The workflow is visualized on the canvas component of the tool's interface, into which nodes are dragged and dropped from the sidebar. The order of the workflow is determined by how the nodes are connected. The example workflow for this chapter is shown in Figure 4. We recommend downloading the workflow for use in the walkthrough via the

KNIME Hub[1], as importing it to your local KNIME installation will also prompt the installation of the needed extensions.

In a KNIME workflow, the data flows along the workflow through the connections made between the nodes. Nodes operate on a "Traffic Light" signal system, where nodes with a red marker beneath them are in need of either an input or configuration, yellow markers indicate a node ready to execute, and green markers indicate nodes that have been successfully executed. The way we work with a KNIME workflow is first designing the process, then assembling the nodes, and finally running them in our decided order so that we can perform whatever task we have in mind with the input data.

When KNIME is first installed, it includes some common or general prebuilt functions. In order for the tool to be useful for specific tasks, such as text analysis, the available nodes must be expanded through the installation of extensions. Extensions are available via the KNIME Hub, and can be installed via the tool or via web browser. The KNIME hub hosts both official KNIME extensions and community extensions, which are created by the users of the tool. Some specialized extensions are not available via the hub, and need a more engaged process which includes manually specifying the correct source repository in order to be installed.

The extensions highlight both a main drawback and a main benefit of KNIME as a tool. The use of extensions allows for the possible use cases for KNIME to be continuously expanded according to the needs of the user community, but it also creates a situation where problems can arise from changes in one extension creating compatibility errors with other extensions that are developed separately but are often used in combination. Furthermore, the functions, processes and tasks afforded by the available nodes and extensions at any given time provide the limit of KNIME's usability, with uptake often lagging a bit behind cutting-edge tools and approaches possible via programming.

In summary, KNIME provides an entry point to computational methods with a significantly lower barrier when compared to programming languages such as Python, but it also lags behind the developments in methods available through such approaches. When used without appropriate consideration, low-code tools like KNIME can also exacerbate closed-box issues. However, the use of the canvas to explicitly connect nodes in order to create a workflow where the data can be inspected and interacted with at any point during the process allows for tasks to be engaged with in a way that supports computational thinking and can help in avoiding haphazard use of methods in cases where the user remains engaged with the different components involved.

1 https://hub.knime.com/s/qDeklEsCiNtekxfv

5 *Case study: Cultural Institutions and the Culture War*

As previously described, public libraries in Sweden have become a focal point in the growing cultural conflict connected to the rise of the far-right. Digital forums and social media play a significant role in fueling these ideologically driven confrontations, which may threaten the libraries' democratic mission. However, there is limited understanding of how these online threats evolve and their connection to real-world events, highlighting the need for further investigation into digital spaces where far-right supporters organize and coordinate actions against libraries. The CICuW project aims to investigate this connection between online and offline interactions between the far-right and the public libraries through analysis of online posts on far-right platforms and offline incidents at public libraries.

The project consists of three sub-projects, of which one is engaged with the types of methods discussed in the current chapter. However, it is important to note this as a component of the project, as it ties into the mixed-methods context in which it takes place. This essentially means that the project as a whole involves both qualitative and quantitative methods. The three interconnected sub-projects involved in the overarching CICuW project are focused on three different components of the research context: digital information activities in far-right spaces, on-site experiences at cultural institutions, and finally the interconnections between the two. The current chapter is based on a pilot study taking place within the first sub-project, dealing with online information activities.

The pilot study was conducted in order to explore the use case for different methods in a smaller sample of relevant materials. The data for the study was collected from an online news platform connected to the Swedish far-right. Articles about public libraries were collected using the site's internal search functionality, and both the articles themselves and the user comments on the individual articles were collected. The different kinds of data collected allowed for the analysis to be conducted on two levels: articles and comments.

The articles were analyzed for topic and sentiments in order to provide a starting point for the topic and sentiment of the exchanges on the platform, essentially acting as the tone-setter for the interaction. In addition, the articles could also be seen as providing both a spatial and temporal grounding for the exchanges in the comments. The comments, on the other hand, could be seen as representing the reaction to the article, and as indicating which topics or sentiments were picked up for further discussion. The number of comments is, of course, also indicative of the level of engagement with a given topic.

Once analyzed using topic modelling and sentiment analysis, the data was further engaged with through qualitative content analysis (Hsieh & Shannon 2005). In practice, this means that the results of the topic analysis was engaged with in order to understand what each of the resultant term collections could refer to in the context of the sample, and a close-reading of relevant pieces was then conducted in order to verify the interpretation. This qualitative assessment also provides the conclusions of the study, as the deeper understanding of the material needed to answer the research questions is produced. This highlights the exploratory use of the digital methods in the context of CICuW, as they serve to filter the data and produce smaller samples to be engaged with directly. While the data we will engage with in the walkthrough is quite different both in nature and language, the process and steps taken are the same.

6 *Example walkthrough using IMDb*

For our example walkthrough, we will collect the descriptions of the top 100 rated movies on the Internet Movie Database (IMDb). We will also collect some additional information, such as user ranking and year of release. However, before we start deciding on exactly what we want to collect, we must first have a look at the site's robots.txt (Figure 5) in order to see if any of the specific information we are interested in is disallowed.

The disallowed items on the page seem to have mainly to do with either dynamically generated content, such as "on this day" or searches, and user-specific content. We can therefore carry on with our intended collection as planned.

The basis for our dataset is going to be the list of movies provided via the "Top 100" link on the website's opening page. The link takes us to a list of films, sorted on user score. We will be using Octoparse to iterate through the list by essentially clicking the URLs in the order provided to us, and collect information from each of the pages. This way we do not have to compile a list of URLs ourselves, and can instead rely on the structure already in place on the site.

6.1 *Collecting the Data with Octoparse 8*

We start by opening Octoparse and clicking "Custom Task" in the hover menu that appears when we hover "New" in the menu on the left. This will prom,pt us to supply either a .csv file with URLs or to input the URL into the text box. Since we will be working with a single URL, we simply copy

```
# robots.txt for https://www.imdb.com properties
User-agent: *
Disallow: /OnThisDay
Disallow: /ads/
Disallow: /ap/
Disallow: /mymovies/
Disallow: /r/
Disallow: /register
Disallow: /registration/
Disallow: /search/name-text
Disallow: /search/title-text
Disallow: /find
Disallow: /find$
Disallow: /find/
Disallow: /tvschedule
Disallow: /updates
Disallow: /watch/_ajax/option
Disallow: /_json/video/mon
Disallow: /_json/getAdsForMediaViewer/
Disallow: /list/ls*/_ajax
Disallow: /list/ls*/export
Disallow: /*/*/rg*/mediaviewer/rm*/tr
Disallow: /*/rg*/mediaviewer/rm*/tr
Disallow: /*/mediaviewer/*/tr
Disallow: /title/tt*/mediaviewer/rm*/tr
Disallow: /name/nm*/mediaviewer/rm*/tr
Disallow: /gallery/rg*/mediaviewer/rm*/tr
Disallow: /tr/
Disallow: /title/tt*/watchoptions
Disallow: /search/title/?title_type=feature,tv_movie,tv_miniseries,documentary,short,video,tv_short&release_date=,202
Disallow: /name/nm*/filmotype/*
Disallow: /user/ur*/ratings
Disallow: /user/ur*/checkins
Disallow: /_json/*

User-agent: Baiduspider
Disallow: /list/*
Disallow: /user/*

User-agent: bingbot
Disallow: /showtimes/location/*
```

Figure 5: IMDb robots.txt document.

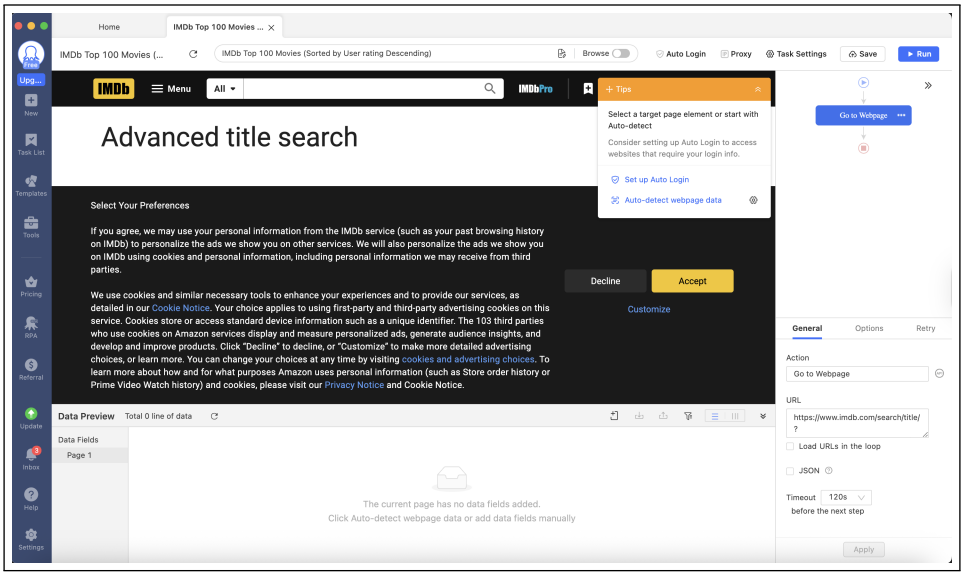


Figure 6: Link opened in the Octoparse browser.



Figure 7: Results of auto-detect feature.

our link into the box. When then click “Save”, which opens up the URL in the Octoparse browser (Figure 6)

As we can see, there is a pop-up in the upper right corner which asks if we want to “Auto-detect webpage data”. Clicking this will start a scan of the page where recurring information elements are detected. Since we are working with a list, the likelihood of it being auto-detected is quite high, and we can make use of the function in order to generate our initial list of URLs. However, once running the scan, the preview indicates some issues. We have only collected information for 50 items, and the title for each film also includes its ranking in the list (Figure 7).

We therefore click the “cancel” button in the auto-detect window, and inspect the site manually. The issue seems to be the inclusion of a button for expanding the list at the bottom of the page, named “50 more”, which is why the web scraper only provided us with 50 lines of data. Clicking the button in the Octoparse browser brings up a dialog box where we can choose the action we want to perform (Figure 8).

Selecting “Click button” adds a “Click item” action to our workflow, and ensures that the list is expanded when the task is later automated. We can now re-try the auto-detection, and see if this solves the issue of the missing movies. Once we have verified that this is the case, we click the “Create workflow” option provided by the dialogue box. The workflow is presented to us on the right side of the screen, and shows that we have now created a loop which collects the information of all items in the list, after having

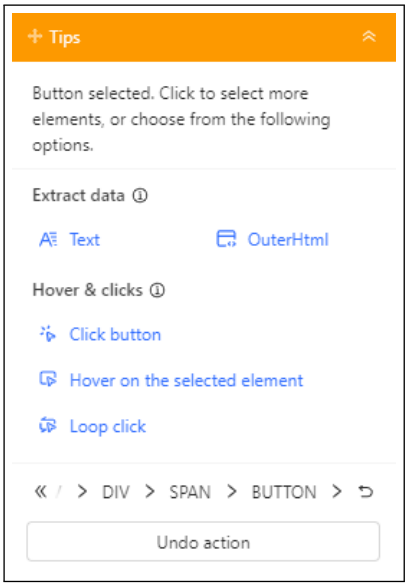


Figure 8: Dialog box resulting from clicking button.

first pressed the "50 more" button. We can now either be happy with the current state of the dataset, or we could click "Select subpage URL" in order to iterate through all of the pages linked in the original list which allows us to select a column in the collected data containing the URL for the subpages.

The latter option will open the first URL in the list, and allows us to select which elements we want to collect from the page. Hovering an element will highlight it in blue, enabling us to see what is included. If we start by clicking the element containing the title, the dialog box asking us which action to perform will appear again. Choosing "Text" will collect the text content of the element, i.e. the title (Figure 9). We then repeat this for the elements containing the short description, the year of release, and the user ranking.

The preview data should now show our selected pieces of information. We can also rename our columns here, to make sure that we know what ends up where in the final output. Once we are happy with the preview data, we click the "Run" button in the top right corner of the window. This will provide us with the different option available for running the task. For this example, we can select "Standard Mode" on our own device. Doing so will bring up a window that allows us to inspect the data as it is being collected (Figure 10). This helps us see if any issues occur without having to run the entire process first.

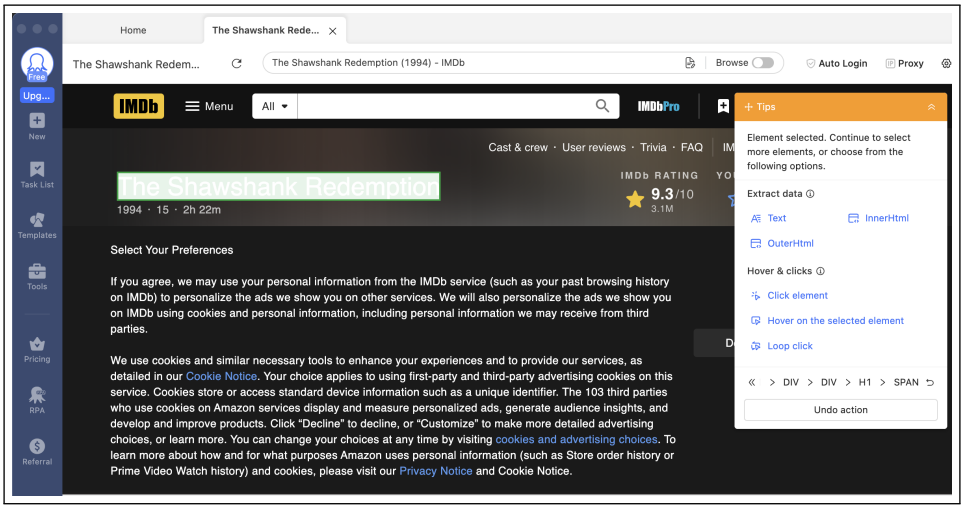


Figure 9: Title selected and text content extracted.

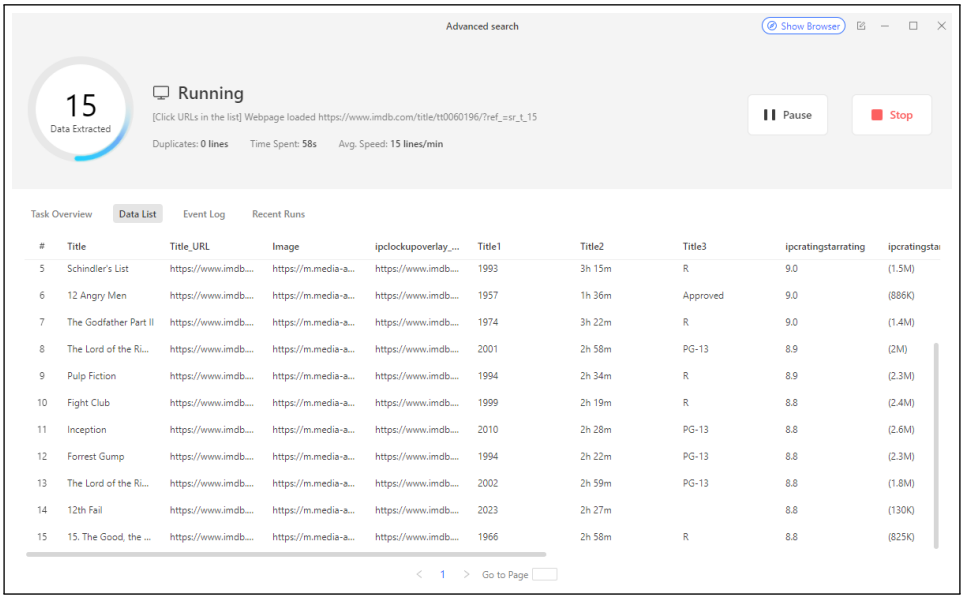


Figure 10: Window allowing for inspection of running task.

Once finished, we can export our data in our format of choice. For this walkthrough, we will choose Excel so that we can easily inspect the results. As we can see, we have created a few duplicate columns by selecting information that was caught by auto-detect in the initial list in our manual selection of elements on each individual page.

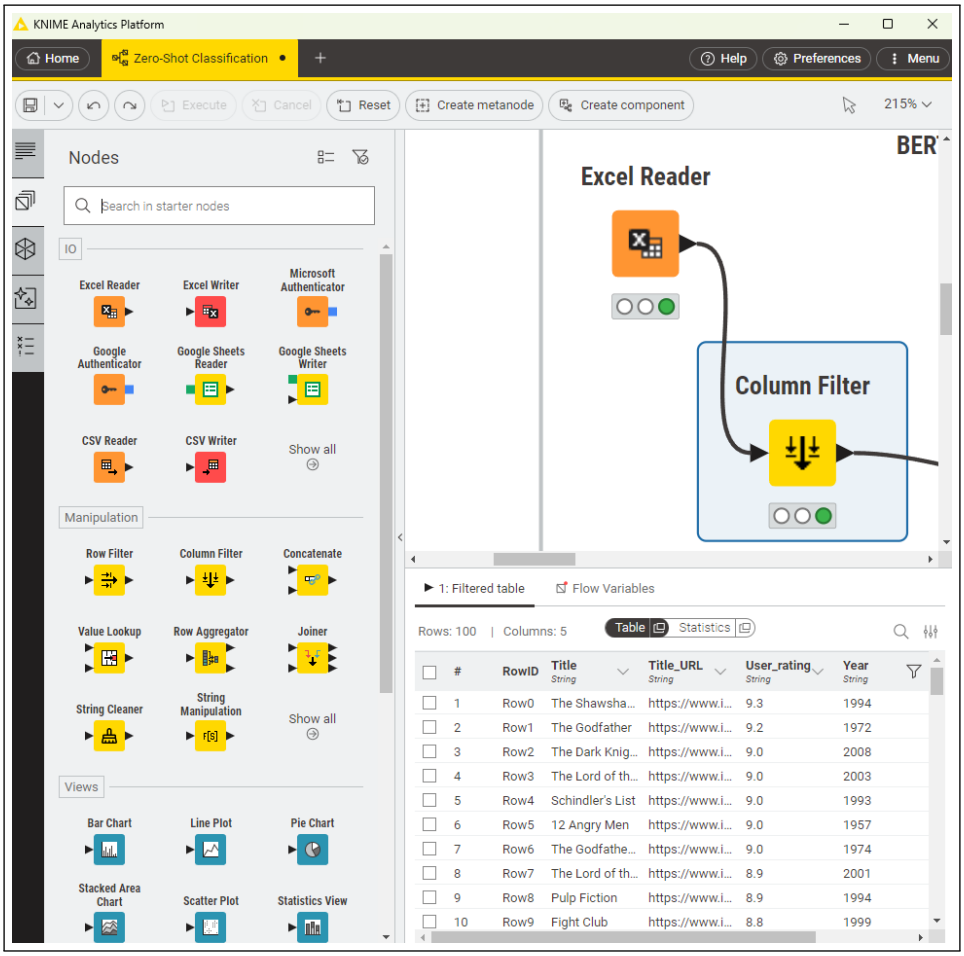


Figure 11: Making the IMDb dataset available in KNIME and removing irrelevant columns.

6.2 *Analysing the Data with KNIME*

We start by opening up a new KNIME workflow, and placing down an "Excel Reader" node so that we can make use of our new IMDb dataset. We then connect our Excel reader to a "Column Filter" node, which will allow us to remove columns that we are not interested in (Figure 11). With these in place, our data is now ready for analysis with KNIME.

We are going to start by performing zero-shot sentiment analysis using a BERT model. In order for us to be able to make use of the nodes related to BERT models, we need to install the Redfield BERT extension for KNIME. Once the extension is installed, we select the "BERT Model Selector", which

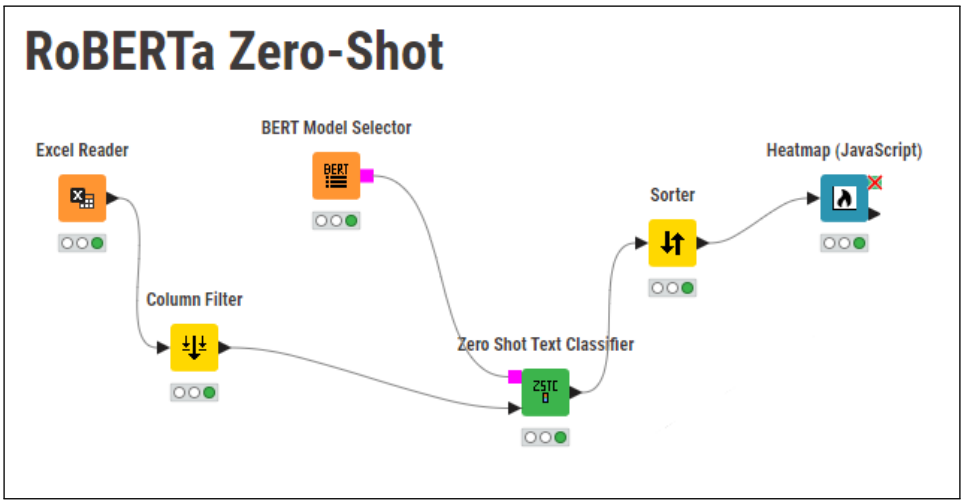


Figure 12: Zero-shot sentiment analysis workflow.

will allow us to select the specific model we want to use. The node allows for the use of a model from HuggingFace, which is a large repository of different models and datasets. In our case, we are going to make use of a model called RoBERTa[2]. We then connect both the model selector and the column filter to a "Zero Shot Text Classifier" node (Figure 12).

Note that the inputs of the "Zero Shot Text Classifier" node are keyed, with one input for the data table and one for the selected model. By right-clicking the "Zero Shot Text Classifier" node, we are able to configure the specific task we want to carry out. In our case, this means selecting the column corresponding to the text we want to analyse, setting up the labels for the classification task, and providing a hypothesis template to indicate the nature of the task. The labels should function as answer alternatives for the question posed by the hypothesis. The configuration for sentiment analysis in the example workflow is shown in Figure 13.

This will add a set of columns to our dataset which will indicate the predicted sentiment, as well as the probability for said prediction of Positive or Negative. We then pass the data along to a "Sorter" node, which allows us to sort the data table according to the predictions. In order to get get a visual overview of the results, we finish the segment of the workflow with a heatmap visualization. In Figure 14, the heatmap is showcasing the probabilities of a Positive label for each of the films.

2 https://huggingface.co/typeform/roberta-large-mnli

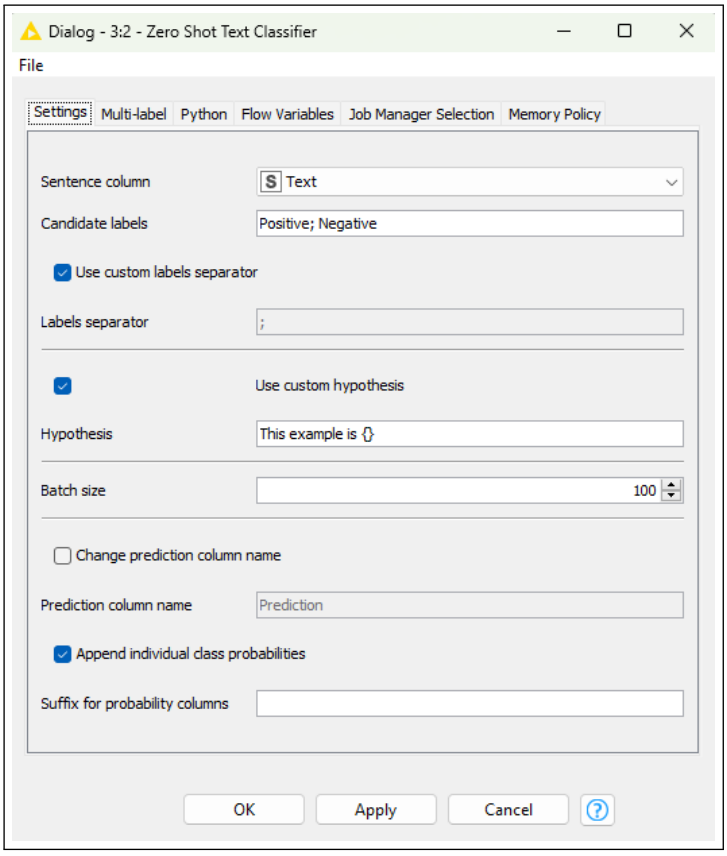


Figure 13: Configuration of "Zero Shot Text Classifier" node for sentiment analysis.

Having performed the sentiment analysis, we then connect our data to the preprocessing stage of the LDA topic modelling section of the workflow. As previously mentioned, preprocessing is a key component for topic modelling, and it is important to consider each step in the context of each project's goals and aims. The preprocessing stage in the example workflow begins by changing the data format, so that it fits with the expected input of the rest of the process.

By using the "Strings to Document" node, we reformat each line of our dataset to a text document. These documents are then preprocessed to remove punctuation, lowercase all words, and remove stop words. Finally, the texts are lemmatized so that each word is converted into its dictionary form. The lemmatization makes use of a SpaCy model (Honnibal & Montani 2017), which is imported through a model selector node similar to the BERT

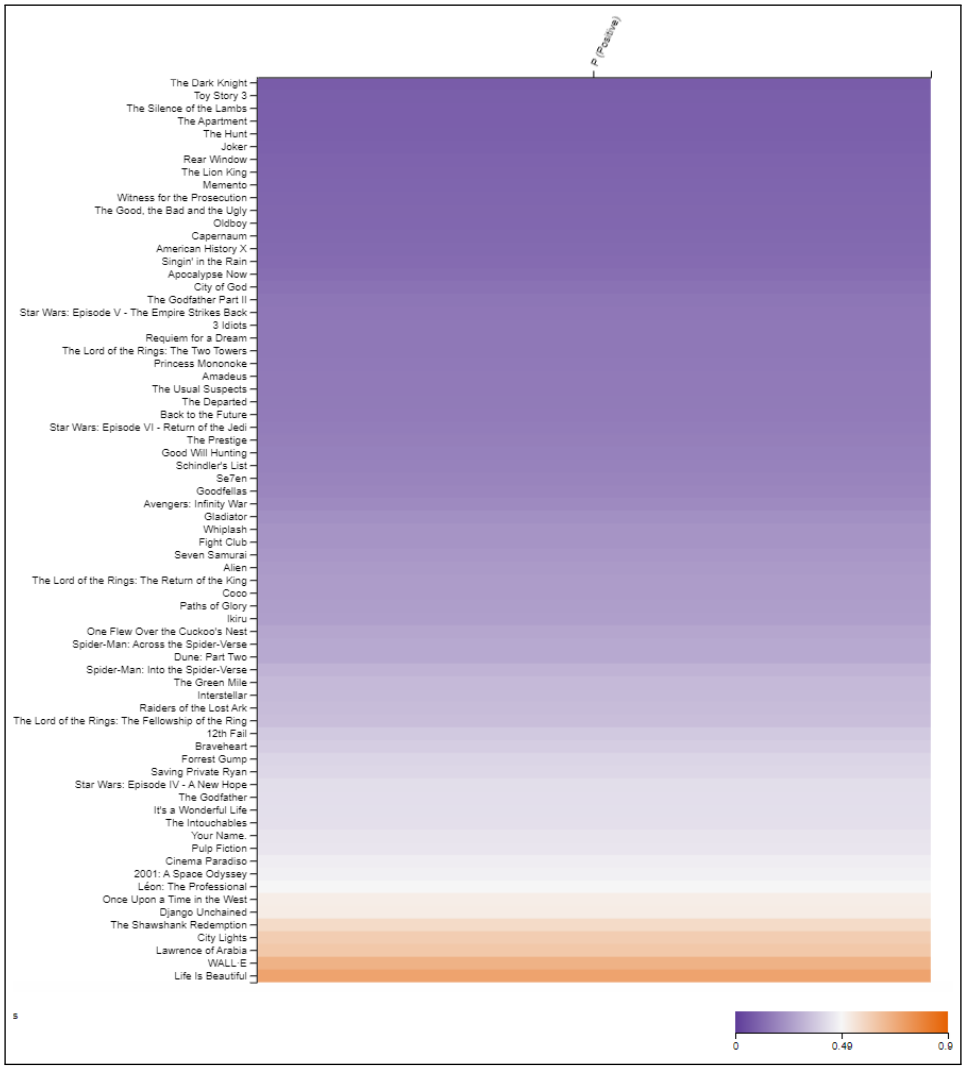


Figure 14: Heatmap showing the probabilities of a Positive label for the IMDb top 100.

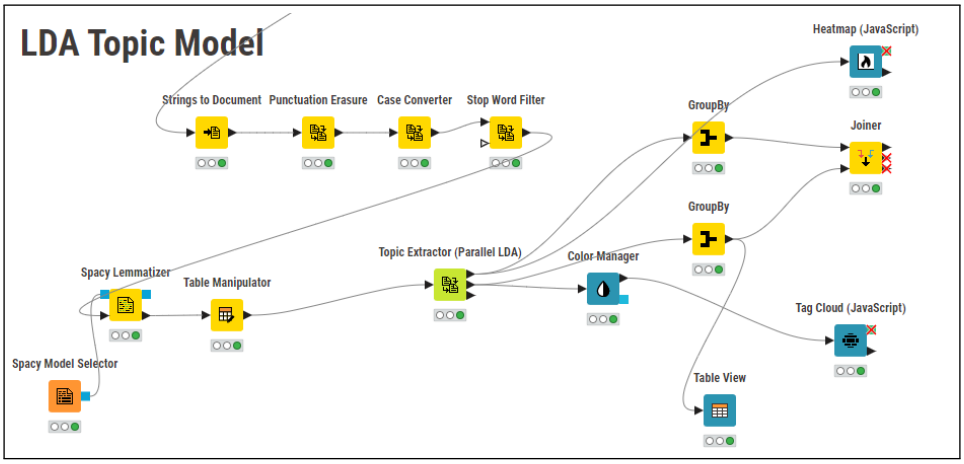


Figure 15: Preprocessing before topic modelling.

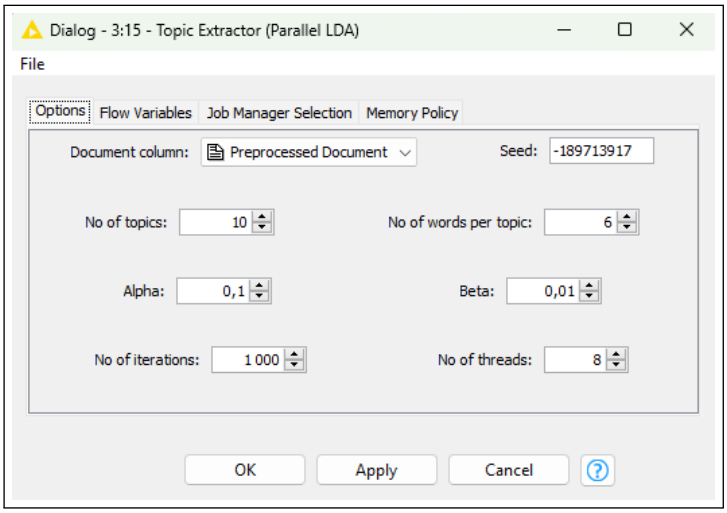


Figure 16: Configuration options for the "Topic Extractor (Parallel LDA)" node.

model selector used for sentiment analysis (Figure 15). As SpaCy models are language-specific, as well as text type-specific to a certain extent, we have to consider the nature of our sample when making this decision. The preprocessed documents are then fed into the "Topic Extractor (Parallel LDA)" node for the actual topic modelling.

In order to run the "Topic Extractor" node, we need to perform some configuration steps (Figure 16). As previously discussed, the LDA method of topic modelling requires us to set a number of topics and the number of

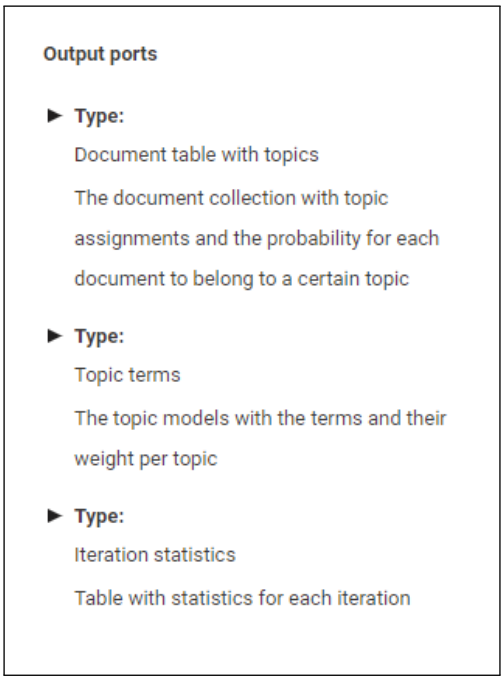


Figure 17: The outputs of the "Topic Extraction (Parallel LDA)" node.

terms per topic before being applied. For our example workflow, the number of topics have been set to 10 and the number of terms to 6. However, feel free to experiment with these settings in order to get as meaningful a result as possible. It is also important to note the column selector at the top of configuration window, as this is where we indicate the column in which we have saved our preprocessed documents.

Once the the node has been executed, we can start looking at the topics we have produced. Here it is important to be aware of the different outputs of the node (Figure 17). The three outputs of the node each produce a different set of information from the process, such as a table of assigned topics with correlating documents, information about the terms included in each topic, and the statistics of the actual extraction process. The input and output specifications are found on the left of the window whenever a node is selected.

These different outputs are each going to inform our understanding of the dataset in a different way. Moving from top to bottom, we start by dealing with the topics assigned for each document. As the internal topic distribution for each document is described numerically, we can visualize

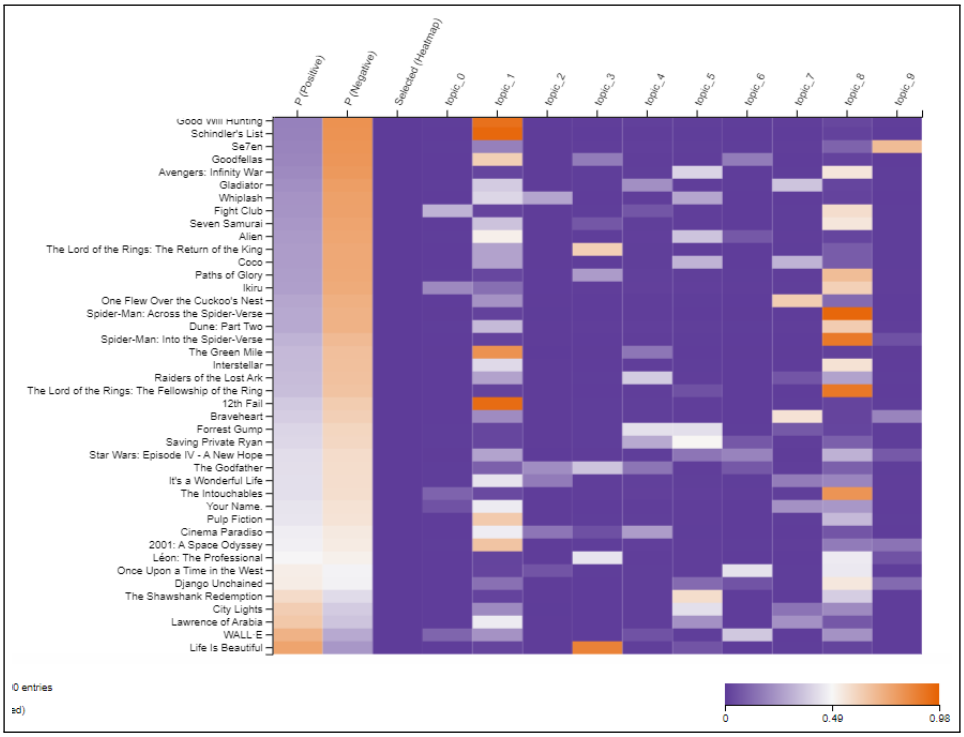


Figure 18: Topic distribution in the IMDb top 100 dataset visualized as a heatmap.

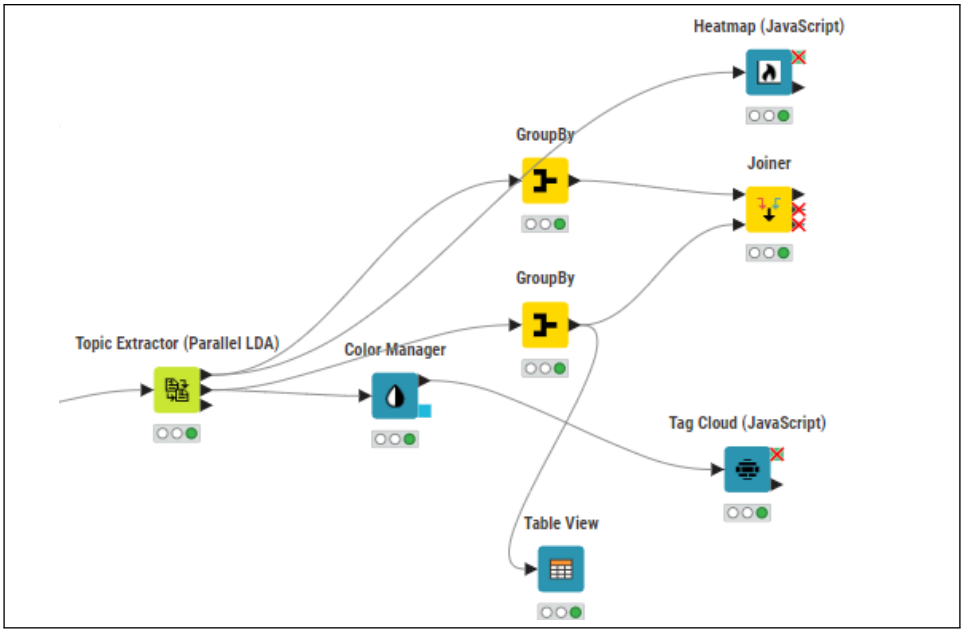


Figure 19: Workflow making use of the outputs of the "Topic Extraction (Parallel LDA)" node to create visualizations and combined data tables.

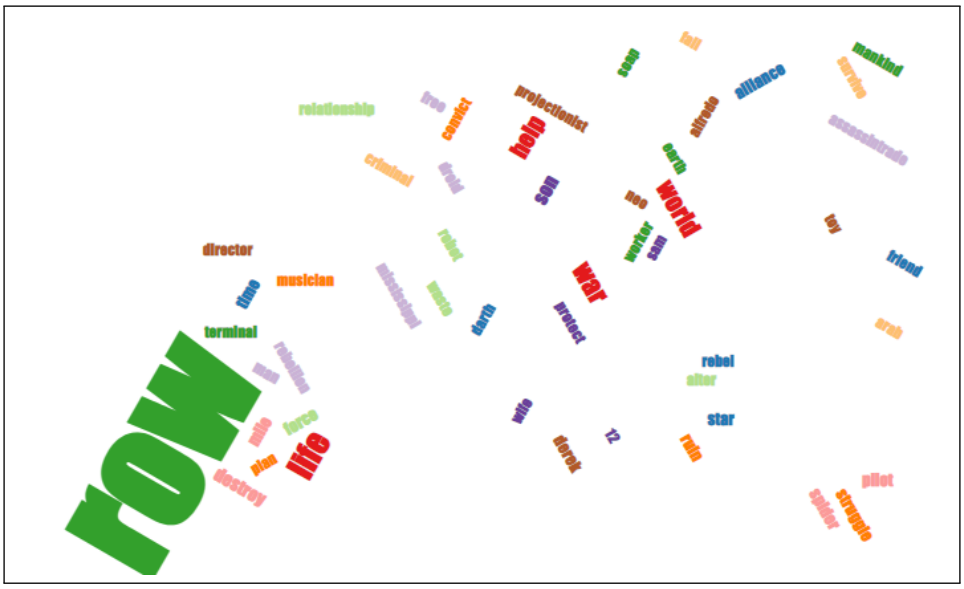


Figure 20: Word cloud created from the topic terms using the "Tag Cloud" node combined with the "Color Manager" node

them as a heatmap (Figure 18). This also allows us to see how topics overlap within individual documents.

We can also combine the top and middle output in order to create a data table containing the documents, their assigned topics, and the terms for the assigned topic. This is done by grouping both outputs according to a shared column, and then joining the two resulting tables together.

We can also make use of the middle output, containing the terms and their weights within the topic, to create a word cloud based on the different terms used. This can serve as an overview of the content of the topics extracted from the dataset. We start by using a "Color Manager" node to assign a consistent color for each of the 10 topics. While this could be done automatically by some of the visualization nodes, the inconsistent use of colors that often result can create ample confusion when several different visualizations are provided based of the same dataset.

After assigning the colors for the topics, we simply pass the output of the "Color Manager" to the "Tag Cloud" node (Figure 19), which creates a word cloud visualization (Figure 20). The configurations for the "Tag Cloud" node allow us to play around with the number of terms included, the sizing based on the weight for the topics, and some other settings that can help us make the visualization as efficient as possible. The word cloud

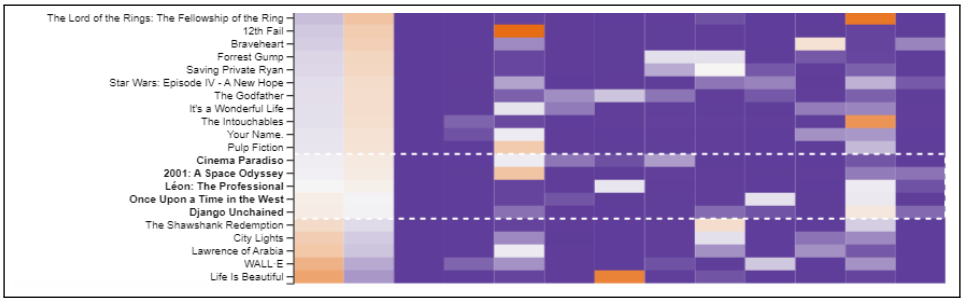


Figure 21: Showcasing a data selection made in the view of the heatmap visualization.

1: Image port ▶ 2: Output table Flow Variables

Rows: 100 | Columns: 9 Table Statistics

RowID	Title String	Title_URL String	User_rating String	Year String	Text String	Prediction String	P (Positive) Number (dou...	P (Negati... Number (dou...	Selected ... Boolean value
Row27	Spirited Away	https://www.i...	8.6	2001	During her fa...	Negative	0.02	0.98	true
Row32	Harakiri	https://www.i...	8.6	1962	When a ronin ...	Negative	0.022	0.978	false
Row84	Eternal Sunsh...	https://www.i...	8.3	2004	When their rel...	Negative	0.023	0.977	true
Row92	Toy Story	https://www.i...	8.3	1995	A cowboy doll...	Negative	0.023	0.977	false

Figure 22: Output data table of the heatmap visualization where current selection is indicated by Boolean value in an appended column.

is also a good way of inspecting the results to find terms that need to be appended to the stop word list. While the stop word list is initially used for removing function words, it can also be used to remove words that either bring nothing of value to the topics or are too frequent amongst the topics, resulting in noisy output.

6.3 *Using Visualizations to Filter*

While there are many different nodes available for different types of visualizations in KNIME, and even more through the use of extensions, the ones used for the example workflow were chosen as they allow for the filtering of data through interacting with the visualizations. In the case of the heatmaps, a selection can be made through clicking "Mouse Mode: Select" in the upper right corner of the view, and then selecting the items of interest. Multiple items can be selected by holding Ctrl (Windows) or Command (Mac) while clicking. Chosen items will be indicated by bolded labels and a dotted line (Figure 21).

Selected items are indicated as such in-place, meaning that the selection does not immediately filter the data. Instead, the use of the interactive visualizations add an additional column where a Boolean value (true or

false) indicates whether or not a specific row has been indicated as selected or not in a given visualization (Figure 22). While this might initially seem cumbersome from the perspective that an additional "Row Filter" node would have to be added to the workflow in order to actually filter the data set, there are benefits as well.

By indicating the selection in place, we have added a new meaningful component to our dataset, which we can make use of further down the line. For instance, it might be useful to know which items were selected based on either sentiment, topic, or both.

The tag cloud offers the same functionality as the heatmap, although selection there is made based on the inclusion of the term selected in the cloud. However, here it is important to keep the input in mind, as the selection column is appended to whatever data table is provided for the node. In the case of the tag cloud used in the example workflow, the input data table only contains the topics and terms. If we are intending to use it for item selection before moving on to further analysis, we might benefit more from using the table produced by the "Joiner" node, as that contains both the terms, topics, and the preprocessed documents.

7 *Concluding remarks*

This chapter has explored the application of low-code tools, specifically Octoparse and KNIME, for web scraping and text analysis within the context of the CICuW project. By demonstrating an accessible workflow that includes data collection, preprocessing, and computational text analysis, we have highlighted how these tools can lower the barrier to entry for researchers with limited programming experience while still enabling sophisticated analytical approaches. The use of zero-shot sentiment analysis and topic modeling with BERT and LDA, respectively, underscores the potential of such methods to generate meaningful insights into large-scale digital discourse.

Despite the advantages of low-code tools in democratizing computational methodologies, we have also emphasized the limitations and challenges they introduce. The "closed-box" nature of such tools can obscure methodological transparency, making it crucial for researchers to critically engage with the processes and assumptions underlying the analyses. Ethical considerations remain a significant concern, particularly regarding web scraping and online data collection, where informed consent is not always straightforward. The adoption of frameworks such as the Association of Internet Researchers' ethical guidelines serves as a necessary step in ensuring responsible research practices. Furthermore, while automated tools can efficiently generate initial

insights, the necessity of combining them with qualitative approaches for contextual interpretation remains paramount, as demonstrated in the CICuW project. Through combining quantitative approaches and qualitative content analysis, the pilot study shows that far-right digital information activities demonstrate a view of public librarians as leftist and as potential threats to the nation. The content analysis of comments from the most commented article, an article about a proposed law requiring public servants to report unauthorized immigrants, indicates a new form of political conflict between national policy and local politics (cf. Harding 2022) that can be connected to the growing national influence of the Sweden Democrats, a far-right party known for nationalism and anti-immigration politics (Dal Bo' et al. 2022).

Ultimately, the methodologies outlined in this chapter illustrate the potential of mixed-methods research in digital humanities, leveraging computational techniques to enhance qualitative inquiry. At the same time, it is important to note that while quantitative methods are necessary when analyzing large amounts of data, qualitative approaches are needed to provide nuanced and contextual understanding. This is particularly important when dealing with contested topics and material rich with irony and sarcasm. By integrating text analysis with exploratory visualization, researchers can identify key themes and focal points for deeper investigation, allowing for a more nuanced understanding of online discourse. As computational tools continue to evolve, future research should aim to refine the balance between accessibility, methodological rigor, and ethical responsibility, especially in cases where closed-box techniques are applied.

References

Berthold, Michael R., Nicolas Cebron, Fabian Dill, Thomas R. Gabriel, Tobias Kötter, Thorsten Meinl, Peter Ohl, Kilian Thiel & Bernd Wiswedel. 2009. KNIME - the Konstanz information miner: Version 2.0 and beyond. *ACM SIGKDD Explorations Newsletter* 11(1). 26–31. DOI: 10.1145/1656274.1656280.

Blei, David M., Andrew Y. Ng & Michael I. Jordan. 2003. Latent Dirichlet allocation. *The Journal of Machine Learning Research* 3. 993–1022.

Carlsson, Hanna, Fredrik Hanell & Joacim Hansson. 2022. "det känns som att jag bara sitter och väntar på att det ska explodera" - politisk påverkan på de kommunala folkbibliotekens verksamhet i sex sydsvenska regioner. *Nordic Journal of Library and Information Studies* 3(1). 26–43. DOI: 10.7146/njlis.v3i1.131852.

Dal Bo', Ernesto, Frederico Finan, Olle Folke, Torsten Persson & Johanna Rickne. 2022. Economic and social outsiders but political insiders: Sweden's populist radical right. *The Review of Economic Studies* 90(2). 675–706. DOI: 10.1093/restud/rdac037.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)*, 4171–4186. Minneapolis, MA. DOI: 10.18653/v1/N19-1423.

franzke, aline shakti, Anja Bechmann, Michael Zimmer, Charles Ess & The Association of Internet Researchers. 2020. *Internet research: ethical guidelines 3.0*. https://aoir.org/reports/ethics3.pdf.

Hanell, Fredrik & Pernilla Jonsson Severson. 2023. An open educational resource for doing netnography in the digital arts and humanities. *Education for Information* 39(2). 155–172. DOI: 10.3233/EFI-230024.

Harding, Tobias. 2022. Culture wars? The (re)politicization of Swedish cultural policy. *Cultural Trends* 31(2). 115–132. DOI: 10.1080/09548963.2021.1971932.

Honnibal, Matthew & Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Hsieh, Hsiu-Fang & Sarah Shannon. 2005. Three approaches to qualitative content analysis. *Qualitative Health Research* 15(9). 1277–1288. DOI: 10.1177/1049732305276687.

Hutto, C. & Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media* 8(1). 216–225. DOI: 10.1609/icwsm.v8i1.14550.

Jones, Steven E. 2016. *Roberto Busa, S. J., and the emergence of humanities computing: The priest and the punched cards*. 1st. Routledge. DOI: 10.4324/9781315643618.

Kozinets, Robert. 2020. *Netnography: The essential guide to qualitative social media research*. 3rd. Thousand Oaks, CA: Sage Publications.

Octoparse. 2024. *Octoparse 8*. https://octoparse.com/.

Scrivens, Ryan, Garth Davies & Richard Frank and. 2020. Measuring the evolution of radical right-wing posting behaviors online. *Deviant Behavior* 41(2). 216–232. DOI: 10.1080/01639625.2018.1556994.

Selenium. 2024. *Selenium Web Driver*. https://selenium.dev/.

Shneiderman, Ben. 1996. The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE symposium on visual languages*, 336–343. DOI: 10.1109/VL.1996.545307.

Tyrkkö, Jukka & Daniel Ihrmark. 2024. Low-code data science tools for linguistics. In Steven Coats & Veronika Laippala (eds.), *Linguistics across disciplinary borders: The march of data*, 40–66. Bloomsbury. DOI: 10.5040/9781350362291.0008.

Wing, Jeannette. 2006. Computational thinking. *Communications of the ACM* 49(3). 33–35. DOI: 10.1145/1118178.1118215.

List of abbreviations

AoIR	Association of Internet Researchers
BERT	Bidirectional Encoder Representations from Transformers
CICuW	Cultural Institutions and the Culture War
DOM	Document Object Model
HTML	HyperText Markup Language
IMDb	Internet Movie Database
IRE	Internet Research Ethics
JSON	JavaScript Object Notation
KNIME	Konstanz Information Miner
LDA	Latent Dirichlet Allocation
RoBERTa	Robustly optimized BERT approach
VADER	Valence Aware Dictionary and sEntiment Reasoner

Corresponding author

Daniel Ihrmark
Department of Cultural Sciences
Linnæus University
daniel.o.sundberg@lnu.se

16 Exploratory Swedish text analysis using notebooks – a *smörgåsbord* of basic corpus linguistic insights

Dimitrios Kokkinakis
University of Gothenburg

Gerlof Bouma
University of Gothenburg

The computational notebook has established itself as a significant tool for conducting exploratory data analysis, which aims at investigating characteristics of a dataset without preformulated expectations. Computational notebooks are a type of interactive document, that supports mixing prose, executable code and its output, such as a calculated result, a table, or a graphic. Data, process, and narrative are effectively integrated into one environment, which makes notebooks ideal for documenting exploratory research. Notebooks also facilitate sharing research in a reproducible way for teaching, collaboration or dissemination.

This chapter demonstrates basic exploratory techniques for Swedish text analysis implemented as *Jupyter notebooks*, a popular computational notebook implementation. Using a selection of documents from a Swedish corpus of COVID-19-related materials, we show some of the kinds of text analysis that can easily be performed using readily available software libraries. The examples in this chapter rely only on automatic annotation, requiring minimal manual processing.

1 *Notebooks for exploratory data analysis in the humanities*

As the field of Digital Humanities continues to evolve, researchers are increasingly exploring new ways to integrate computational tools into traditional humanistic study. The significant expansion in the volume of accessible materials has fuelled the development of methodologies that leverage data-driven analysis, visualization techniques, and machine learning to uncover new insights into historical, literary, and cultural phenomena. A critical

HUM INFRA Dimitrios Kokkinakis & Gerlof Bouma. 2025. Exploratory Swedish text analysis using notebooks – a *smörgåsbord* of basic corpus linguistic insights. In Gerlof Bouma, Dana Dannélls, Dimitrios Kokkinakis & Elena Volodina (eds.), *Huminfra handbook: Empowering digital and experimental humanities* (NEALT Proceedings Series 59), 541–563. University of Tartu Library. DOI: 10.58009/aere-perennius0185 ©The authors, CC BY 4.0

area of growth is also the use of open-source platforms and collaborative tools, which allow scholars to perform complex data analysis while sharing methodologies transparently.

Large-scale analysis of textual sources is one of the core methodologies of Digital Humanities. Computational linguistics and corpus linguistics have contributed greatly to this methodology and will continue to do so. This chapter showcases some basic techniques from these linguistic fields and their use in exploratory data analysis in the context of text-based Digital Humanities (York 2017). Exploratory data analysis is "a collection of techniques for identifying the main characteristics of a novel dataset, about which one may initially know nothing" (York 2017: 462) and involves looking at the materials that form a study's empirical basis from many different angles and at different levels of abstraction, qualitatively and especially, quantitatively. Documenting this iterative process is very important. It supports the researchers themselves, but it also provides a means to communicate to others how the researchers came to draw conclusions from the dataset or formulate hypotheses based on it. *Computational notebooks* have proven to be an effective way of creating this documentation as part the analysis process.

1.1 *Introduction to notebooks*

A computational notebook is an interactive document that lets the notebook author combine formatted text, source code and program output into one flow. The researcher can thus perform exploratory data analysis inside the notebook environment, and add explanatory narrative, be it private or public, along the way.

Because it places code and explanatory narrative on equal footing, the computational notebook concept aligns within the paradigm of *literate programming* (Knuth 1984). In this paradigm documented computer programs are written in the form of coherent texts focussed on the human reader, which include the executable source code of the program that is being documented. Notebooks add to this paradigm the ability to directly execute code and display its results as part of such documents. A computational notebook thus consists of three types of content, or "cells" in notebook terminology: formatted text, program code and program output. The latter may be in the form of text, tables, visualizations, etc. During creation of a notebook, the text and code cells are directly editable. Notebooks can be converted to formats suitable for publishing, for instance PDFs or webpages. These are typically not dynamic in the same way, although some notebook implementations may allow for interactive content in published notebooks. Also, when the underlying notebook files themselves are made accessible

– an increasingly common dissemination strategy – other researchers may directly use them as dynamic documents to edit and build upon.[1]

Notebooks are actively deployed in a wide range of scientific contexts, including education, economics, engineering, data science, (corpus) linguistics and digital humanities (Alderson 2021, Barba et al. 2019, Granger & Pérez 2021, Hardebolle 2023, Dombrowski et al. 2019, Bednarek et al. 2024). Interestingly, although notebooks can be and are used for dissemination, their primary use is, as their name suggests, as personal documentation and support in data exploration (Rule et al. 2018). Notebooks are also valued as pedagogical tools (Blanke et al. 2023), amongst other things because they allow instructive text to be interspersed with editable and executable code. The adoption of notebooks in Digital Humanities research represents a clear method advancement in the field. By combining the power of computer-supported data analysis with the flexibility of narrative and collaboration, these tools reshape how scholars may engage with large and complex datasets. Computational notebooks provide a versatile platform that enhances both the analytical process and the communication of research findings. Through their support for iterative analysis and dynamic visualization, these tools lower the threshold for addressing complex research questions in the humanities using data-intensive methods. They are increasingly central to interdisciplinary research, bridging computational and humanistic approaches (Ahnert et al. 2023). In the quickly evolving practices of digital humanists, notebooks can supplement "fixed" graphical user interfaces to give more technical users flexible access to archives and datasets (Melgar-Estrada et al. 2019).

1.2 *Implementations and deployment of notebooks*

There are many modern implementations of the notebook idea; open-source ones such as Jupyter Notebook[2] (Granger & Pérez 2021), currently the de-facto standard, and Spyral (Land et al. 2021) as well as proprietary ones such as MATLAB Live Editor[3] and the seminal notebooks of Wolfram Mathematica.[4] Depending on the implementation, notebooks can be created and viewed locally or online. For the widely used Jupyter Notebook, commercial online services include Google Colaboratory[5] (Colab), Kaggle[6], and Mi-

1 As just a tiny example of this, https://github.com/quinnanya/dh-jupyter list 100 notebooks / notebook-based resources with Digital Humanities relevance.
2 https://jupyter.org
3 https://www.mathworks.com/products/matlab/live-editor.html
4 https://www.wolfram.com/notebooks/
5 https://colab.research.google.com/
6 https://www.kaggle.com

crosoft's GitHub Codespaces and Azure Machine Learning.[7] One advantage of using an online service is that there is no need to install any software: the functionality to run the notebooks is supplied by the service and typically there is a rich, pre-installed ecosystem of auxiliary packages that add ready-made capabilities to the notebook. Moreover, online services may provide access to powerful computing servers that allow the researcher to work on a scale that is not easily achievable on their local hardware. Some services also offer tools to facilitate collaboration, for instance through joint notebook editing and publication. On the other hand, the advantage of running notebooks locally is that the researcher keeps full control over code and data, imperative when working with sensitive or otherwise restricted materials, and does not need a user account, crucial when researcher anonymity is important. Also, no internet connectivity is needed for local execution, even when Jupyter Notebook uses the browser as their user interface.

In this chapter, we hope to demonstrate notebooks in exploratory analysis of (large-scale) text corpora in a way that is both efficient for the researcher and capable of supporting insightful interpretations. We will use Jupyter Notebook as our notebook platform and Python as our programming language. Figure 1 contains an overview of the Jupyter Notebook interface. This chapter will not include detailed instructions on how to install or use Jupyter Notebook, neither will it introduce the Python language. For this we refer the reader to the many online tutorials on these topics, to Dombrowski et al. (2019) for an article length instructive introduction to Jupyter Notebook for researchers in the humanities, and to Karsdorp et al. (2021) for a book length introduction to humanities data analysis using Python. The notebooks will be available from spraakbanken.gu.se/en/projects/huminfra/exploratory-text-analysis.

2 *Research scenario*

The following sections present an outline of the various text analysis components that are implemented in our notebooks. These components not only serve as a guide to the implemented methodologies but also offer a dynamic showcase of the diverse analyses conducted using the platform. Furthermore, they stimulate hands-on exploration and experimentation by individuals with minimal programming skills, thereby democratizing access to advanced data analysis tools and enabling broader participation in computational research. This transparency ensures the reproducibility of the research, allowing others to access and verify the data sources used. All

7 https://visualstudio.microsoft.com/vs/features/notebooks-at-microsoft/

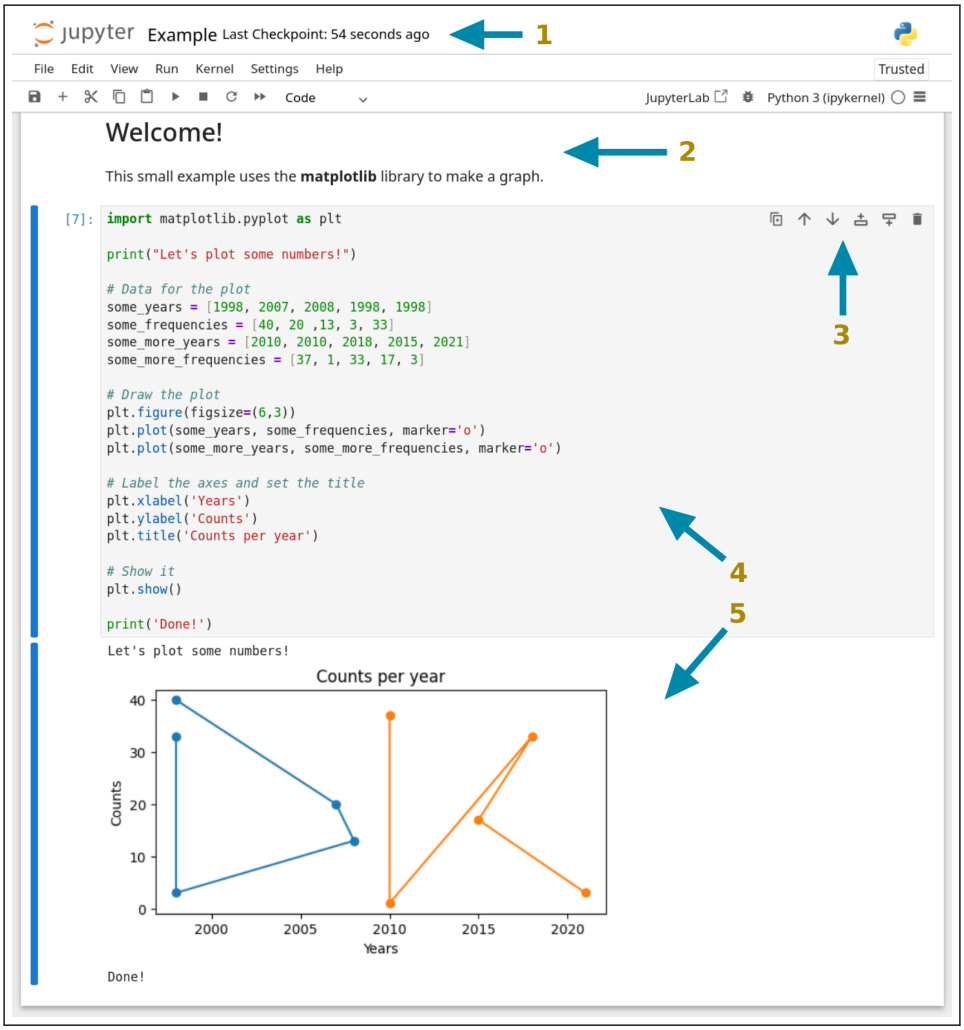


Figure 1: The user interface in Jupyter Notebook: (1) The name of the notebook and information about the last manual save. (2) A *markdown cell* containing formatted prose text. (3) Controls to move the active cell or add new cells around it. (4) A *code cell* containing Python code. The coloured code formatting is provided by the notebook software. (5) The output of the code cell above it, containing both textual and graphical elements.

resources used in the demonstrations, the notebooks as well as the used text corpus and lexical resources, are available online and are linked from the notebooks.

Although interactive notebooks are available for many different programming languages, we use Python as our language becuase of its accessability

and the rich ecosystem of libraries for data analysis, NLP and graphing. In our notebooks, we use for instance NLTK[8], spaCy[9] and BERTopic[10] for NLP, Pandas[11] or Polars[12] for data handling, NumPy[13] and SciPy[14] for numerical computing, and Matplotlib[15] for graphing.

After a brief introduction of the dataset used in all of our notebooks in Section 2.1, Sections 2.2–2.6 discuss the different case studies on this data.

2.1 *Dataset overview*

The text corpus used in the experiments below consist of approximately 2 700 documents published online in 2021, during the COVID-19 pandemic. This dataset is part of the sv-COVID-19 corpus (Kokkinakis 2021), which contains articles published in Swedish on the internet and covers a range of topics related to the COVID-19 pandemic from different sources. The corpus has been systematically classified by the first author into eight distinct stylistic genres, determined by the original publication medium and contextual factors surrounding each text.[16] The figure between parenthesis is the number of documents for each genre:

- Authorities (64): e.g., https://skr.se/
- Blog (189): http://www.islandsbloggen.com/
- Medical (460): e.g., https://www.lif.se/
- News (850): e.g., https://www.gp.se/
- Public media (759): e.g., https://sverigesradio.se/
- Periodicals (338): e.g., https://borsvarlden.com/
- Research (44): e.g., https://www.forskning.se/
- Social media (8): e.g., https://www.familjeliv.se/

These genres provide different perspectives on the pandemic, highlighting the varied approaches to the topic across different platforms. The ability to perform genre-specific searches facilitates the exploration of how different

8 https://www.nltk.org/
9 https://spacy.io/
10 https://maartengr.github.io/BERTopic/
11 https://pandas.pydata.org/
12 https://pola.rs/
13 https://numpy.org/
14 https://scipy.org/
15 https://matplotlib.org/
16 This categorization enables targeted analysis and querying within SpråkBanken Text's word research platform *Korp*: https://spraakbanken.gu.se/korp/#?corpus=sv-covid-19.

types of media and publication contexts influence the portrayal and discussion of COVID-19. Researchers can also use the data to examine shifts in language use, thematic focus, and, to some extent, public sentiment over time. Analysis of data like these may offer insights into pandemic-related communication and its broader impact, and hopefully deepen our understanding of how the crisis was discussed and perceived.

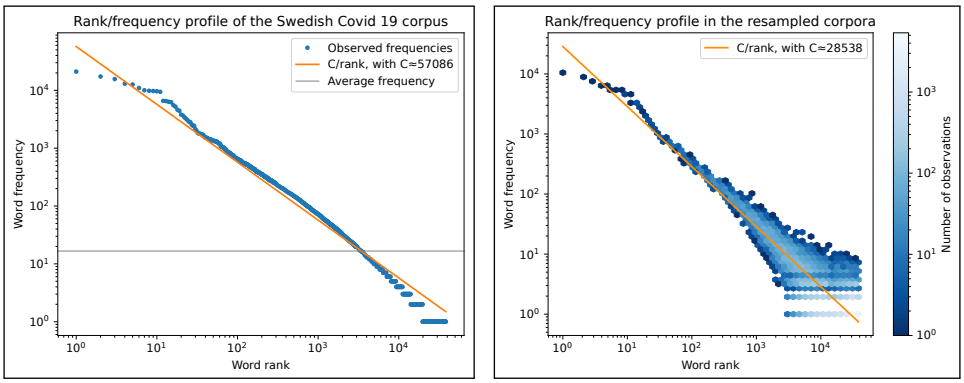
2.2 *Word distributions and word frequencies: Zipf and beyond*

In linguistics, the study of word distributions and word frequencies, particularly in large corpora, has been significantly shaped by what has become to be known as Zipf's Law (Zipf 1949). This model posits that the frequency of a word is inversely proportional to its rank: A small number of high-frequency words dominate, while the vast majority of words occur only very infrequently. Easy access to text materials at scale for many languages has allowed for precise word frequency analysis, revealing both adherence to and deviations from Zipf's model (Piantadosi 2014).

Figure 2a shows a classical plot of the *Zipfian distribution* of word frequencies (rank–frequency plot) in the comparatively small, examined dataset. The distribution is plotted on logarithmic scales for both axes; the Zipfian model says that the graph should follow a straight line in this case. The model's predicted relation between rank and frequency of is included in the graph in orange. A few function words and closed class words, such as *att* 'to/that' (infinitive marker/subordinator), *i* 'in', *och* 'and', *det* 'it', and *som* 'that/as' (subordinator), are used extremely often, while an abundance of words are rarely used and thus appear in the right tail of the diagram.

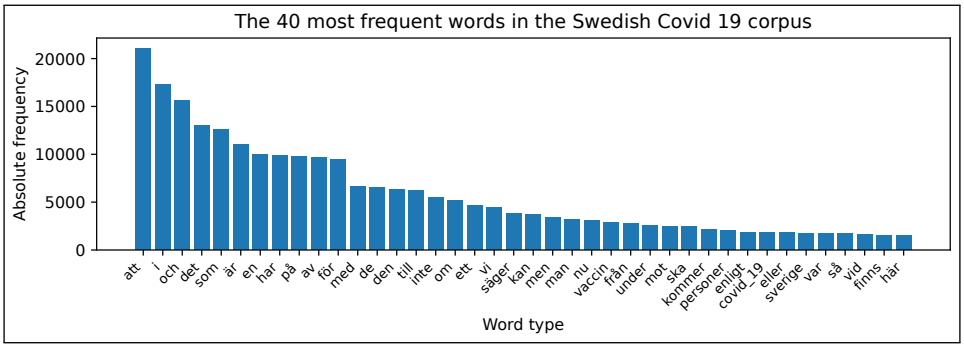
Piantadosi (2014) argues that estimating both rank and frequency from exactly the same data is a mistake, and proposes to create two random splits of a dataset: one for the estimation of a word's frequency and one for the estimation of its rank. The resulting graph is in Figure 2b. The overall shape is very similar, with a strong linear trend on the log–log scale. But the right tail now fans out, because at the low frequency end, small differences in frequency may lead to relatively large differences in rank. This uncertainty in the relation between rank and frequency is not shown at all in the classic way of constructing the Zipf plot, but it shows (correctly) in this alternative method.

The next two graphs connect the previous Zipfian discussion with an alternative visualisation of word frequencies. Figure 3a shows the frequencies of the top 40 most frequent words as a bar chart. What is a linear trend in the Zipf plots, can be seen here as exponential decay. The sharp drop after *för* 'for' (at 11th place) in the bar chart can be recognized as a bend in the

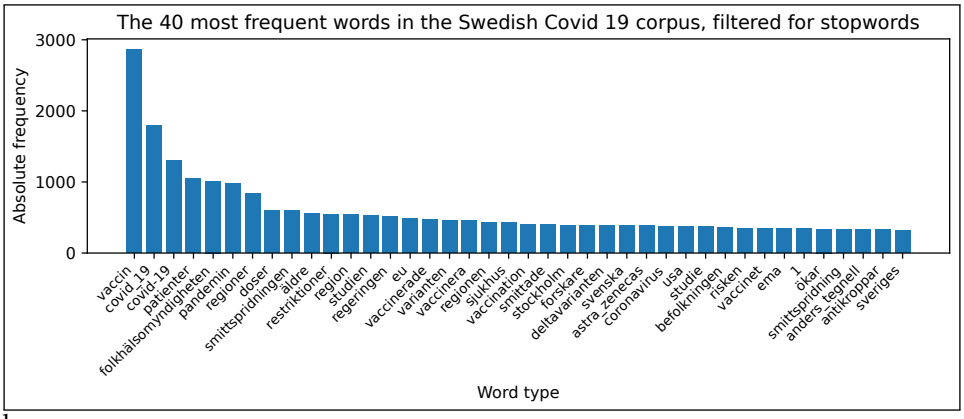


a. b.

Figure 2: The Zipfian distribution of the words in the dataset, in (a) the classic style and (b) the style of Piantadosi (2014) with rank and frequency estimated from different random subsets of the corpus.



a.



b.

Figure 3: The 40 most frequent unigrams in the dataset, (a) unfiltered, (b) without stopwords.

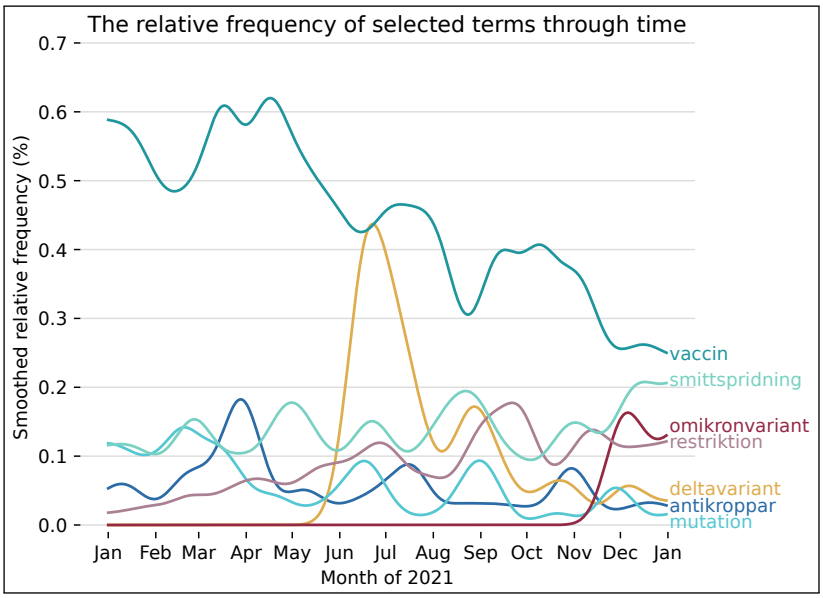


Figure 4: Smoothed word frequency trends for a small number of selected words.

Zipf plot. Figure 3b shows the same, but after so-called "stop words" have been filtered out. Stop words typically provide minimal semantic value in isolation, making them less relevant for tasks like text classification, sentiment analysis, and information retrieval, and are often filtered out during preprocessing stages, as their high frequency across texts can overshadow more meaningful content. We see that the overall shape of the distribution is similar (quickly dropping in frequency initially, more slowly towards the right-hand side), but from a completely different starting point in terms of the frequency of the most common words. Looking at the words themselves, they are not just general domain content words, but many of them are highly specific to the corpus' common denominator of the COVID-19 pandemic, such as *vaccin* 'vaccine', *covid-19*, and *folkhälsomyndigheten* 'Public health agency of Sweden'.

The corpus contains materials collected over a whole year, and it is therefore possible to study the changes in word frequencies over this period. Figure 4 shows the word frequency distributions over the examined year for a small number of selected words, namely *omikronvariant* 'omicron variant', *deltavariant* 'delta variant', *vaccin* 'vaccine', *restriktion* 'restriction', *smittspridning* 'spread of infection', *antikroppar* 'antibodies', and *mutation*, including inflected forms of these words. In this diagram one can clearly see fluctu-

ations and areas where word frequencies spike, like the brief heightened frequency of mention of the *delta variant* in the middle of the year and the rise of the *omicron variant* at the end of 2021. These observations can be linked to real-world events, such as the dominance of these variants in Sweden during this year.

2.3 *Significant word associations and collocations*

Collocations are words that tend to occur in each other's company. At a macro-level, the study of collocations can tell us things about the nature of the text material we are studying and at a micro-level, we can learn about the meaning of a word by studying its collocates.[17] Looking at collocations may also provide the empirical foundation for lexical studies of multiword expressions such as idioms, phrasal verbs, proper names, etc.

The notion of collocation can be operationalized in many different ways, for instance by the choice of *association measure*, that is, a score to summarize how often words co-occur and how surprising this level of co-occurrence is. These measures typically rely not just on the observed frequency of co-occurence, but also on an expected frequency, based for instance on what we would expect to see if the distribution of two words were not related to eachother.

Table 1 shows the 20 bigrams, directly adjacent words, from the COVID-19 corpus that score highest on the association measure of *log-likelihood ratio* (LLR),[18] after some preprocessing like removing case distinctions and filtering out combinations containing stop words.[19] This LLR top 20 contains personal (*Sara Byfors*) or geographic (*Västra Götaland*) proper names, but we can furthermore see a recurrent pattern descriptive noun–proper name (*statsepidemiolog Anders Tegnell* 'chief epidemiologist Ander Tegnell', *socialminister Lena Hallengren* 'minister of social affairs Lena Hallengren', *läkemedelsmyndigheten EMA* 'medicinal product authority EMA', amongst others).

The table also gives scores for four other association measures: frequency, point-wise mutual information (PMI), t-score, and chi-squared (χ^2). To show that these measures may disagree wildly about which bigrams form strong collocations, we have included the rank of each bigram for the re-

17 Or, as the famous and ubiquitously cited adage from Firth (1957: 11) has it: "You shall know a word by the company it keeps!"

18 We refer the reader to Evert (2009) for definitions of LLR and the other association measures mentioned here, and for a good general discussion of the collocation concept.

19 The used corpus contains some preprocessing that combines first name–last name combinations. For simplicity, we have treated these underscore-separated combinations as one "word". This means that the list of bigrams may contain combinations of more than two words, such as in *astra_zenecas vaccin* 'Astra Zeneca's vaccine'.

Table 1: The 20 strongest bigram collocations in the dataset according to the log-likelihood ratio (LLR), for pairs that do not contain stopwords and occur at least 5 times. Scores (and ranks) for four other association measures are also given.

Bigram	LLR	Freq	PMI	t-score	χ^2
astra_zenecas vaccin	3043.7	315 (1)	7.4 (965)	17.6 (1)	52513.6 (366)
statsepidemiolog anders_tegnell	2068.1	140 (2)	10.7 (558)	11.8 (2)	230143.3 (166)
tegmark wisell	1829.0	95 (6)	12.5 (355)	9.8 (6)	557066.6 (1)
socialminister lena_hallengren	1698.6	99 (5)	11.9 (413)	10.0 (5)	374281.3 (108)
europeiska läkemedelsmyndigheten	1544.5	102 (4)	11.1 (509)	10.1 (4)	217260.4 (176)
sara byfors	1444.5	78 (10)	12.5 (352)	8.8 (10)	468851.2 (49)
karin tegmark	1386.6	83 (8)	11.8 (418)	9.1 (8)	305002.5 (137)
västra götalandsregionen	1203.8	73 (12)	11.6 (452)	8.6 (12)	220265.8 (172)
västra götaland	1203.8	73 (12)	11.6 (452)	8.6 (12)	220265.8 (172)
johan carlson	1097.9	74 (11)	10.8 (539)	8.6 (11)	136540.0 (223)
region stockholm	1092.7	108 (3)	8.3 (888)	10.4 (3)	34492.6 (439)
emma spak	1027.0	56 (20)	12.7 (338)	7.5 (20)	370954.6 (109)
sveriges kommuner	1018.7	83 (8)	9.7 (687)	9.1 (9)	67857.1 (333)
matti sällberg	991.3	50 (22)	13.3 (279)	7.1 (22)	492057.8 (44)
kristine rygge	954.7	47 (24)	13.4 (264)	6.9 (24)	526653.4 (20)
läkemedelsmyndigheten ema	947.6	68 (14)	10.6 (572)	8.3 (14)	102524.6 (273)
låga nivåer	936.8	65 (16)	10.9 (526)	8.1 (16)	127783.3 (235)
karolinska universitetssjukhuset	914.4	57 (18)	11.8 (422)	7.6 (18)	207728.8 (179)
brittiska varianten	902.2	84 (7)	8.8 (836)	9.2 (7)	36223.9 (424)
johan bratt	875.0	57 (18)	11.0 (519)	7.6 (19)	115870.4 (252)

maining association measures. We see for instance that LLR, frequency and t-score agree to a high extent, since the latter two also give prominent ranks to the cases that LLR brings up. PMI and χ^2 on the other hand would place the shown combinations much further down on the list. An important part of studying collocations is choosing the association measure that gives the relevant insights into the data.

Collocations need not just concern directly adjacent words. Figure 5 is a network visualisation of the most common words to occur within a window of four words before or after the focus terms of *positiv*, *positivt*, *positiva* 'positive(ly)' and *negativ*, *negativt*, *negativa* 'negative(ly)'. The placement of nodes and the thickness of connecting edges reflect how common a collocation is. The lack of a connection between nodes indicates that the frequency of co-occurrence falls below a pre-set threshold (and may be zero). Only

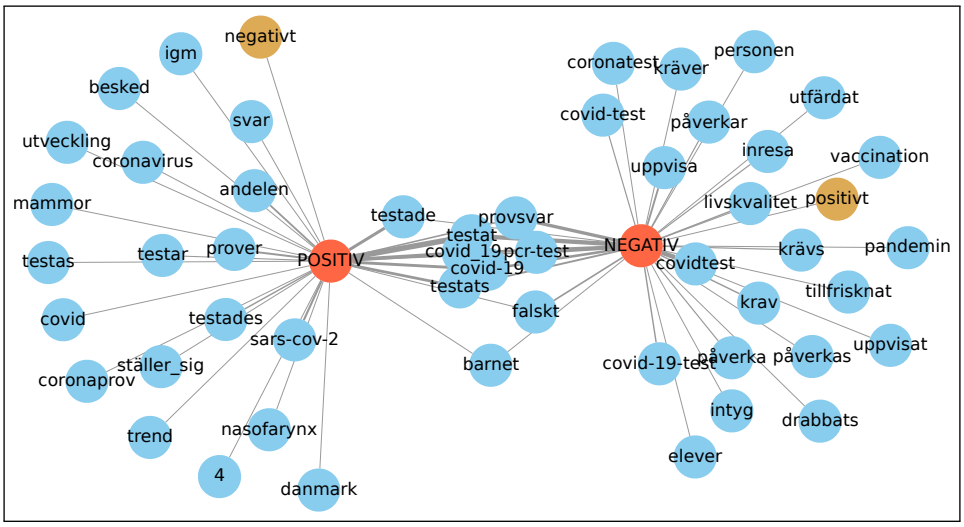


Figure 5: Terms strongly collocating with "positive" and "negative" as a network graph.

a small set of words are collocations of both *positive* and *negative*. A clear real-world effect on the collocation pattern can be seen in the collocates of *negative* that have to do with restrictions on free movement and the demands on having proof of recent negative test results, such as: *intyg* 'certificate', *krav* 'demand', *krävs* 'is demanded', *inresa* 'entry', *uppvisa* 'show'.

Association measures can also be used to study the association between words and a category. The heatmap in Figure 6 visualises the association between singular personal pronouns and genre. The map uses a normalized variant of pointwise mutual information, NPMI, which falls between -1 (never cooccur) and 1 (always cooccur). The map shows for instance that the first and second person pronouns (*jag* 'I', *mig* 'me', *du* 'you.sub.sg', *dig* 'you.obj.sg') are associated with the Social media and Blog genres, and that the gender neutral *hen* '(singular) they' is associated with the Medical genre. The second person pronouns are also associated with the Research and Authorities data. This high-level overview can be the starting point for a fine-grained, qualitative investigation for the mechanisms behind these associations.

2.4 *Grammatical analysis and Universal Dependencies*

SpaCy is an open-source, industrial-strength NLP pipeline with models for many languages. It offers for instance facilities for part-of-speech tagging,

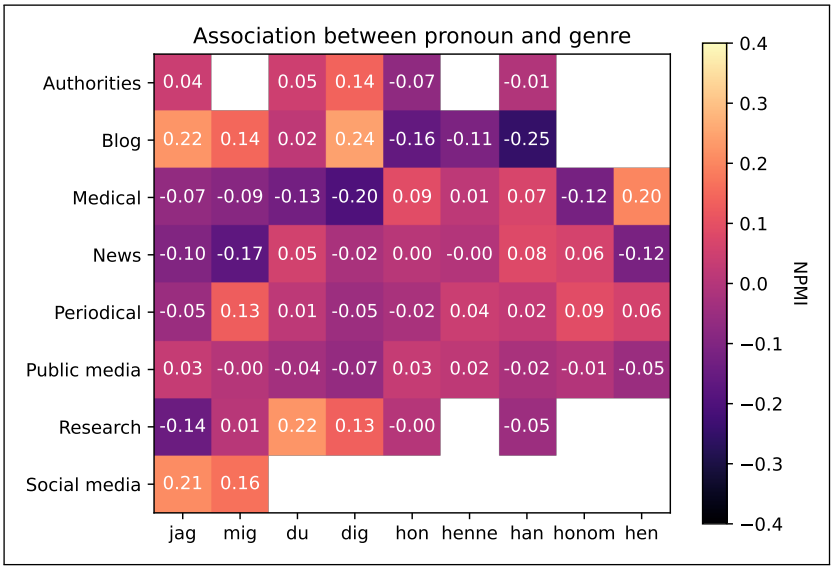


Figure 6: Association between singular person pronouns and genre in the dataset. Cells without observations (by definition NPMI $= -1$) are left uncoloured to avoid clutter.

dependency parsing, named-entity recognition and text classification. Here we have used SpaCy's Swedish model to part-of-speech tag and parse the COVID-19 data.[20] Table 2 shows the distribution of part-of-speech tags in the data. Nouns are most common part of speech with just over 22%, whereas verbs make up about 16% (main and auxiliary verbs) of the data.

The syntactic structures assigned by SpaCy are dependency structures following the Universal Dependencies framework (de Marneffe et al. 2021). An example parse can be seen in Figure 7. The automatic parser manages to correctly recognize many grammatical relations in this sentence, such as the PP postmodifying structures in *exemplaren av nytrycket av boken* 'copies of the reprint of the book', but it also makes mistakes (in red in the figure), for instance when it treats *Corona-PLANdemin* 'Corona PLANdemic' as an adverbial, rather than part of the NP *boken Corona-PLANdemin* 'the Corona PLANdemic book', or when it treats *har* 'has' as the root instead of as a temporal auxiliary to the main verb *kommit* 'come'.

An example use of parses is to construct a measure of syntactic complexity of the material. One such measure is the average number of finite subordinate structures per sentence: having more such subordinate clauses is associated

20 The Swedish pipeline in SpaCy can be found here: https://spacy.io/models/sv (version: 3.7.0). The Swedish model applied is called `sv_core_news_sm`.

Table 2: Distribution of part-of-speech labels in the corpus according to the spaCy analysis.

Part of speech		Count	%
NOUN	Noun	157512	22.3
ADP	Preposition	83873	11.8
VERB	Main verb	81221	11.5
PUNCT	Punctuation	71037	10.0
ADJ	Adjective	57588	8.1
PRON	Pronoun	48894	6.9
ADV	Adverb	46357	6.5
AUX	Auxiliary verb	34772	4.9
DET	Determiner	30295	4.2
PROPN	Proper noun	27915	3.9
CCONJ	Coordinator	22755	3.2
SCONJ	Subordinator	19233	2.7
PART	Particle	15646	2.2
NUM	Numeral	7269	1.0
SYM	Other symbol	521	0.0
INTJ	Interjection	211	0.0

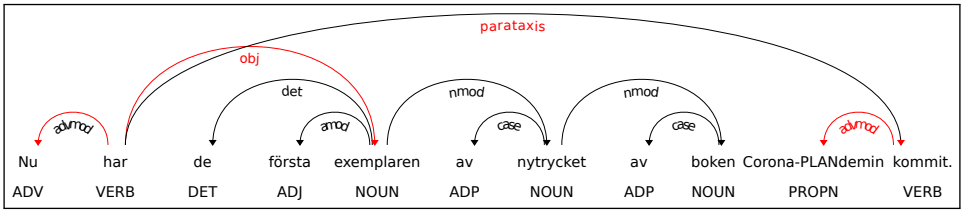


Figure 7: An example of a dependency tree from the dataset using the Swedish SpaCy model.

with increased complexity of the text. Subordinate clauses can be recognized by looking at the grammatical functions assigned by the parser. Figure 8 plots the relation between the number of subordinate clauses and sentence length for the Blog data and the News data. Perhaps somewhat surprisingly, the Blog data is more complex on several accounts: the average number of subordinate clauses is higher and, additionally, the subordinate clauses themselves are on average longer.

2.5 *Sentiment analysis*

Sentiment analysis is a computational technique for evaluating the polarity of opinions expressed in textual data. It enables scholars to track shifts in

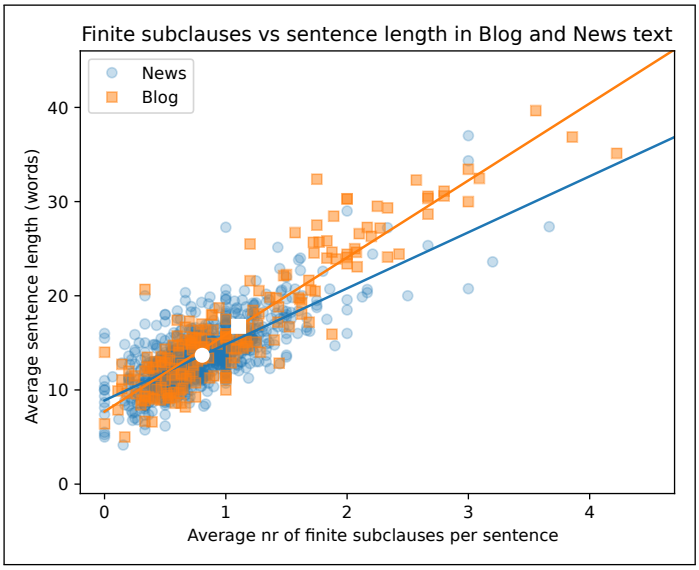
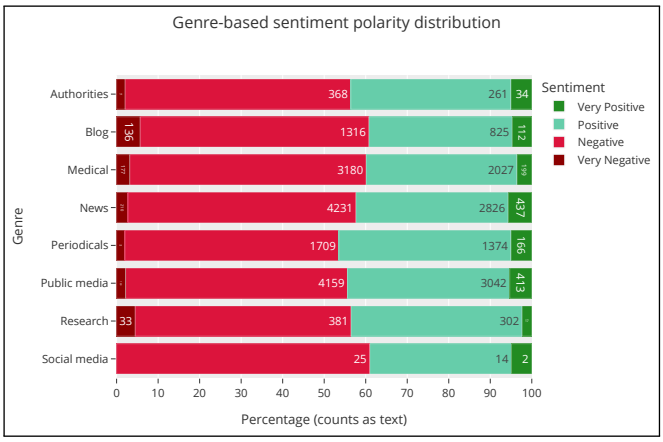


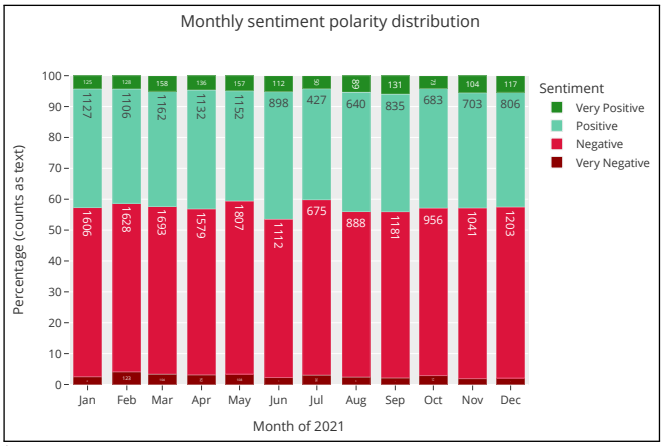
Figure 8: Texts from the Blogs genre are more complex than those from the News genre, both in terms of average sentence length and average number of finite subordinate clauses (the white points). Moreover, each additional finite subordinate clause is longer in Blog text than in News texts by on average 2.3 words (difference in regression line slopes, 95% confidence interval: 1.6–2.9 words/clause).

public opinion over time, assess attitudes toward significant historical events, and analyze the portrayal of individual characters in literature and other narratives, on a large scale. This quantitative approach offers a complementary lens through which to view qualitative sources, enabling researchers to uncover patterns and trends that may not be immediately apparent through traditional analysis alone.

A common method of sentiment analysis looks for the use of emotionally charged words from a so-called *sentiment lexicon*. However, scholars must carefully apply these computational tools to avoid overlooking the inherent subtleties and complexities in the texts under analysis. Literary works may contain complex meanings that require a deep understanding of their contexts. Texts from various historical periods or genres may use language and rhetorical strategies that complicate such direct sentiment analysis. Nevertheless, the "distribution of emotion-laden words" (Öhman 2021) can still provide valuable insights into the emotional tone of a text, serving as a foundation for more nuanced analysis. More advanced approaches to sentiment analysis involve machine learning, NLP and hybrid lexical-based



a.



b.

Figure 9: The sentiment distribution per genre (a) and per month (b) for 2021 in the dataset.

techniques, which enhance the ability to interpret sentiment by considering context, tone, and even the cultural background of the text. These advanced methods do come at the price of added complexity and development cost.

We apply a lexicon-driven approach to our material, using a list of about 2000 words marked up with polarity and strength (Nusko et al. 2016, Språkbanken Text 2017). We distinguish four categories of emotion-laden words, ranging from very negative to very positives. Results are visualised in Figure 9. The most striking thing about both pictures is how even the distribution is, throughout genres as well as over time. The (very) negative words have a fairly constant slight overhand (between 50-60%). The Blog

and Social media genres have the highest share of (very) negative words amongst emotion-laden words, followed by the Medical genre. If we were to pursue the topic further, we would probably look into what makes the Medical genre so negative. In addition, the Blog data contains the highest proportion of very negative terms.

2.6 *Topic modeling*

Topic modeling aims to uncover themes or topics in a collection of texts and is a hugely popular method in Digital Humanities.[21] Texts in a collection are stochastically assigned to a number of undefined topics on the basis of their linguistic content. These topics can then be interpreted by the researcher, by looking at the words associated with them, and their distribution throughout the material can be studied. Topic modelling can be contrasted with the way we have used genres in our dataset: whereas our "genres" are a classification imposed by us (by design) on the data, the categories from topic modelling come from the distributional properties of the data itself.

There are many implementations of the topic modeling idea, the most established technique being *Latent Dirichlet Allocation* (LDA, see Jelodar et al. 2019 for an overview). We, however, use the topic modelling library BERTopic, Grootendorst (2022), which presuposes a BERT-style language model that can create numerical meaning representations of the documents in the collection. For our Swedish data, we have access to the BERT models developed at KBLab at the National Library of Sweden (Rekathati 2021).[22] BERTopic requires minimal data preprocessing and its availability as a Python library means it is easy to integrate in our setup.[23]

The topics found using BERTopic in our data are summarized in Figure 10. We can interpret these topics by looking at the words associated with them. For instance, Topic 1, the most commonly present topic, appears to deal with pressure on the hospital care system (*patienter* 'patients', *krislägesavtalet* 'emergency labour agreement [in hospitals]', *IVA* 'ICU'), Topic 4 with speculations about the origin of the virus (*Kina* 'China', *Wuhan*, *ursprung* 'origin'), and Topic 7 with tracking the spread of the virus in the population (*avlopsvattnet* 'sewage', *PCR*, *smittspårning* 'contact tracing')

Looking at the distribution of these topics in the dataset now gives us new perspectives on the data. For instance, Figure 11 shows the frequency of

21 Already more then 10 years ago, Schmidt (2012: 49) wrote: "So many scholars in humanities departments are turning to the tool [of topic modeling] in their research that it is sometimes described as part of the digital humanities in itself."

22 Available from https://huggingface.co/KBLab/sentence-bert-swedish-cased.

23 See Haffenden & Sikora (2025; Chapter 2 in this handbook) for more information about the use of KBLab's BERT models.

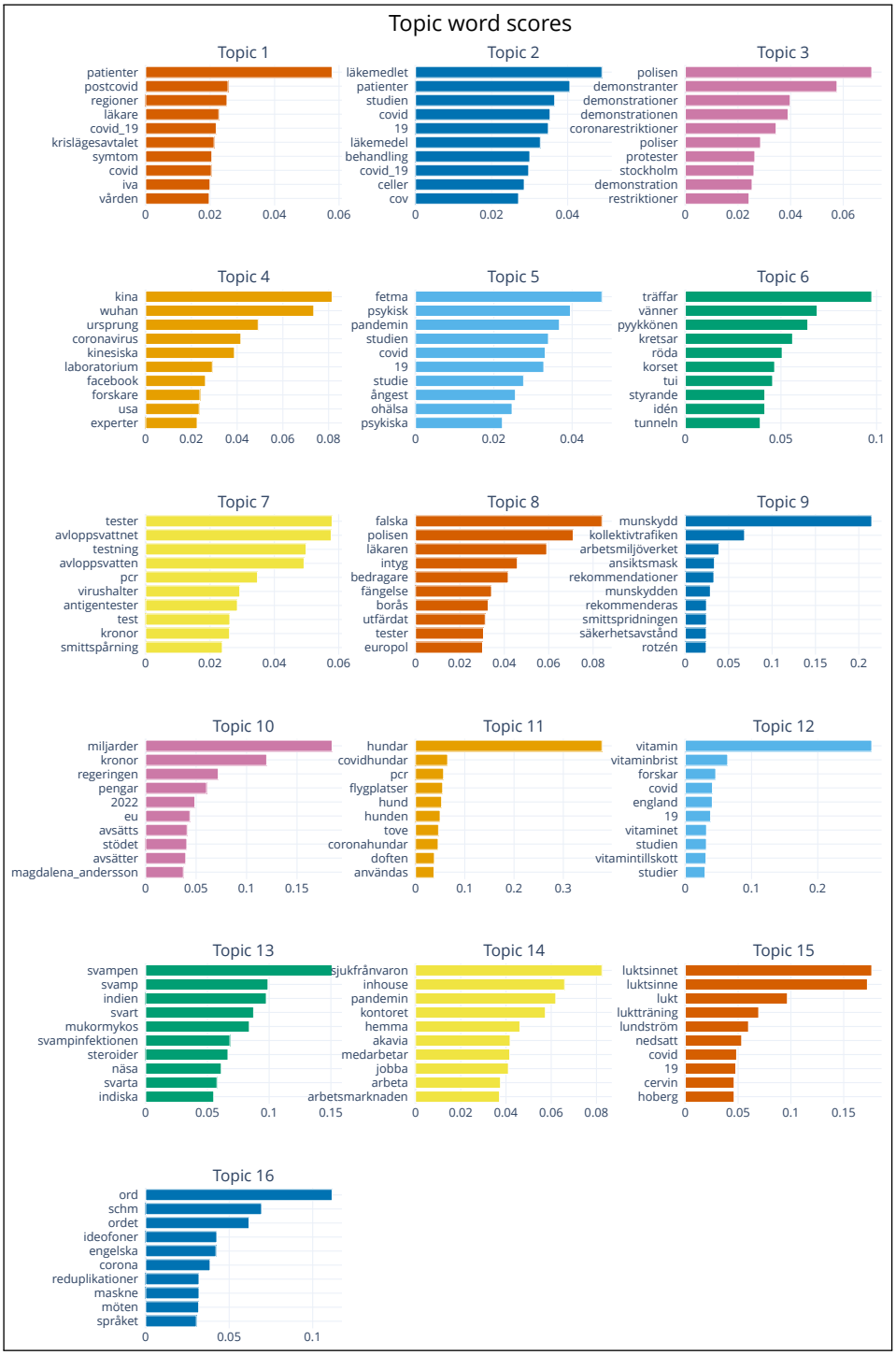


Figure 10: Topic word scores for the generated topics.

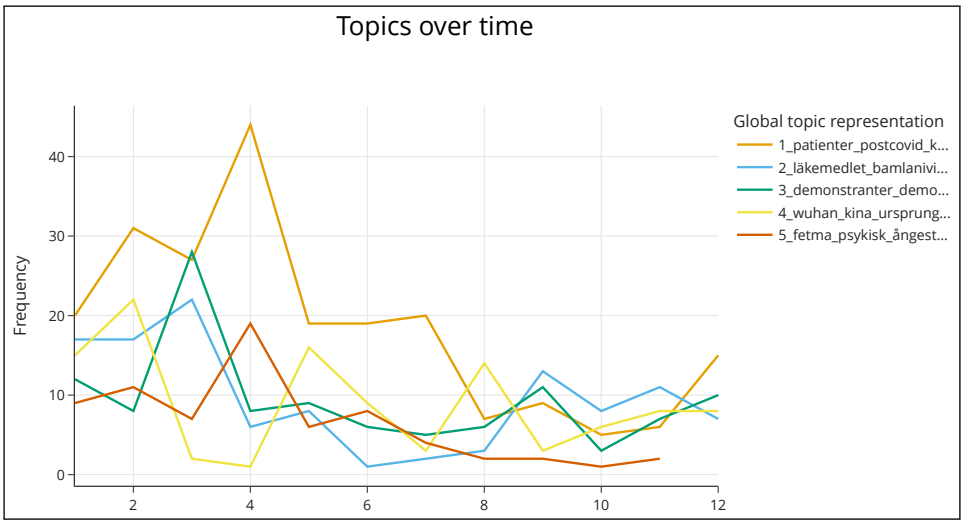


Figure 11: The distribution of the top 5 most common topics over time.

the five most common topics throughout the year. All of these "big topics" decrease in frequency, suggesting that the discourse becomes less dominated by them towards the end of the year, and more diverse.

3 *Final remarks*

In this chapter, we have given several basic examples of using interactive notebooks to perform text exploration tasks. Notebooks can be a powerful tool for researchers in Digital Humanities as pedagogical, analytical, and scholarly tools (De Keulenaar et al. in press; Talboom & Bell 2022, Chun & Elkins 2023), offering a flexible and efficient environment with rich textual documentation alongside the code, ease of collaboration and code interactivity and data visualization capabilities to help convey findings and insights. The ease with which interactive notebooks can be shared and published can also improve reproducibility, allowing other researchers to reproduce results, build upon existing work, and contribute to a collective understanding. This transparency is important in ensuring the robustness and credibility of humanities research.

Many tasks covered in this chapter, such as trend analysis of selected words or part-of-speech tagging, can be performed using dedicated tools like

AntConc[24], LancsBox[25], or the Korp and Mink tools from Språkbanken Text[26]. The notebooks presented in this chapter belong instead to a more general coding approach, suitable for those who wish to customize and automate their analysis. These tools are valuable for quick, hands-on exploration, but coding provides more flexibility for complex or large-scale analyses.

Acknowledgments

The research presented here was supported by *Huminfra*, a Swedish national infrastructure for the Humanities, funded by the Swedish Research Council and the consortium nodes (grant numbers 2021-00176 and 2023-00171), as well as by the Swedish national research infrastructure *Språkbanken* – jointly funded by its 10 partner institutions and the Swedish Research Council (grant numbers 2017-00626 and 2023-00161).

References

Ahnert, Ruth, Emma Griffin, Mia Ridge & Giorgia Tolfo. 2023. *Collaborative historical research in the age of big data: Lessons from an interdisciplinary project* (Elements in Historical Theory and Practice). Cambridge University Press. DOI: 10.1017/9781009175548.

Alderson, David. 2021. Interactive computing for accelerated learning in computation and data science. *INFORMS Transactions on Education* 22(2). 130–145. DOI: 10.1287/ited.2021.0261.

Barba, Lorena, Lecia Barker, Douglas Blank, Jed Brown, Allen Downey, Timothy George, Lindsey Heagy, Kyle Mandli, Jason Moore, David Lippert, Kyle Niemeyer, Ryan Watkins, Richard West, Elizabeth Wickes, Carol Willing & Michael Zingale. 2019. *Teaching and learning with Jupyter*. https://jupyter4edu.github.io/jupyter-edu-book/index.html.

Bednarek, Monika, Martin Schweinberger & Kelvin Lee. 2024. Corpus-based discourse analysis: From meta-reflection to accountability. *Corpus Linguistics and Linguistic Theory* 20(3). 539–566. DOI: 10.1515/cllt-2023-0104.

Blanke, Tobias, Giovanni Colavizza & Zarah van Hout. 2023. An open educational resource to introduce data analysis in python for the humanities. *Education for Information* 39(2). 105–119. DOI: 10.3233/EFI-230020.

24 https://www.laurenceanthony.net/software/antconc/
25 https://lancsbox.lancs.ac.uk/
26 https://spraakbanken.gu.se/en/tools

Chun, Jon & Katherine Elkins. 2023. The crisis of Artificial Intelligence: A new Digital Humanities curriculum for human-centred AI. *International Journal of Humanities and Arts Computing* 17(2). 147–167. DOI: 10.3366/ijhac.2023.0310.

de Marneffe, Marie-Catherine, Christopher D. Manning, Joakim Nivre & Daniel Zeman. 2021. Universal dependencies. *Computational Linguistics* 47(2). 255–308. DOI: 10.1162/coli_a_00402.

Dombrowski, Quinn, Tassie Gniady & David Kloster. 2019. Introduction to Jupyter Notebooks. *Programming Historian* 8. DOI: 10.46430/phen0087.

Evert, Stefan. 2009. Corpora and collocations. In *An international handbook*. Anke Lüdeling & Merja Kytö (eds.). Berlin: De Gruyter Mouton. Chap. 58, 1212–1248. DOI: 10.1515/9783110213881.2.1212. https://doi.org/10.1515/9783110213881.2.1212.

Firth, John. 1957. A synopsis of linguistic theory, 1930–55. In *Studies in linguistic analysis. Special volume of the Philological Society*, 1–32. Oxford: Blackwell.

Granger, Brian & Fernando Pérez. 2021. Jupyter: Thinking and storytelling with code and data. *Computing in Science & Engineering* 23(2). 7–14. DOI: 10.1109/MCSE.2021.3059263.

Grootendorst, Maarten. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure.

Haffenden, Chris & Justyna Sikora. 2025. Doing digital research at KBLab: A practical introduction to using the National Library of Sweden's data lab. In Gerlof Bouma, Dana Dannélls, Dimitrios Kokkinakis & Elena Volodina (eds.), *Huminfra handbook: Empowering digital and experimental humanities* (NEALT Proceedings Series 59), 17–56. University of Tartu Library. DOI: 10.58009/aere-perennius0171.

Hardebolle, Cécile. 2023. *Online interactive textbooks with Jupyter Notebooks*. https://www.epfl.ch/education/educational-initiatives/cede/teaching-interactively/jupyter-notebooks-for-education/teaching-and-learning-with-jupyter-notebooks/online-interactive-textbooks-with-jupyter-notebooks/.

Jelodar, Hamed, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li & Liang Zhao. 2019. Latent Dirichlet allocation (LDA) and topic modeling: Models, applications, a survey. *Multimedia Tools and Applications* 78(11). 15169–15211. DOI: 10.1007/s11042-018-6894-4.

Karsdorp, Folgert, Mike Kestemont & Allen Riddell. 2021. *Humanities data analysis: Case studies with Python*. See https://www.humanitiesdataanalysis.org/ for the book's contents in notebook form. Princeton University Press.

de Keulenaar, Emillie, Thomas Poell, Anne Helmond, Bernhard Rieder & Jasmijn Van Gorp. In press. Computational cross-media research: Trac-

ing divergences between normative Dutch television and social media discourses on the 'refugee crisis' (2013-2018). *Convergence*. DOI: 10.1177/13548565241258956.

Knuth, D. E. 1984. Literate programming. *The Computer Journal* 27(2). 97–111. DOI: 10.1093/comjnl/27.2.97.

Kokkinakis, Dimitrios. 2021. Insights on a Swedish covid-19 corpus. In Monica Monachini & Maria Eskevich (eds.), *CLARIN Annual Conference proceedings*, 31–34. https://office.clarin.eu/v/CE-2021-1923-CLARIN2021_ConferenceProceedings.pdf.

Land, Kaylin, Geoffrey Rockwell & Andrew MacDonal. 2021. Spyral notebooks as a supplement to voyant tools. In *CSDH-SCHN 2021: Making the net work, virtual*. Canadian Society for Digital Humanities. DOI: 10.17613/2bsr-xp53.

Melgar-Estrada, Liliana, Marijn Koolen, Kaspar Beelen, Hugo Huurdeman, Mari Wigham, Carlos Martinez-Ortiz, Jaap Blom & Roeland Ordelman. 2019. The CLARIAH Media Suite: A hybrid approach to system design in the humanities. In *Proceedings of the 2019 conference on human information interaction and retrieval*, 373–377. Glasgow, Scotland UK: Association for Computing Machinery. DOI: 10.1145/3295750.3298918.

Nusko, Bianka, Nina Tahmasebi & Olof Mogren. 2016. Building a sentiment lexicon for Swedish. In *Digital humanities 2016. from digitization to knowledge 2016: Resources and methods for semantic processing of digital works/texts*, vol. 126 (Linköping Electronic Conference Proceedings 6), 32–37. LiU Electronic Press. https://ep.liu.se/ecp/126/006/ecp16126006.pdf.

Öhman, Emily. 2021. The validity of lexicon-based sentiment analysis in interdisciplinary research. In Mika Hämäläinen, Khalid Alnajjar, Niko Partanen & Jack Rueter (eds.), *Proceedings of the workshop on natural language processing for digital humanities*, 7–12. NIT Silchar, India: NLP Association of India (NLPAI). https://aclanthology.org/2021.nlp4dh-1.2/.

Piantadosi, Steven. 2014. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review* 21(5). 1112–1130. DOI: 10.3758/s13423-014-0585-6.

Rekathati, Faton. 2021. *Introducing a Swedish sentence transformer*. The KBLab Blog. https://kb-labb.github.io/posts/2021-08-23-a-swedish-sentence-transformer/.

Rule, Adam, Aurélien Tabard & James Hollan. 2018. Exploration and explanation in computational notebooks. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, 1–12. New York, NY, USA: Association for Computing Machinery. DOI: 10.1145/3173574.3173606.

Schmidt, Benjamin. 2012. Words alone: Dismantling topic models in the Humanities. *Journal of Digital Humanities* 2(1). 49–66.

Språkbanken Text. 2017. *Sentimentlexikon*. Dataset. DOI: 10.23695/1yyf-6409.
Talboom, Leontien & Mark Bell. 2022. Keeping it under lock and keywords: Exploring new ways to open up the web archives with notebooks. *Archival Science* 22(3). 393–415. DOI: 10.1007/s10502-022-09391-6.
York, Christopher. 2017. Exploratory data analysis for the Digital Humanities: The Comédie-Française registers project analytics tool. *English Studies* 98(5). 459–482. DOI: 10.1080/0013838X.2017.1332024.
Zipf, George Kingsley. 1949. *Human behavior and the principle of least effort: An introduction to human ecology*. New York: Addison-Wesley.

List of abbrevations

BERT	Bidirectional Encoder Representations from Transformers
Colab	Google Colaboratory
LLR	Log-likelihood ratio
NLP	Natural Language Processing
NPMI	Normalized pointwise mutual information
PMI	Pointwise Mutual Information

Corresponding author

Dimitrios Kokkinakis
Språkbanken Text
Department of Swedish,
Multilingualism, Language
Technology
University of Gothenburg
dimitrios.kokkinakis
@svenska.gu.se

A PRACTICAL GUIDE TO HUMINFRA

Research in the Humanities is increasingly interdisciplinary, quantitative, and multi-methodological. The Swedish national research infrastructure Huminfra brings together 12 nodes in the domains of digital and experimental Humanities. The goal is to facilitate the discovery of and access to existing Swedish instruments, tools, expertise and materials, to provide training in relevant techniques and methods, and to host the Swedish membership in the pan-European Digital Research Infrastructure for the Arts and Humanities (DARIAH).

The Huminfra Handbook is the result of Huminfra's first conference, HiC 2024. Designed as a practical textbook, it presents user-friendly tools, infrastructural components and introductions to some of the labs and units within Huminfra.

Whether you are a beginner or an experienced researcher, this handbook offers guidance on how to make the most of the resources available through Huminfra.

Find available resources here!