# 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

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 |

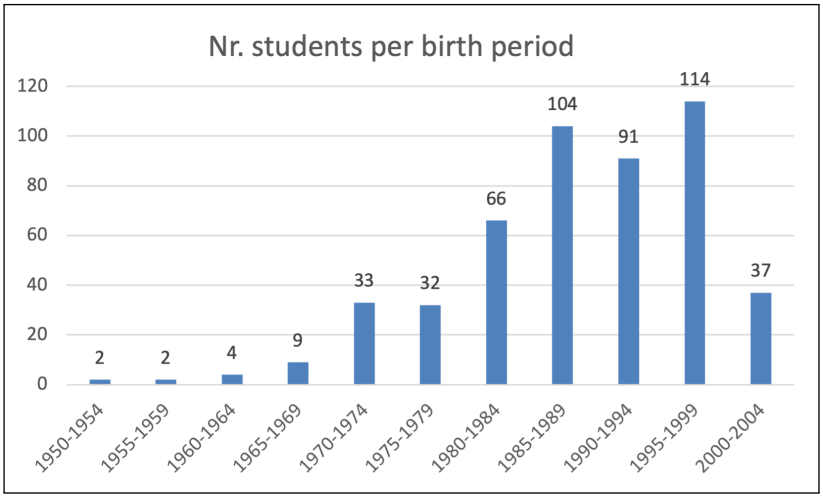



Figure 1: Distribution of age groups in the SweLL-gold data. All essays were collected during 2017–2021.

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).

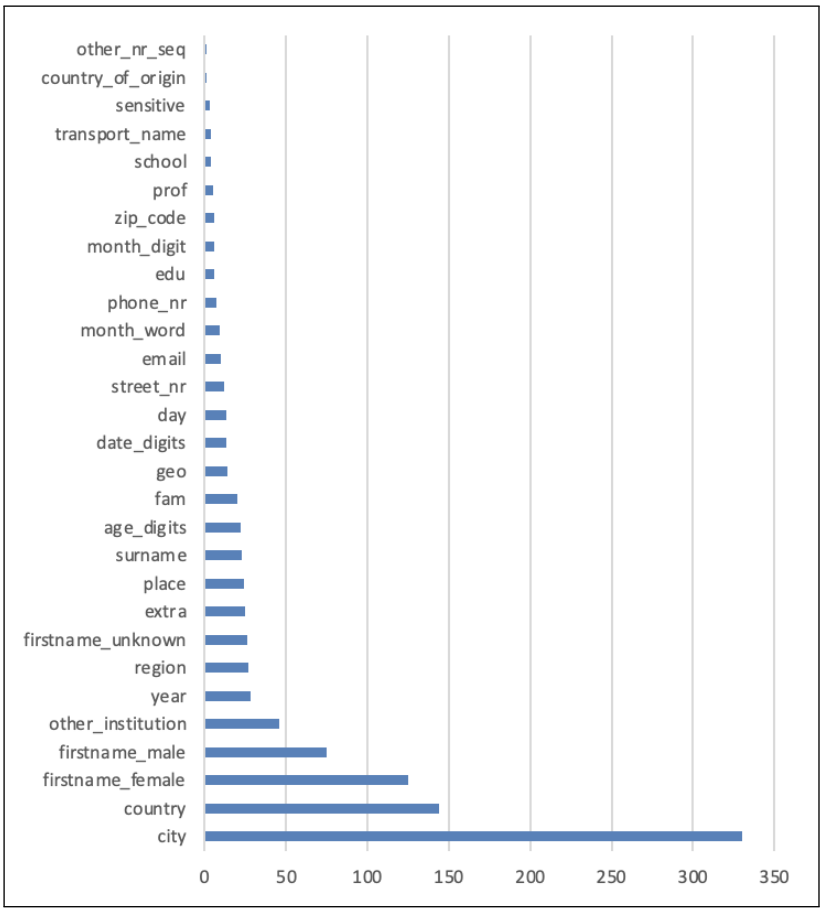



Figure 2: Overview of pseudonymization tags in the SweLL-gold data.

### 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.

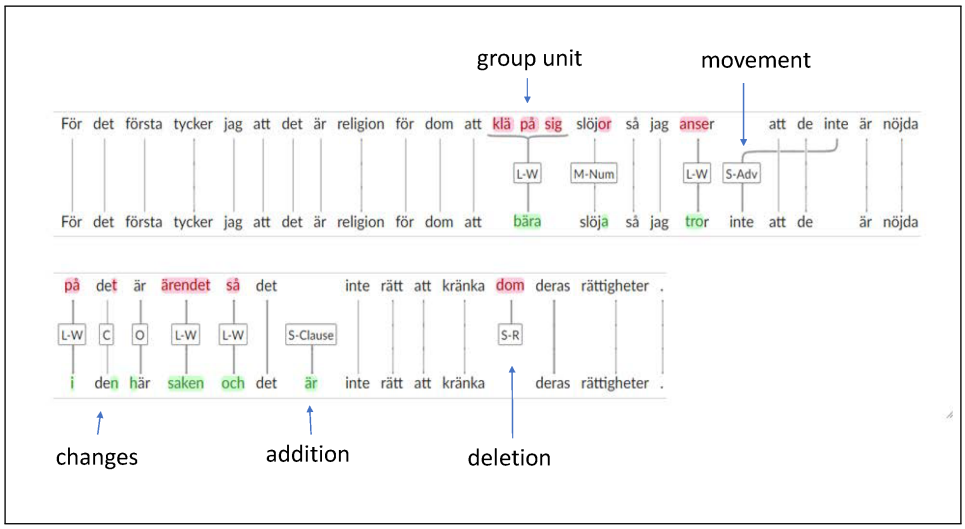



Figure 3: Types of correction and unit types in the SVALA annotation tool. The top level represents an original learner text, the lower level represents its normalized version.

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).

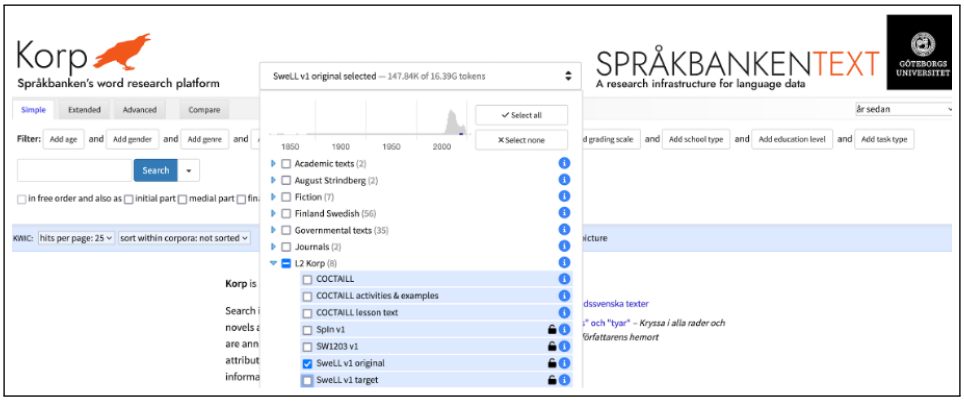


Figure 7: Corpus selection in Korp user interface.

## 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/

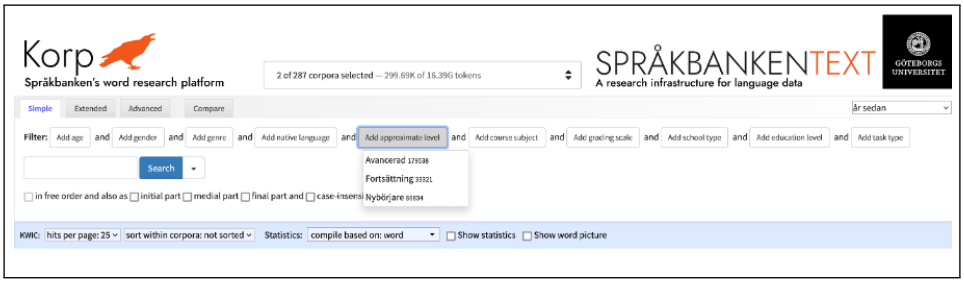



Figure 8: L2-specific filter in Korp user interface.

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 |

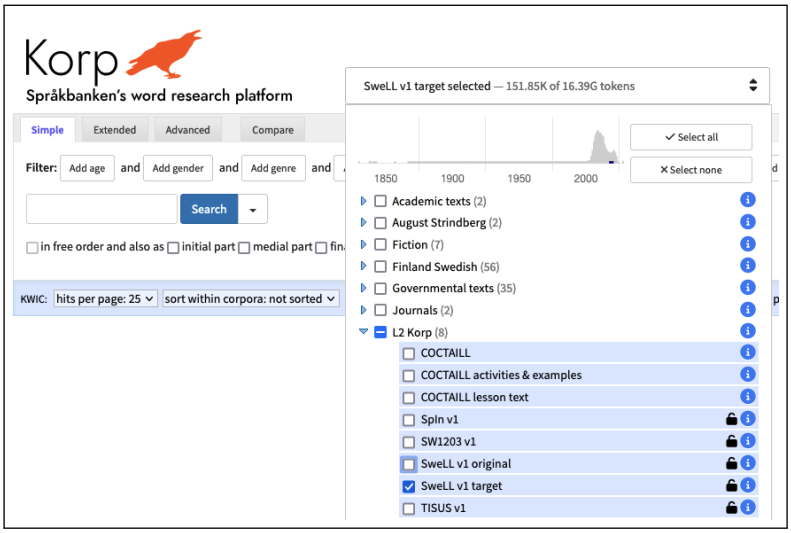



Figure 9: Select "SweLL v1 target" to compare frequency of use for the three verbs in normalized texts to usage in original texts.

"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

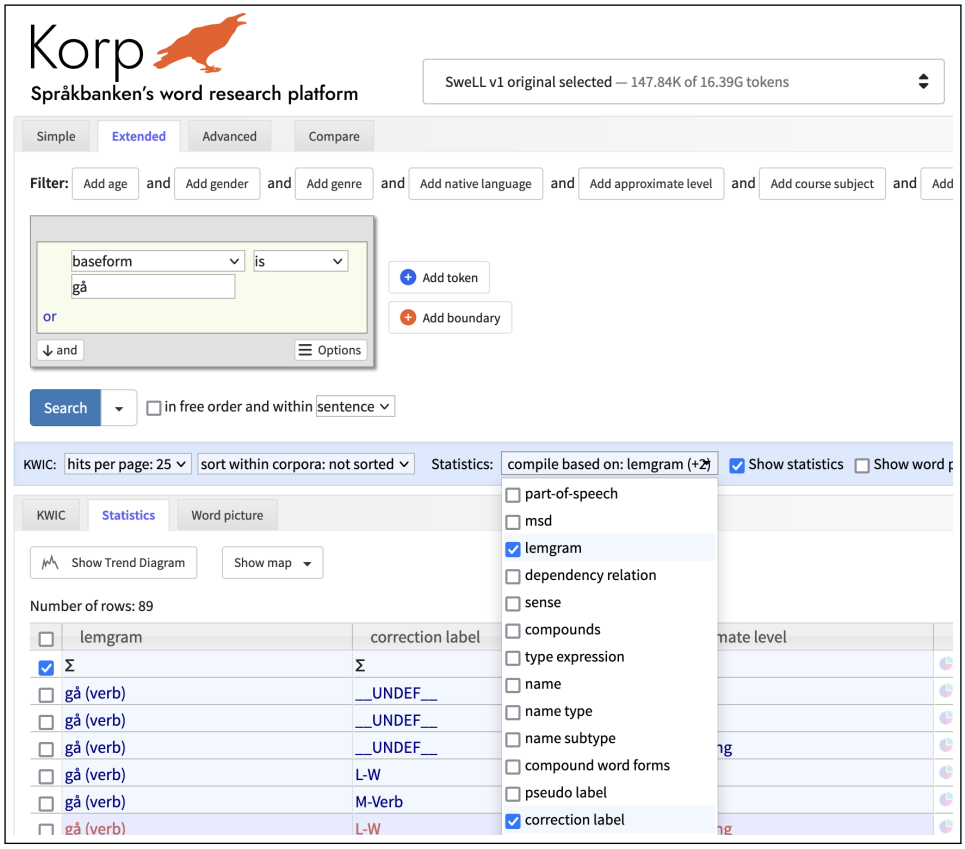



Figure 10: Define a search for base forms and compile statistics for *lemgram* (a combination of baseform and part of speech), correction label and approximate level.

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

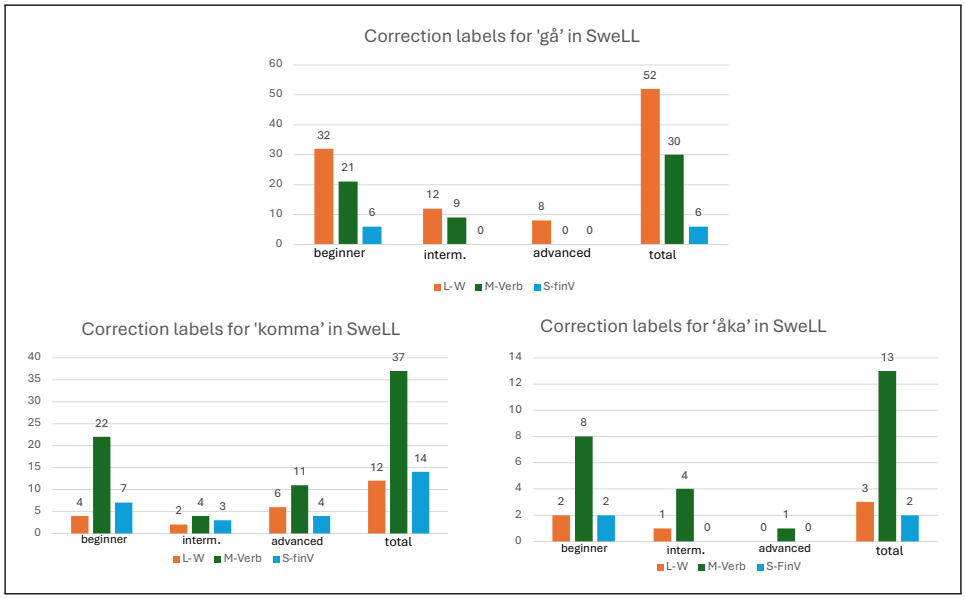



Figure 11: Frequency of correction labels for *gå*, *komma* and *åka* in SweLL-gold original at the three learner levels.

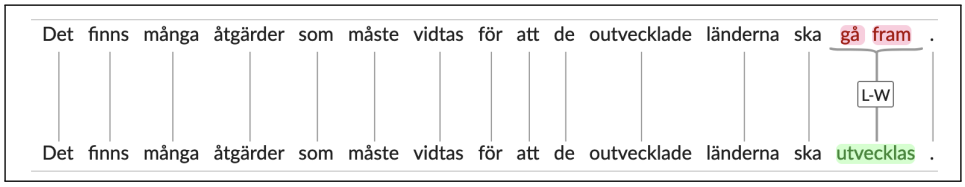



Figure 12: Target hypothesis *utvecklas* 'evolve' for *gå fram* 'go forward' in SweLL-gold (Advanced level).

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

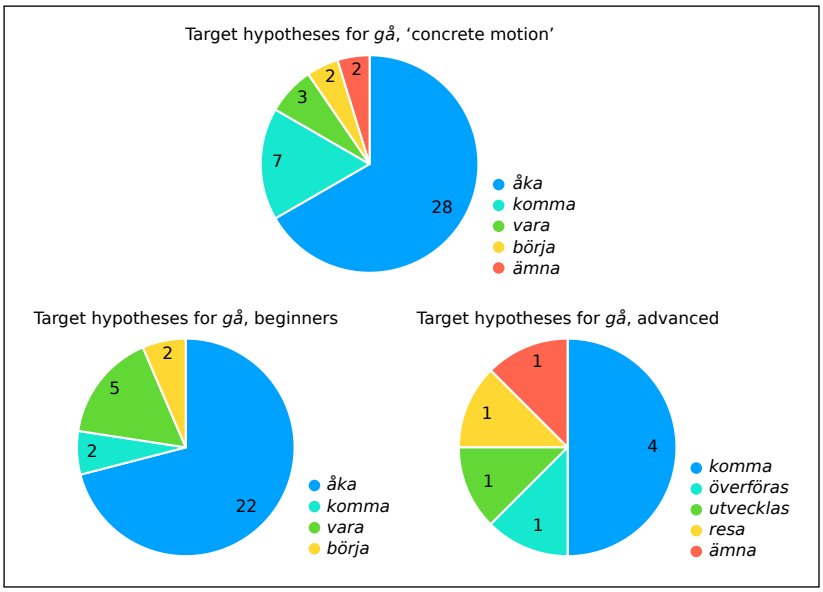



Figure 13: Target hypotheses for *gå* in SweLL target.

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
[...]
```

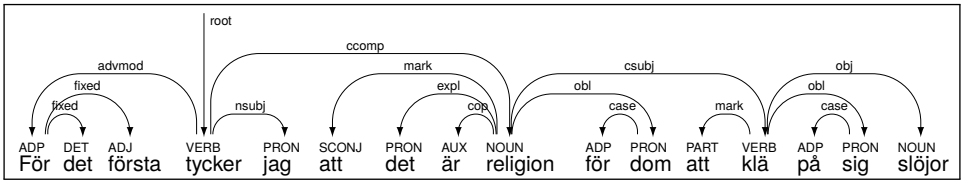
Figure 18: Excerpt from the test split of the dataset used in the MultiGEC-2025 shared task. Full essays are stripped from their metadata (except for the sentence ID) and stored in two separate Markdown files, one (partially reproduced here) containing original essays, and one containing their normalized 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