10 From text to insight: Uncovering linguistic patterns with SWEGRAM

Beáta Megyesi
Stockholm University

Rex Ruan
Stockholm University

Empirical linguistic analysis provides valuable insights into textual data for researchers in the humanities and social sciences, enabling them to identify patterns and trends within large datasets. SWEGRAM is a freely available tool designed to annotate and analyze Swedish and English texts without requiring programming skills or a user account. Users can upload one or more texts for linguistic analysis, extracting morphological and syntactic features. The linguistically annotated texts can then be used for quantitative linguistic analysis, allowing researchers to systematically explore textual characteristics. Additionally, the tool visualizes syntactic relations between words in sentences and provides detailed insights into the distribution of syntactic functions and relations within the text. Users can also create their own linguistically annotated text collections and generate statistical summaries of the linguistic properties of their texts. The tool is available as both a web-based service, which requires no user login or account, and a downloadable version for local use when data privacy and security are a priority. This dual availability ensures accessibility and flexibility for diverse research needs.

1 *Introduction*

Linguistic annotation and text analysis play a crucial role in the humanities and social sciences, offering a systematic approach to capturing and examining language use across different contexts. By assigning linguistic features—such as morphological, syntactic, semantic, and discourse-related properties—to words, phrases, sentences, paragraphs, and texts, researchers can identify patterns and trends that deepen our understanding of human communication, cultural practices, and social interaction. This approach

HUM INFRA Beáta Megyesi & Rex Ruan. 2025. From text to insight: Uncovering linguistic patterns with SWEGRAM. In Gerlof Bouma, Dana Dannélls, Dimitrios Kokkinakis & Elena Volodina (eds.), *Huminfra handbook: Empowering digital and experimental humanities* (NEALT Proceedings Series 59), 307–353. University of Tartu Library. DOI: 10.58009/aere-perennius0179

allows for the creation of large, annotated corpora that can be analyzed quantitatively and qualitatively, offering insights into language use, social dynamics, and culture. The empirical foundation of these linguistic annotations ensures that analyses are grounded in actual language use, which is essential for developing robust theories and models in linguistics, history, history of ideas, sociology, psychology, anthropology, and related fields.

Moreover, empirical linguistic analysis facilitates interdisciplinary research, bridging gaps between traditional humanities scholarship and the social sciences. For instance, sociolinguistic studies that combine annotated linguistic data with demographic information can reveal how language reflects and reinforces social stratification or how linguistic practices vary across different communities. Similarly, in digital humanities, the analysis of annotated texts enables the exploration of literary styles, authorship attribution, and the evolution of genres over time. This empirical approach not only enriches our understanding of language as a social and cultural phenomenon but also enhances the methodological rigor of research in the humanities and social sciences, making it possible to address complex questions with greater precision and depth.

Several powerful tools are available for the automatic annotation and analysis of texts (e.g., Bird et al. 2009, Kilgarriff et al. 2014, Manning et al. 2014, Honnibal & Montani 2017), each offering unique features tailored to different research needs. In this chapter, we describe SWEGRAM,[1] a web-based tool designed for the automatic annotation and linguistic analysis of texts in both English and Swedish. It not only facilitates detailed linguistic analysis but also allows users to create and manage their own text corpora, thus serving as a valuable tool for researchers working with language data.

SWEGRAM was originally developed for Swedish (Näsman et al. 2017, Megyesi et al. 2019) and through a collaboration between the Department of Linguistics and Philology and the Department of Scandinavian Languages at Uppsala University, Sweden. The tool was adapted later to English. Since 2024, SWEGRAM has been hosted by the Department of Linguistics at Stockholm University, Sweden. SWEGRAM has been an integral part of the Swe-CLARIN[2] project, which serves as the Swedish node of the European CLARIN infrastructure, with financial support from the Swedish Research Council. The primary goal of SWEGRAM is to make language-based text material accessible as primary research data for the humanities and social sciences, utilizing advanced processing tools for both written and transcribed spoken language.

1 http://swegram.ling.su.se/
2 https://sweclarin.se

In the following sections, we will introduce the components of the tool and provide guidance on how it can be utilized for empirical linguistic analysis. We begin by outlining some related work to set the context.

2 *Related work*

Numerous robust tools have been developed to support the automatic annotation and analysis of texts, each equipped with unique features that cater to diverse research needs in the humanities and social sciences. Tools like Stanford CoreNLP (Manning et al. 2014) and spaCy (Honnibal & Montani 2017) offer comprehensive natural language processing (NLP) pipelines capable of performing a wide range of tasks, including tokenization, part-of-speech tagging, named entity recognition, and syntactic parsing. These tools are highly scalable and integrate seamlessly into larger data analysis workflows, making them indispensable for complex linguistic studies.

The Natural Language Toolkit (NLTK) (Bird et al. 2009) is another widely used resource, particularly valued in educational and research settings for its extensive documentation and user-friendly design. NLTK supports a broad spectrum of text processing tasks, while specialized tools like VADER (Hutto & Gilbert 2014) or MALLET (McCallum 2002) excel in areas such as sentiment analysis and topic modeling. These tools are especially well-suited for analyzing social media data and conducting large-scale text mining projects. In the realm of spoken language annotation, ELAN (Wittenburg et al. 2006) and EXMARaLDA (Schmidt & Wörner 2014) are popular for their ability to synchronize audio or video recordings with transcriptions, annotations, and metadata. Collectively, these tools significantly enhance the efficiency and precision of linguistic annotation and analysis, empowering researchers to handle large volumes of text data and uncover meaningful patterns.

Beyond these specific tools, a variety of others are available for different types of text analysis. Xaira (Burnard 2006) is an open-source software package that supports the indexing and analysis of corpora using Extensible Markup Language (XML); originally developed for the British National Corpus (BNC) (BNC Consortium 2007), it has become a standard tool for corpus-based research. Similarly, BNCWeb (Hoffmann et al. 2008) offers a web-based interface for the British National Corpus, enabling users to conduct in-depth searches and analyses.

Concordance programs such as AntConc (Laurence 2011), ProtAnt (Anthony & Baker 2015), The Sketch Engine and WebCorp (Kilgarriff et al. 2014) are also popular in the field of text analysis, valued for their ability to dis-

play various text-related features like word frequencies, collocations, and keywords. WordSmith Tools (Scott 2016) is another widely used suite of text analysis software, offering functionalities that include the creation of word lists, concordance lists, clusters, collocations, and keyword analysis. These tools collectively contribute to a comprehensive ecosystem for text analysis, supporting both qualitative and quantitative research across different languages and text types.

In the context of Swedish linguistic research, several specialized tools have been developed to facilitate the automatic annotation and analysis of Swedish texts. SPARV, developed by Språkbanken at the University of Gothenburg (Borin et al. 2016, Hammarstedt et al. 2022), is a powerful tool offering a comprehensive suite of annotation services tailored for Swedish. It provides functionalities such as tokenization, sentence segmentation, morphological analysis, named entity recognition, and dependency parsing. SPARV's high degree of customizability allows users to fine-tune the annotation pipeline according to specific research requirements.

Another notable tool from Språkbanken is Korp (Borin et al. 2012), a corpus search engine that grants access to extensive annotated corpora of Swedish texts. Korp enables users to conduct complex searches across various linguistic features, including word forms, lemmas, and syntactic structures, facilitating in-depth analysis and exploration of language usage in corpora that are available through Språkbanken's web portal.

In parallel with the development of Korp which later also featured an annotation pipeline called Annotation Lab (Borin et al. 2012), the first version of SWEGRAM (Näsman et al. 2017, Megyesi et al. 2019) was released in 2015. At that time, there was no tool available for the automatic annotation of Swedish that was accessible to researchers without programming skills. SWEGRAM aimed to enrich the landscape of Swedish linguistic tools by offering comprehensive analysis capabilities for Swedish and later for English texts, providing services such as tokenization, part-of-speech tagging, lemmatization, and syntactic parsing. Another equally important feature of SWEGRAM is its ability to conduct linguistic analysis based on annotations within the same tool, streamlining the research process. Additionally, SWEGRAM allows users to create and manage their own corpora, making it a versatile resource for both corpus creation and analysis. Over the past few years, SWEGRAM has been updated and recently released as SWEGRAM 2.0 with new linguistic features, enhanced visualization options, and a redesigned user interface, further solidifying its utility in linguistic research, which we will describe in the subsequent sections.

In summary, linguistic annotation tools are crucial for advancing research, offering robust resources for both small-scale and large-scale studies. These

tools streamline text processing and enable detailed, scalable linguistic investigations, benefiting the broader humanities and social sciences. The array of tools for automatic annotation and analysis enhances researchers' ability to conduct large-scale empirical studies.

3 *SWEGRAM*

SWEGRAM (Näsman et al. 2017, Megyesi et al. 2019) was developed to provide automatic annotation and analysis of texts, enabling empirical linguistic analysis of large text corpora within a single tool. The primary goal was to create a web-based platform that allows users to perform annotation and statistical linguistic analysis without requiring any programming skills. Emphasis was placed on making the tool user-friendly, offering a web service that allows users to upload texts without needing to log in or provide personal information. The tool is designed to be easily accessible for researchers and others interested in the humanities and social sciences and requires no knowledge of automated text processing, nor any programming skills.

In SWEGRAM, the user can upload one or more texts and have them linguistically analyzed for morphological and syntactic features. These linguistically annotated texts can then be used to perform quantitative linguistic text analysis or distant reading with the tool providing statistics on sentence length, total word count, readability metrics, part-of-speech (PoS) distribution, and the frequency of lemmas, PoS categories, or misspelled words. Additionally, SWEGRAM supports qualitative text analysis in a close reading scenario, allowing users to closely examine linguistic patterns and structures within individual texts or across a corpus. The tool visualizes the syntactic relationships between words in sentences and provides detailed information about the distribution of various syntactic functions and relations in the text. Moreover, users can easily compile their own linguistically annotated text collection (corpus) and generate comprehensive statistics on its linguistic characteristics.

SWEGRAM 1.0, initially introduced in 2015, has since been employed in numerous research projects (Megyesi et al. 2016, 2019). Recently, SWEGRAM has been updated to support the processing of English texts. In addition, we have expanded the range of linguistic features for both English and Swedish, incorporating readability metrics for English texts and second language acquisition analysis, inspired by Ildikó Pilán's PhD thesis (Pilán 2018). The visualization of morphosyntactic analysis has also been enhanced, now offering graphical representations of syntax trees for individual sentences. SWEGRAM now features an upgraded back-end, enabling faster and more

efficient text processing. In the latest version, even the analysis of relatively large texts—though still requiring considerable processing time—can be displayed incrementally. In the online version, this is facilitated by temporarily storing annotated texts in a database, reducing the burden of data loading during processing. These temporary records are automatically removed after a few days. Users are fully responsible for ensuring compliance with copyright and any other applicable restrictions when uploading data to the system. For non-public or sensitive data, SWEGRAM can be run locally using the downloadable version (see Section 6).

Next, we give an overview of the various functionalities of SWEGRAM 2.0, from text upload and linguistic annotation to text analysis, visualization of features, and exporting annotated text and statistics for further processing.

4 *Using SWEGRAM*

SWEGRAM comprises two major components, the annotation taking care of the labeling on various linguistic levels and the analysis part, deriving statistics of a wide range of linguistic features. Figure 1 illustrates the two major components of the tool. SWEGRAM offers functionalities for language selection, text upload, linguistic annotation pipeline, linguistic analysis, and the export of text and data files. We provide an overview of each. A more detailed description of these functionalities can be found in the SWEGRAM 2.0 Guidelines (Megyesi & Ruan 2024).

4.1 *Preparing for processing*

Before processing begins, the user must select the language to be processed: English or Swedish. Depending on the chosen language, the system applies the appropriate language-specific pipeline. Then, one or more files can be uploaded for annotation (up to 10 MB raw text), preferably in the Universal Character Encoding Standard Unicode UTF-8 format. Alternatively, the user can choose to paste text directly into the designated field, as illustrated in Figure 2. The uploaded files are displayed in the text selection area under the *Statistics* or *Visualize* modules.

To enable text comparison, the uploaded texts can also contain optional metadata of any type. Each line of metadata must follow a certain format to be processed correctly by the tool. The format is illustrated in Figure 3a. Each line of the metadata must start with < and end with >. Inside the metadata line, the various types of information are structured as `FEATURE:VALUE` pairs, with a colon (`:`) as a separator between the feature and its value. Several

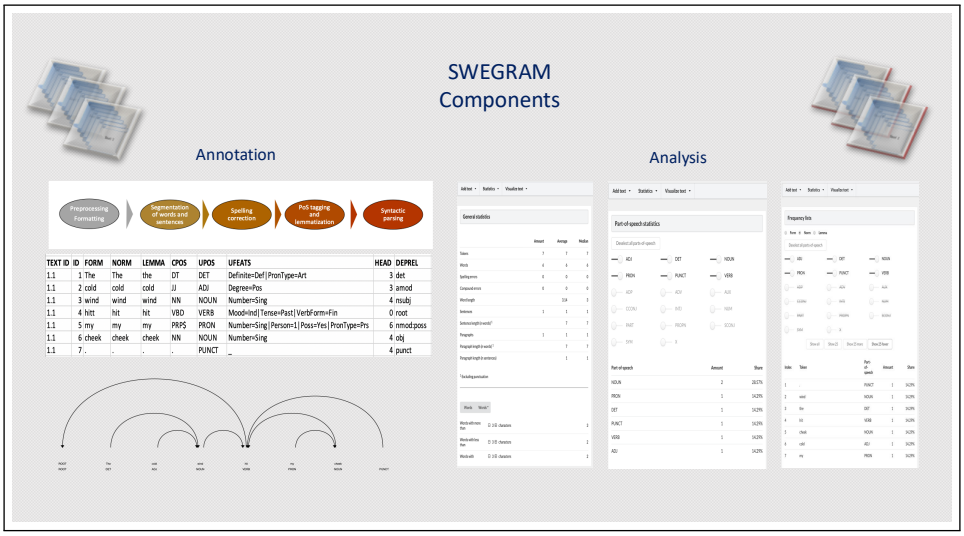

Figure 1: The SWEGRAM components.

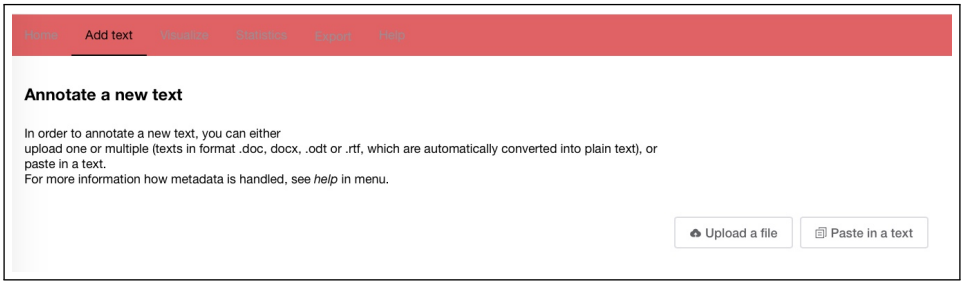

Figure 2: Upload texts.

pairs are separated by a semicolon (`;`). For example, a metadata line might appear as follows:

```
<AUTHOR:John Joe; GENRE:Narrative; YEAR:2023>
```

In this example, `AUTHOR`, `GENRE`, and `YEAR` are features, while `John Joe`, `Narrative`, and `2023` are their corresponding values.

If the uploaded file contains several texts, the metadata is displayed at the beginning of each text, as shown in Figure 3b. The metadata line can be also used for text selection, as illustrated in Figure 4. The system filters texts that match the specified feature-value pairs and proceeds with them for further analysis.

Lines can be added as additional information without being treated as part of the text to be analyzed by marking them with a hashtag (#). SWEGRAM excludes all lines starting with a hashtag during processing.

```
1 <FEATURE1:VALUE1;FEATURE2:VALUE2>
2
3 This is a text to be annotated and analyzed.
```

a. *Plain text*

b. *Visualized text with metadata*

Figure 3: Metadata label format in text files and in analysis.

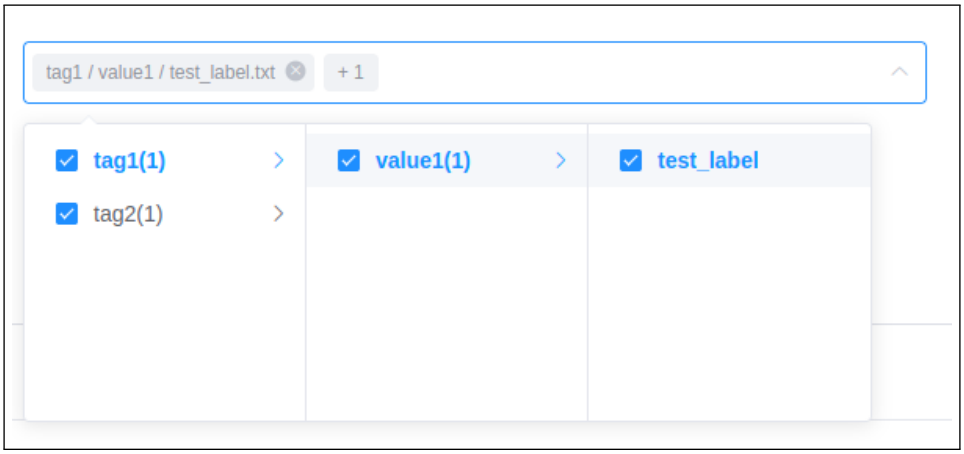

Figure 4: Metadata for selection.

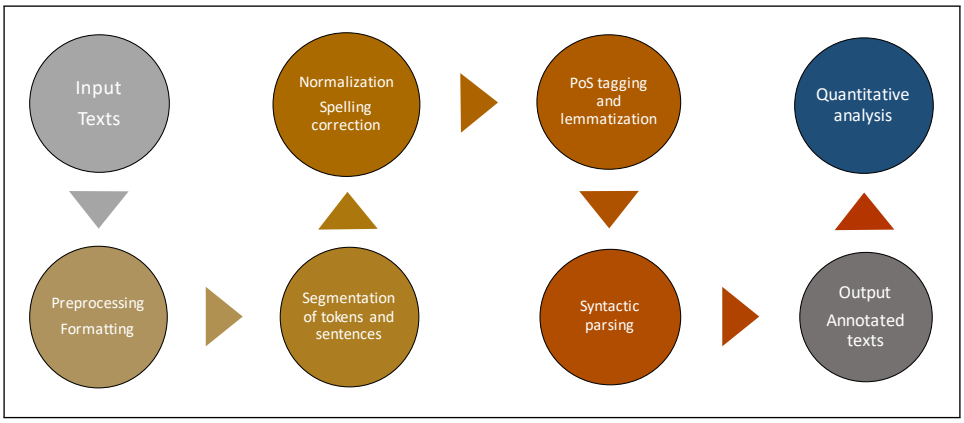

Figure 5: The components for linguistic annotation in SWEGRAM.

SWEGRAM can also be used to modify or add annotations to an already annotated text by selecting the "This is an annotated text" option. For this purpose, the CSV file format is accepted. When enabling this option, the uploaded file will not go through the standard annotation process; instead, the system will call for the annotation checker to ensure that the file is in the required format. All errors detected by the annotation checker are displayed with an index reference, indicating the type of errors that the file might have.

4.2 *Linguistic annotation*

Once the files are uploaded, linguistic annotation can be performed. SWEGRAM provides a comprehensive chain of tools to annotate a text on various linguistic levels. These include tokenization for the segmentation of words and sentences, normalization to correct misspelled words, and PoS tagging/-parsing for the morpho-syntactic analysis of sentences.

The components of the annotation pipeline are illustrated in Figure 5. This pipeline is built on a chain of automated tools for the analysis of the English and Swedish languages. We use state-of-the-art tools developed in the fields of natural language processing to automatically process and annotate texts with proven accuracy.

The system processes all running texts in the file except the lines marked with a hashtag (#).

Texts annotated using the tool are represented in a standardized format, with annotations at various linguistic levels adhering to specific guidelines. We use Unicode (UTF-8) for character encoding. To represent annotations at various linguistic levels and comply with international standards, we use

Table 1: Annotation representation.

TEXT-ID	Paragraph and sentence index, integer starting at 1 for each new paragraph and sentence.
ID	Token index, integer starting at 1 for each new sentence; may be a range for originally multiword tokens that have been split due to misspelling.
FORM	Word form or punctuation mark, so called token.
NORM	Corrected/normalized token for misspelled words.
LEMMA	Lemma of word form.
UPOS	Part-of-speech based on **Universal PoS tagset**.
XPOS	Part-of-speech based on the **Stockholm-Umeå Corpus** PoS tagset for Swedish and **Penn Treebank** PoS tagset for English.
CFEATS	List of morphological features from the **Stockholm-Umeå Corpus**; "_" if feature is missing. This is only valid for Swedish text analysis.
UFEATS	List of morphological features from the **Universal PoS tagset**; "_" if feature is missing.
HEAD	Head of the current word, which is either a value of ID or zero (0) if the word is the ROOT of the sentence.
DEPREL	Dependency relation to the HEAD based on the **Universal dependency relations**.
MISC	Any other annotation.

the CoNLL-U tab-separated format. Each token (i.e., each separated word and punctuation mark) occupies its own line along with its analysis, with each new sentence beginning on a new line shift.

Every token is analyzed at various levels, which are represented on the same line as the token itself in dedicated columns, separated by tabs. These columns contain both the ID number for the paragraph, sentence, and token, and the linguistic analysis. The linguistic analysis is represented at both the word level, through lemma, PoS annotation and morphological information, and the sentence level, through syntactic analysis. Table 1 provides a summary of annotation representation. On the left-hand side, the names of the features listed for each token in the columns are described, followed by the description of each feature.

Output from the linguistic annotation is illustrated for the sentence *Lixards like to eat bugs.*, which includes the misspelled token *Lixards* that SWEGRAM's normalization module corrected to *Lizards*. The linguistic annotation labels in the CoNLL-U format are shown in Figure 6 and the syntax tree with PoS tags and syntactic functions in Figure 7.

Next, we describe each component involved in the linguistic analysis in detail.

TEXTID	ID	FORM	NORM	LEMMA	UPOS	XPOS	U-FEATS	HEAD	DEPREL	MISC
1.1	1	Lixards	Lizards	Lizard	NOUN	NNS	Number=Plur	2	nsubj	_
1.1	2	like	like	like	VERB	VBP	Mood=Ind|Tense=Pres|VerbForm=Fin	0	root	_
1.1	3	to	to	to	PART	TO	_	4	mark	_
1.1	4	eat	eat	eat	VERB	VB	VerbForm=Inf	2	xcomp	_
1.1	5	bugs	bugs	bug	NOUN	NNS	Number=Plur	4	obj	_
1.1	6	.	_	.	PUNCT	.	_	2	punct	_

Figure 6: Output – linguistic annotation labels in CoNLL-U format.

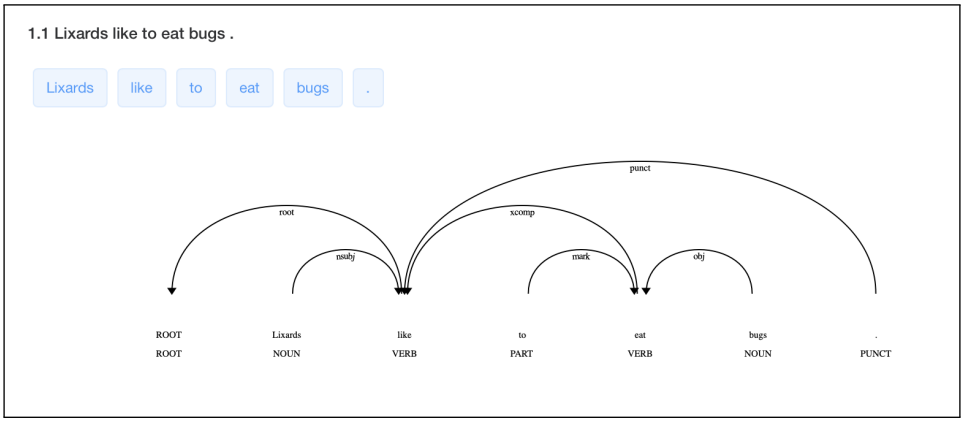

Figure 7: Output – annotated syntax tree.

4.2.1 *Tokenization and sentence segmentation*

In the formatted and uploaded text, words are separated from punctuation and segmented into linguistic units using tokenization, leaving each word and punctuation mark on its own line and each sentence followed by an empty line, see Figure 8. Each token in the sentence is then numbered in numerical order to receive an ID-number. Tokenization and sentence segmentation are based on the tokenization module used for the Swedish Treebank (Cap et al. 2016), which is also a component of EFSELAB (Östling 2016) applied for linguistic analysis.

4.2.2 *Normalization, spelling and separated compound words*

Misspelled words are identified and corrected by the normalization component. The normalized word is added to the analysis and displayed in a separate field to ensure that the original word form is kept and displayed. In Figure 9, the original word form is shown in the `FORM` column, while the corrected, normalized version appears in the `NORM` column. This shows the original and corrected text side by side, token by token.

```
This
is
e.g.
tokenized
.

The
sentences
are
segmented
.
```

Figure 8: An example of a tokenized and sentence segmented text.

English translation	TEXT ID	ID	FORM	NORM
My	1.1	1	Min	Min
	1.1	2–3		inspirationskälla
inspiration	1.1	2	inspirations	
source	1.1	3	källa	
comes	1.1	4	kommer	kommer
from	1.1	5	från	från
text	1.1	6	texten	texten
.	1.1	7	.	.

Figure 9: Normalized words (NORM).

The module also corrects incorrectly separated compound words in Swedish, see Figure 9. In such cases, the corrected form is indexed with a hyphen (2–3) and shown in the NORM column in Figure 9. The corrected version is used for further annotation, and the original form is used for linguistic analysis.

Spelling corrections are handled by a modified version of Hist-Norm (Pettersson et al. 2013), a tool designed for normalizing historical text spellings. While effective, the tool may not correct every error, so manual review is highly recommended for reliable analysis.

4.2.3 *Part-of-speech analysis*

The corrected, normalized text is analyzed using a part-of-speech (PoS) tagger, which annotates each token—both words and punctuation—with its appropriate PoS tag and morphological information. PoS annotation includes

Table 2: Universal PoS tags (UPOS).

Tag	Explanation	Example
ADJ	Adjective	*fine*
ADP	Adposition (preposition)	*on*
ADV	Adverb	*quick*, *very*
AUX	Auxiliary verb	*have*
CCONJ	Conjunction	*and*
DET	Article/Determiner	*a*
INTJ	Interjection	*yeah*
NOUN	Noun	*car*
NUM	Numeral	*two*
PART	Particle	*out*
PRON	Pronoun	*he*
PROPN	Proper noun	*Jenny*
PUNCT	Punctuation mark	, .
SCONJ	Subordinating conjunction	*that*
SYM	Symbol	☺
VERB	Verb	*in*
X	Other	*xbbe*

two types: universal tagsets, as defined by the Universal Dependency (UD) framework (Nivre et al. 2016) and listed in Table 2 as well as language-specific tagsets. The Stockholm-Umeå Corpus (SUC) tagset (Ejerhed et al. 1992, Gustafson-Capková & Hartmann 2006) is utilized as the language-specific tagset for Swedish as specified in Table 3, and the Penn Treebank Tagset (Mitchell et al. 1993) for English as illustrated in Table 4.

The PoS analysis with UD tags (see Table 2) is shown in column UPOS, and the language specific PoS tags (see Table 3 and 4) are shown in the column XPOS.

Morphological analysis reproduces a set of features that describe the word's lexical and grammatical characteristics. Lexical features include subcategories of PoS, such as nouns/proper nouns/types of proper noun. Grammatical features describe the categorization of PoS for a given word form; for example, gender, numeral expression, case and species for nominal word classes (noun and pronoun) or case, species, comparative degree and numeral expression for adjectives. We use the UD formalism for both English and Swedish, with an additional morphological feature set using SUC for Swedish.

Universal morphological features are stated in the form of Feature = Value pairs, where the feature indicates the morphological category (which may

Table 3: PoS tags in SUC 2.0 (XPOS).

Tag	Explanation	Example
AB	Adverb	*inte* 'not'
DT	Determiner	*ett* 'an'
HA	Interrogative/relative adverb	*när* 'when'
HD	Interrogative/relative determiner	*vilken* 'which'
HP	Interrogative/relative pronoun	*som* 'as'
HS	Interrogative/relative possessive pronoun	*vars* 'whose'
IE	Infinitive marker	*att* 'to'
IN	Interjection	*ja* 'yes'
JJ	Adjective	*fin* 'pretty'
KN	Conjunction	*och* 'and'
MAD	Major delimiter	. ? ! :
MID	Minor delimiter	, - ; / *
NN	Noun	*bil* 'car'
PAD	Pairwise delimiter	( )
PC	Participle	*dansande* 'dancing'
PL	Particle	*in* 'in'
PM	Proper noun	*Jenny*
PN	Pronoun	*han* 'he'
PP	Preposition	*på* 'on'
PS	Possessive pronoun	*hennes* 'her'
RG	Cardinal number	*två* 'two'
RO	Ordinal number	*andra* 'second'
SN	Subjunction	*att* 'to'
UO	Foreign word	*nota bene*
VB	Verb	*fira* 'celebrate'

be in abbreviated form) and the value describes the actual features of the word. Some of the most important morphological features in UD for English and Swedish are listed in Table 5.

Unlike UD, SUC does not explicitly state the name of the morphological feature but only provides its corresponding value. In contrast, UD spells out both the feature name and its value. SUC's morphological features are described in Table 6.

The interested reader is referred to UD's description (Nivre et al. 2016), the SUC 2.0 Manual (Gustafson-Capková & Hartmann 2006) and the Penn Treebank documentation (Mitchell et al. 1993) for further details. Annotation of PoS with morphological features is performed with the aid of EFSELAB (Östling 2016) and represented in the column UFEATS for the UD annotation and CFEATS for the SUC morphological annotation, as illustrated in Figure 10.

Table 4: Penn Treebank PoS tagset (`XPOS`.)

PoS Tag	Description
CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PRP	Personal pronoun
PP$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	infinitival to
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Wh-determiner
WP	Wh-pronoun
WP$	Possessive wh-pronoun
WRB	Wh-adverb

Table 5: Examples of morphological features in UD (UFEATS).

Feature	Value	Explanation	PoS
Case	Nom	Nominative	ADJ, NOUN, PRON, PROPN
	Acc	Accusativ	
	Gen	Genitive	
Definiteness	Ind	Indefinite	ADJ, DET, NOUN, PRON, PROPN
	Def	Definite	
Gender	Com	Utrum	ADJ, DET, NOUN, PRON, PROPN
	Neut	Neutrum	
Number	Sing	Singular	ADJ, DET, NOUN, PRON, PROPN
	Plur	Plural	
Possessive	Yes	Possessive	DET
Degree	Pos	Positive	ADJ, ADV
	Cmp	Comparative	
	Sup	Superlative	
Mood	Ind	Indicative	AUX, VERB
Tense	Pres	Present	AUX, VERB
	Past	Past	
Verb form	Fin	Finite	AUX, VERB
	Inf	Infinite	
Voice	Act	Active	AUX, VERB
	Pass	Passive	
Abbreviation	Yes	Abbreviation	ADV

4.2.4 *Lemmatization*

The lemma of each word is identified, reducing each token to its base form, known as its "lemma." Unlike stemming, which simply removes word endings, lemmatization takes into account the context and meaning of the word to return a valid dictionary form, as shown in the column labeled LEMMA in Figure 10. Lemmatization uses vocabulary and morphological analysis to accurately transform words based on their part of speech and is performed in conjunction with PoS tagging, utilizing EFSELAB (Östling 2016).

Table 6: Morphological features in SUC 2.0 (CFEATS).

Feature	Value	Explanation	PoS
Gender	UTR	Common/Utrum	DT, HD, HP, JJ, NN, PC, PN, PS, (RG, RO)
	NEU	Neuter	
	MAS	Masculine	
Number	SIN	Singular	DT, HD, HP, JJ, NN, PC, PN, PS, (RG, RO)
	PLU	Plural	
Definiteness	IND	Indefinite	DT, (HD, HP, HS), JJ, NN, PC, PN, (PS, RG, RO)
	DEF	Definite	
Case	NOM	Nominative	JJ, NN, PC, PM, (RG, RO)
	GEN	Genitive	
Tense	PRS	Present	VB
	PRT	Past	
	SUP	Supinum	
	INF	Infinite	
Voice	AKT	Active	VB
	SFO	Passive	
Mood	KON	Conjunctive	VB
Participle	PRS	Present	PC
	PRF	Perfect	
Degree	POS	Positive	(AB), JJ
	KOM	Comparative	
	SUV	Superlative	
Pronoun	SUB	Subject form	PN
	OBJ	Object form	
	SMS	Compound	All PoS

```
TEXTID  ID FORM        NORM     LEMMA   UPOS    XPOS   U-FEATS
1.1     1  Lixards     Lizards  Lizard  NOUN    NNS    Number=Plur
1.1     2  like        like     like    VERB    VBP    Mood=Ind|Tense=Pres|VerbForm=Fin
1.1     3  to          to       to      PART    TO     _
1.1     4  eat         eat      eat     VERB    VB     VerbForm=Inf
1.1     5  bugs        bugs     bug     NOUN    NNS    Number=Plur
1.1     6  .           _        .       PUNCT   .      _
```

Figure 10: Linguistic annotation labels for lemma (LEMMA), PoS (UPOS, XPOS) and morphological features (UFEATS).

4.2.5 *Syntactic analysis*

Finally, sentences are syntactically analyzed by a parsing module that identifies the relationships between words and their grammatical functions within the sentence. This analysis follows the Universal Dependency formalism, version 2 (UD2)[3]. In the UD framework (Nivre et al. 2016), words are linked in pairs, forming binary relationships where one word serves as the head and the other as its dependent. The verb typically occupies the central role in the sentence and acts as the head. Unlike other dependency formalisms, where the finite verb (including auxiliary verbs) is the head, in UD Version 2 (UD2), and as used by SWEGRAM, the verb carrying the semantic content is considered the head. All other tokens (including words and punctuation marks) are connected either directly or indirectly to the semantic verb through directed arcs. In UD2, these arcs point from the dependent words to their respective heads, showing the relationship between the head and its dependents through linked arcs, with words acting as nodes. The syntactic or grammatical functions are typically specified on these links between each head-dependent pair. The syntactic dependency relationships are listed in Table 7.

We illustrate the syntactic analysis for the sentence *The cold wind hit my cheek* both in CoNLL-U format (shown in Figure 11) with the syntactic analysis in the columns `HEAD` and `DEPREL` and as a dependency graph in Figure 12. Each word in the sentence forms a pair, with one word serving as the head and the other as its dependent, linked by a specific syntactic function. Dependents point toward their head words, with the syntactic relationship clearly stated. Every sentence is organized around an intended root node, labeled as 0, which serves to connect the words in the sentence. Each word is sequentially numbered based on its ID (in our example, 1–7). The head word in the sentence is the semantic verb *hit* (word number 4), which has several direct dependents: the subject of the clause *wind* (`NSUBJ`), the direct object *cheek* (`OBJ`), and the punctuation mark "." (`PUNCT`). In turn, these words also have their own dependents. For instance, *wind* (word number 3) has *The* (word 1) and *cold* (word 2) as its dependents, functioning as determiner (`DET`) and adjectival modifier (`AMOD`), respectively. Similarly, *my* (word number 6) has *cheek* (word number 7) as its head.

4.3 *Correction of automatic analysis*

Modularity has been a key factor in the development of the tool, allowing users to customize the annotation process by enabling or disabling specific

3 https://universaldependencies.org/v2/

Table 7: Syntactic relations in UD2 (DEPREL).

Name	Description	Name	Description
acl	clausal modifier of noun (adjectival clause)	fixed	fixed multiword expression
advcl	adverbial clause modifier	flat	flat multiword expression
advmod	adverbial modifier	goeswith	goes with
amod	adjective modifier	iobj	indirect object
appos	appositional modifier	list	list
aux	auxiliary verb	mark	infinitive marker
case	case marking	nmod	nominal modifier
cc	conjunction	nsubj	nominal subject
ccomp	clausal complement	nummod	numerical modifier
clf	classifier	obj	object
compound	compound	obl	oblique nominal
conj	conjunction	orphan	orphan
cop	copula	parataxis	parataxis
csubj	clausal subject	punct	punctuation
dep	unspecified	reparandum	reparation
det	determiner	root	root
discourse	discourse element	vocative	vocative
dislocated	dislocated element	xcomp	open clausal complement
expl	expletive		

ID	FORM	NORM	LEMMA	U-POS	X-POS	U-FEATS	HEAD	DEPREL
1	The	The	the	DET	DT	Definite=DEF|PronType=Art	3	det
2	cold	cold	cold	ADJ	JJ	Degree=Pos	3	amod
3	wind	wind	wind	NOUN	NN	Number=Sing	4	nsubj
4	hitt	hit	hit	VERB	VBD	Mood=Ind|Tense=Past|VerbForm=Fin	0	root
5	my	my	my	PRON	PRPS	Number=Sing|Person=1|Poss=Yes|PronType=Prs	6	nmod:poss
6	face	face	face	NOUN	NN	Number=sing	4	obj
7	.	.	.	PUNCT	.	_	4	punct

Figure 11: Annotation in CoNLL-U format.

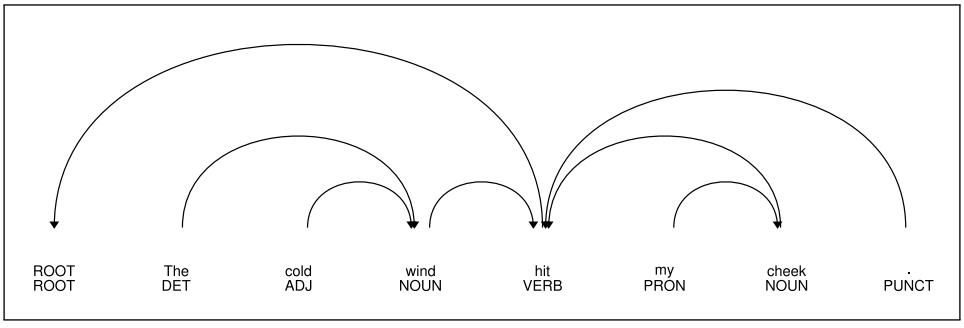

Figure 12: Example of syntactic analysis with dependency relations.

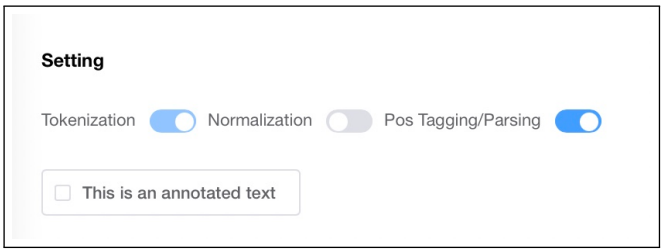

Figure 13: Choosing annotation components.

modules. This flexibility supports both tailored text analysis and manual correction of automatically assigned labels for subsequent processing.

The tool performs automatic linguistic analysis through several core components: tokenization, normalization, PoS tagging, and syntactic analysis (parsing), as shown in Figure 13. By default, all components except normalization are activated from the start.

To facilitate manual correction, users can download the automatically annotated file, review incorrectly assigned labels, and replace them with the correct ones, ensuring consistency with the tagset defined in the corresponding column. This step enhances the accuracy of annotations before further processing, making the tool adaptable to specific research needs.

Once corrections have been made, users can upload the revised annotations in the CoNLL-U format, as previously defined, for further automatic analysis by downstream components. This approach enables a refined workflow, where automatic linguistic analysis can be manually corrected and seamlessly reintegrated into the processing pipeline, ensuring both accuracy and flexibility.

4.4 *Storage of annotated texts*

The annotated texts are stored in a database on the server for a limited period. The system automatically deletes them after a few days. Since user registration is not required, the system does not track either the annotated texts or their statistics. Therefore, users should download and save their annotated texts and analyses for future processing.

4.5 *Linguistic analysis*

Once the uploaded texts are analyzed, SWEGRAM can perform quantitative analysis of various linguistic features based on the annotations. Statistics are calculated at multiple levels, including individual texts, several texts within

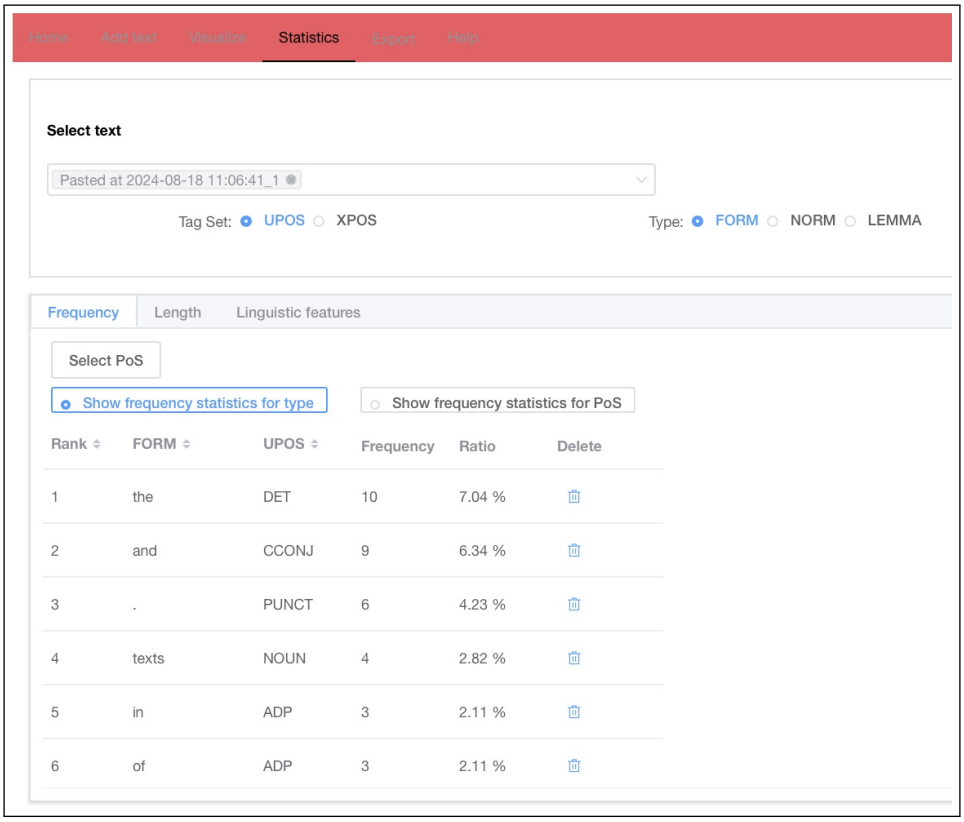

Figure 14: Frequency analysis.

a single file, or multiple uploaded files. General statistics can be obtained on word, sentence, and paragraph length, misspellings, PoS distribution, as well as readability metrics. Individual texts can also be studied using the visualization tool. Additionally, search filters can be applied to display statistics for selected texts among those that have been uploaded. Further details on statistics, filtering and visualization of various features are provided below.

4.5.1 *Frequencies*

SWEGRAM provides a display of tokens along with their PoS tags, which includes both the universal dependency tagset and a chosen language-specific tagset. The display also includes the total number of occurrences and the ratio (percentage) relative to the selected text(s). Figure 14 illustrates an example of the start of the frequency list for the abstract of this chapter, which will also serve as a demo text for the subsequent sections to illustrate the various features offered by SWEGRAM.

By default, the frequency analysis is based on the normalized word form (NORM). However, if this form is not available in the annotation, the analysis defaults to the surface form (FORM). Alternatively, the lemma can be selected in the top row of the frequency window. The frequency list can be sorted either alphabetically by token or by part-of-speech category.

Additionally, the frequency can be displayed separately for each PoS type, as shown in Figure 15. The search can be refined by deselecting unwanted PoS categories using the toggle buttons.

4.5.2 *Word length*

The statistics of word length excluding punctuation marks are shown in the *Length* feature set. SWEGRAM visualizes the number of tokens associated with specific PoS categories (displayed in columns) according to their character length, from 1 up to the length of the longest token in the text (displayed in rows). In the table, the rows represent word lengths, while the columns correspond to the parts of speech. The cells show the absolute frequency of tokens based on their PoS and length. The cells also show the specific words contributing to the counts and how often they appear in the selected texts. The list can be sorted by increasing or decreasing frequency.

Figure 16 presents the statistics for the length features based on the abstract of this chapter. In this specific example, there are eight nouns with a character length of four, six nouns with a length of five, two nouns with a length of six, and one noun with a length of three. The last row shows that there are 38 nouns in total. The last column lists the total number of words for each corresponding length.

4.5.3 *Syllables*

Some linguistic features and readability measures consider the number of syllables a word contains, and SWEGRAM estimates syllable counts using a rule-based heuristic approach for both English and Swedish. In English, the function `_syllable_count_en()` first converts the word to lowercase and determines the initial syllable count based on whether the first letter is a vowel. It then scans the word, increasing the count when a vowel is followed by a consonant, as this transition often marks the beginning of a new syllable. Additionally, it accounts for silent final *e*, subtracting one syllable if the word ends in *e*, and ensures that every word has at least one syllable by returning the higher value between the computed count and one. In Swedish, the function `_syllable_count_sv()` follows a similar approach but primarily counts the number of vowels in a word, as Swedish syllable structure typically

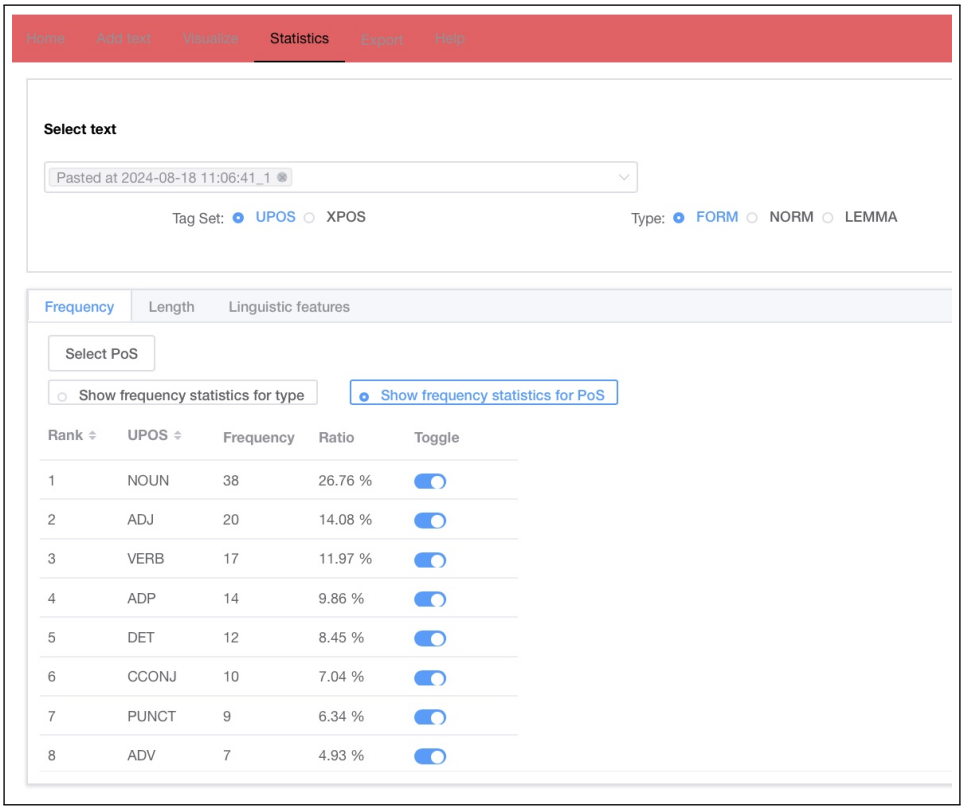

Figure 15: Frequency analysis of PoS categories.

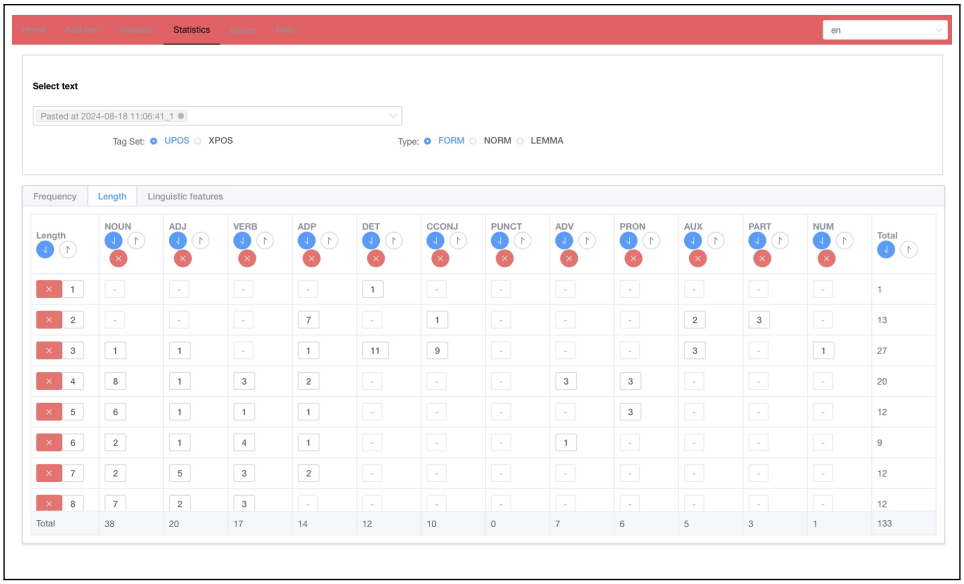

Figure 16: The analysis of character length sorted by PoS.

aligns with vowel occurrences. A notable exception is the word *journalist*, where the *ou* sequence is treated as a single syllable, reducing the overall count by one. If no vowels are found, the function guarantees a minimum syllable count of one. While these methods efficiently approximate syllable counts, they do not capture irregular pronunciations or complex phonetic patterns, making them useful but approximate tools for syllable extraction.

4.5.4 *Linguistic features*

In addition to frequency list of tokens and PoS tags, and the analysis of word length distribution of the uploaded text(s), SWEGRAM provides quantitative measurements on linguistic features generated from the annotated text, paragraph and sentence levels.

The content of the text under analysis—whether it is a full text, paragraph, or sentence—is displayed in the *Content* section, as illustrated in Figure 17 for the abstract of the article.

General features

Clicking on *Detail* in the top right corner displays the linguistic features in a popup window. The features for the specific content are computed immediately after the text is uploaded to the system. Once the level of analysis is selected, the statistics in *Overview* are displayed as a list consisting of *General*, *Readability*, *Lexical*, *Morphological*, and *Syntactic features*.

The general features of the text under analysis include details such as the total number of tokens, types, sentences, paragraphs, misspellings, separated compound words, as well as word, sentence, and paragraph lengths. Additionally, the median and mean values are provided, calculated based on the specific level at which the statistics are generated. Figure 18 illustrates the presentation of these general statistics for the abstract of this chapter.

Readability

SWEGRAM also calculates various readability, word variation, and nominality metrics. These are gathered under *Readability features*, which cover some common readability measures. Since readability measures are language dependent, two different sets have been developed, one for Swedish and another for English.

For Swedish, the Readability index (Lix) (Björnsson 1968), Word variation index (Ovix) (Hultman & Westman 1977), Type-token ratio (TTR) as discussed by (Johansson 2009), Simple nominal ratio (Af Geijerstam 2006: 108), and Full nominal ratio (Hultman & Westman 1977) are calculated.

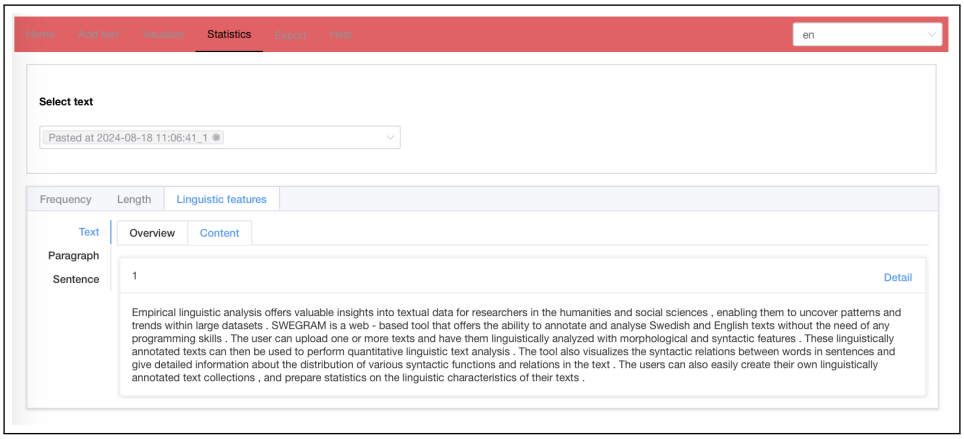

Figure 17: The analyzed text shown in *Content*.

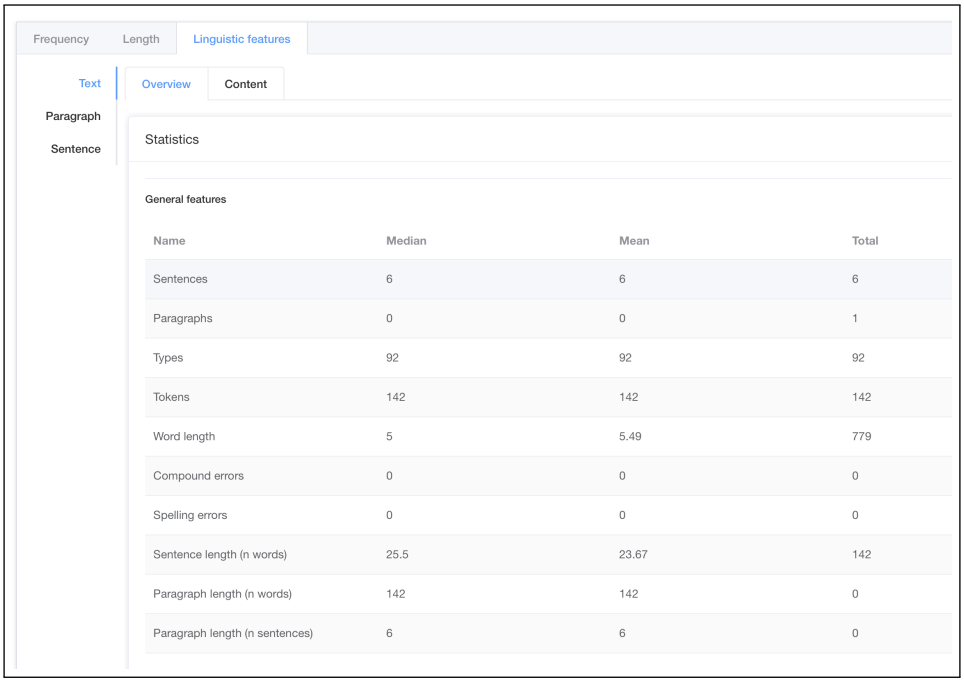

Figure 18: General features.

Table 8: Lix values, different genres.

Type of text	Lix	Words/ sentences	Long words	Interpretation
Books for children and young people	27	12	15	Very easy
Fiction	33	15	18	Easy
Daily and weekly press	39	14	25	Intermediate
Factual literature	47	18	29	Difficult
Nonfiction	56	20	35	Very difficult

Lix LIX, short for *läsbarhetsindex* in Swedish (readability index in English), is a formula used to evaluate the readability of a text, specifically how easy or difficult it is to comprehend (Björnsson 1968). It is calculated as shown in Equation 1, where a long word is defined as any word with more than six characters. The readability scores can be interpreted according to Table 8 (Melin & Lange 2000).

$$\text{LIX} = \frac{\text{total number of tokens}}{\text{total number of sentences}} + \frac{\text{total number of long tokens} \times 100}{\text{total number of tokens}} \quad (1)$$

Ovix Ovix stands for *ordvariationsindex* (English, word variation index) (Hultman & Westman 1977) and measures the total number of lexemes (minimal meaningful unit of language) in relation to the total number of tokens in a text. The formula used is defined in Equation 2. A higher Ovix value implies a greater word variation in relation to the length of the text (Hultman & Westman 1977: 56).

$$\text{OVIX} = \frac{\ln(\text{total number of tokens})}{\ln\left(2 - \frac{\ln(\text{total number of types})}{\ln(\text{total number of tokens})}\right)} \quad (2)$$

Type-token ratio The type-token ratio, abbreviated TTR, and discussed by Johansson (2009) is a metric used to measure the lexical diversity of a text. It is calculated by comparing the number of unique words (types) to the total number of tokens in a given text, as defined in Equation 3. A higher TTR indicates greater lexical diversity, meaning the text uses a wider variety of words while a lower TTR suggests that the text repeats the same words more often, indicating less lexical diversity. TTR can be affected by text length, as

shorter texts tend to have a higher TTR due to fewer opportunities for word repetition, while longer texts naturally tend to have a lower TTR.

$$\text{TTR} = \frac{\text{total number of types}}{\text{total number of tokens}} \tag{3}$$

Instead of using a simple ratio of the total number of types and tokens to mitigate the risk of the results affected by text length, bi-logarithm and square root TTR (see Equations 4) are also implemented. Here, logarithmic and square root transformations are applied to the number of types and the number of tokens, making the metrics more stable across texts of varying lengths.

$$f(\text{Bi-logarithm TTR}) = \frac{\ln(\text{total number of types})}{\ln(\text{total number of tokens})} \tag{4a}$$

$$f(\text{Square Root TTR}) = \frac{\text{total number of types}}{\sqrt{\text{total number of tokens}}} \tag{4b}$$

Simple nominal ratio The simple nominal ratio (Af Geijerstam 2006: 108) gives an indication of the text's information density by indicating the relationship between nouns (`NN`) and verbs (`VB`). The calculation of simple nominal ratio is defined in Equation 5. The higher percentage of nouns compared to verbs, the more information dense the text is.

$$\text{Simple nominal ratio} = \frac{\text{total number of nouns (NN)}}{\text{total number of verbs (VB)}} \tag{5}$$

Full nominal ratio The full nominal ratio offers an advanced measure of nominality. It highlights the relationship between nouns and noun-related parts of speech on one side, and verbs, adverbs, and pronouns on the other. Elements such as nouns (`NN`), prepositions (`PP`), and participles (`PC`) (both present and past) are associated with higher nominality and greater information density. In contrast, verbs (`VB`), adverbs (`AB`), and pronouns (`PN`) tend to dilute the informational content of a text. The formula for the full nominal ratio is presented in Equation 6.

$$\text{Full nominal ratio} = \frac{\text{total number of NN+PP+PC}}{\text{total number of VB+AB+PN}} \tag{6}$$

To measure readability for English texts, two types of type-token ratio, Bilogarithm and Root (Guiraud 1954), as well as the Coleman Liau Index (Coleman

& Liau 1975), Flesch Reading Ease (Flesch 1948), Flesch Kincaid Grade level (Kincaid et al. 1975), Automated Readability Index (Kincaid & Delionbach 1973), and Simple Measure of Gobbledygook (SMOG) index (McLaughlin 1969) are implemented.

Coleman Liau Index Coleman–Liau readability test (CLI) (Coleman & Liau 1975) is a formula used to evaluate the readability level of a text, specifically estimating the U.S. grade level required to understand the text. It is based on characters per tokens and sentence length indices and is calculated as shown in Equation 7. The formula first computes the average number of characters per 100 tokens to estimate word complexity. Then, it calculates the average number of sentences per 100 tokens to estimate sentence length and structure complexity.

$$\text{CLI} = 5.88 \times \frac{\text{total number of characters}}{\text{total number of tokens}} - 29.6 \times \frac{\text{total number of sentences}}{\text{total number of tokens}} - 15.8 \quad (7)$$

The resulting number is an estimate of the U.S. school grade level necessary to understand the text. The interpretation of the scores is shown in Table 9.

Flesch Reading Ease The Flesch Reading Ease (FRE) (Flesch 1948) analyzes the sentence lengths and word complexity by calculating the average length of the sentences measured by the number of words and the average number of syllables per word in the text.

$$\text{FRE} = 206.835 - 1.015 \times \frac{\text{total number of words}}{\text{total number of sentences}} - 84.6 \times \frac{\text{total number of syllables}}{\text{total number of words}} \quad (8)$$

It provides a score between 0 and 100; the higher the reading score, the easier the text is to read. The interpretation of the scores is described in Table 10.

Flesch Kincaid Grade level Flesch-Kincaid Grade Level (FKG) (Kincaid et al. 1975) is a variation of the Reading Ease formula with readjusted weights. The measure assesses the approximate reading grade level of a text, based on average sentence length and word complexity, see Equation 9.

Table 9: Coleman Liau Index levels.

Score	School level	Comprehension
5	5th grade and below	Very easy to read
6	6th grade	Easy to read
7	7th grade	Quite easy to read
7–10	8th, 9th, and 10th grade	Conversational English
11–12	11th and 12th grade	Quite hard to read
13–16	College	Difficult to read
17+	Professional	Very hard to read

Table 10: Flesch Reading Ease levels.

Score	Reading level
90–100	very easy to read, easily understood by an average 11-year-old
80–90	easy to read
70–80	fairly easy to read
60–70	easily understood by 13- to 15-year-old
50–60	fairly difficult to read
30–50	difficult to read, best understood by college graduates
0–30	very difficult to read, best understood by university graduates

Table 11: Flesch Kincaid scores and US grade levels.

Score	Reading level	School level	Age
0–3	Basic	Kindergarden	5–8
3–6	Basic	Elementary	8–11
6–9	Average	Middle school	11–14
9–12	Average	High school	14–17
12–15	Advanced	College	17–20
15–18	Advanced	Post-grad	20+

$$\text{FKG} = 0.39 \times \frac{\text{total number of words}}{\text{total number of sentences}} + 11.8 \times \frac{\text{total number of syllables}}{\text{total number of words}} - 15.59 \quad (9)$$

The scores correspond to US grade levels and age as listed in Table 11.

Automated Readability Index The Automated Readability Index (ARI) (Kincaid & Delionbach 1973) considers the average number of characters per

Table 12: ARI Scores and US grade levels.

ARI Score	Grade level	Reading level	Ages
1–5	Kindergarten	Extremely easy	5–6 y
1–5	1st grade	Extremely easy	6–7 y
6–7	2nd grade	Very easy	7–8 y
8–9	3rd grade	Very easy	8–9 y
10–11	4th grade	Easy	9–10 y
12–13	5th grade	Fairly easy	10–11 y
14–15	6th grade	Fairly aasy	11–12 y
16–17	7th grade	Average	12–13 y
18–19	8th grade	Average	13–14 y
20–21	9th grade	Slightly difficult	14–15 y
22–23	10th grade	Somewhat difficult	15–16 y
24–25	11th grade	Fairly difficult	16–17 y
26–27	12th grade	Difficult	17–18 y
28+	College	Very difficult	18–22 y

token and the average number of tokens per sentence. Characters include any letters, numbers, symbols, etc. with the exception for white space between characters.

$$\text{ARI} = 4.71 \times \frac{\text{total number of characters}}{\text{total number of tokens}} + 0.5 \times \frac{\text{total number of tokens}}{\text{total number of sentences}} - 21.43 \quad (10)$$

Table 12 summarizes the ARI scores and their corresponding interpretations.

SMOG The Simple Measure of Gobbledygook (McLaughlin 1969) is a measure of readability that estimates the years of education needed to understand a piece of writing. The SMOG formula scores a text based on the complexity of the sentences and words by considering the number of polysyllabic words, i.e. words with three or more syllables and the number of sentences in a text. The SMOG formula is defined In Equation 11.

$$\text{SMOG} = 1.043 \times \left(\sqrt{\frac{\text{total number of polysyllabic words} \times 30}{\text{total number of sentences}}} + 3.1291 \right) \quad (11)$$

The interpretation of the SMOG scores is detailed in Table 13.

Table 13: SMOG Scores and US grade levels.

SMOG Score	Approx. Grade level (+1.5 grades)
1–6	5
7–12	6
13–20	7
21–30	8
31–42	9
43–56	10
57–72	11
73–90	12
91–110	13
111–132	14
133–156	15
157–182	16
183–210	17
211–240	18

Table 14: Readability features for English texts. Scores are calculated on the chapter's abstract.

Name	Median	Mean	Total
SMOG	17.26	17.26	17.26
Root TTR	10.83	10.83	7.72
Bilogarithm TTR	0.98	0.98	0.91
Coleman Liau Index	17.30	17.30	17.30
Flesch Reading Ease	22.13	22.13	22.13
Flesch Kincaid Grade Level	15.68	15.68	15.68
Automated Readability Index	17.24	17.24	17.24

Each of the readability measures is calculated for the abstract of this chapter and shown in Table 14. All metrics indicate a professional text which is difficult to read with the exception of the ARI and SMOG scores, resulting in a grade level of 7.

Lexical features

To measure lexical diversity and proficiency, the Swedish KELLY-list[4] (Volodina & Kokkinakis 2012) is consulted. This list is based on the project *KEywords for Language Learning for Young and Adults Alike* (KELLY). The KELLY-list is

4 https://spraakbanken.gu.se/en/projects/kelly. Available on Sep 14, 2020.

Table 15: Example entries from the Swedish KELLY-list.

CEFR	Lemma	PoS	WPM
A1	*all* 'all'	pronoun	2975.47
	för 'for'	adverb	421.08
A2	*påstående* 'statement'	noun-ett	63.24
	för 'therefore'	conjunction	44.36
B1	*avgå* 'resign'	verb	23.46
	stabilitet 'stability'	noun-en	23.32
B2	*ödmjuk* 'humble'	adjective	11.86
	i natt 'tonight'	adverb	11.85
C1	*foster* 'fetus'	noun-ett	7.06
	enastående 'outstanding'	adjective	7.05
C2	*allergisk* 'allergic'	adjective	3.40
	eskalera 'escalate'	verb	3.40

a vocabulary sample that represents vocabulary usage, with each selected word or phrase classified into one of the six levels of the *Common European Framework of Reference for Languages* (CEFR) (Pilán & Volodina 2016) in ascending order of difficulty: A1, A2, B1, B2, C1, and C2. The Swedish KELLY-list contains 8,409 items generated from a corpus of web texts, as illustrated in Table 15 with examples of two entries for each CEFR level. Each word entry is represented by its lemma, part of speech, relative word frequency measured in words per million (WPM), followed by its English translation.

The distribution of lemmas according to their difficulty levels (A1, A2, B1, B2, C1, C2) found in the text is presented first, along with their median, mean, and total values. The number of "Difficult Words," referring to any token with a CEFR level of B1 or higher, and "Difficult NOUN & VERB," referring to nouns or verbs with a CEFR level of B1 or higher are provided. Finally, the number of tokens in the text that are not part of the KELLY list is listed. Table 16 illustrates the lexical features of the abstract of this chapter, indicating that the text contains many difficult words, both nouns and verbs.

Morphological features

The frequency of a part of speech or a morphological feature relative to other parts of speech or morphological features is measured using relative

Table 16: Lexical features. Scores are calculated on the chapter's abstract.

Name	Median	Mean	Total
A1 Lemma	415.49	415.49	415.49
A2 Lemma	133.80	133.80	133.80
B1 Lemma	98.59	98.59	98.59
B2 Lemma	14.08	14.08	14.08
C1 Lemma	14.08	14.08	14.08
C2 Lemma	49.30	49.30	49.30
Difficult Word	176.06	176.06	176.06
Out of Kelly list	274.65	274.65	274.65
Difficult Noun or Verb	126.76	126.76	126.76

frequencies, expressed through the incidence score (INCSC) (Graesser et al. 2004). The INCSC computation is defined as:

$$\text{INCSC} = 1000 \times \frac{N_t}{N_c} \tag{12}$$

where N_c is the number of tokens that belong to a certain category we are interested in and N_t is the number of tokens used for comparison. In most cases, N_c is a subset of N_t.

SWEGRAM covers 30 morphological features grouped into six subcategories based on part-of-speech (PoS) usage in the INCSC calculation. These linguistic features have been found to correlate with language development in both first and second language research (Pilán 2018). The features are defined using Universal Dependency (UD) tags.

SWEGRAM calculates the proportion of the following parts of speech relative to all tokens: adjectives (`ADJ`), adverbs (`ADV`), nouns (`NOUN`), particles (`PART`), verbs (`VERB`), punctuation (`PUNCT`), and subordinating conjunctions (`SCONJ`).

In addition, SWEGRAM analyzes the distribution of one part of speech in relation to others (PoS-PoS), namely: noun-to-verb (`NOUN - VERB`), pronoun-to-noun (`PRON - NOUN`), and pronoun-to-preposition (`PRON - ADP`).

SWEGRAM also computes the proportion of three specific subcategories of parts of speech (SUBPoS) relative to all tokens in the text: passive verb forms (s-verbs), third-person singular personal pronouns (`PRON 3SG`), and neuter nouns (`NOUN NEU`).

Additionally, SWEGRAM calculates the proportion of adjectives (`ADJ`), adverbs (`ADV`), nouns (`NOUN`), and verbs (`VERB`) relative to the sum of these four parts of speech (PoS-Multiple PoS).

Table 17: Specification of categories of N_c and N_t given the specific morphological features. *ALL* refers to all parts of speech categories.

Feature	N_c	N_t
Functional token	Functional PoS	ALL PoS
Lexical token	Lexical PoS	ALL PoS
Conjunction och subjunction	PoS = CCONJ+SCONJ	ALL PoS
Interrogative and relative	PoS = Int+Rel	ALL PoS
Lexical - functional token	Lexical PoS	Functional PoS
Nominal - Verbal	PoS = NOUN+ADP+PC	PoS = PRON+ADV+VERB

Different types of verb forms (N_c) are analyzed with respect to modality, tense, and aspect in relation to the total number of verbs.

Finally, SWEGRAM also calculates the ratio of groups of parts of speech relative to other word groups. For instance, it can compute the ratio between lexical and functional words. Lexical parts of speech in the UD framework include adjectives (ADJ), adverbs (ADV), interjections (INTJ), nouns (NOUN), proper nouns (PROPN), and verbs (VERB). Functional parts of speech include prepositions (ADP), auxiliaries (AUX), coordinating conjunctions (CCONJ), determiners (DET), numbers (NUM), particles (PART), pronouns (PRON), subordinating conjunctions (SCONJ), punctuation (PUNCT), symbols (SYM), and other (X). The distribution of these parts of speech groups, based on N_c and N_t, is detailed in Table 17.

Table 18 summarizes the morphological features used in the abstract of this chapter.

Syntactic features

The syntactic analysis, as provided by the UD2 annotation, is based on the number of dependency arcs in the sentence, along with the type and frequency of the dependency relations. Five syntactic features are extracted from statistics on the dependency arcs:

- Dependence length: The total number of dependency arcs from each token in the sentence to the ROOT.
- Longest dependency length: The longest dependency arc from any token to the ROOT.
- Dependency arcs longer than 5: The total number of tokens with dependency arcs that exceed a length of 5.

Table 18: Lexical features. Scores are calculated on the chapter's abstract.

Name	Median	Mean	Total
POS-ALL ADJ INCSC	140.85	140.85	140.85
POS-ALL ADV INCSC	49.30	49.30	49.30
POS-ALL NOUN INCSC	267.61	267.61	267.61
POS-ALL PART INCSC	21.13	21.13	21.13
POS-ALL VERB INCSC	119.72	119.72	119.72
POS-ALL PUNCT INCSC	63.38	63.38	63.38
POS-ALL SCONJ INCSC	0	0	0
POS-POS NOUN to VERB	2235.39	2235.39	2235.39
POS-POS PRON to NOUN	157.89	157.89	157.89
POS-POS PRON to PREP	6000.00	6000.00	6000.00
SUBPOS-ALL S-VERB INCSC	0	0	0
POS-MultiPOS ADJ Variation	243.90	243.90	243.90
POS-MultiPOS ADV Variation	85.37	85.37	85.37
VERBFORM Past VERB to VERB	45.45	45.45	45.45
MultiPOS-MultiPOS Rel INCSC	7.04	7.04	7.04
POS-MultiPOS NOUN Variation	463.41	463.41	463.41
POS-MultiPOS VERB Variation	207.32	207.32	207.32
VERBFORM Modal VERB to VERB	227.27	227.27	227.27
VERBFORM Supine VERB to VERB	0	0	0
VERBFORM Present VERB to VERB	227.27	227.27	227.27
MultiPOS-MultiPOS Lex to Token	577.46	577.46	577.46
MultiPOS-MultiPOS Nominal Ratio	2000.00	2000.00	2000.00
MultiPOS-MultiPOS Lex to Non-lex	1366.67	1366.67	1366.67
VERBFORM Past Participle to VERB	181.82	181.82	181.82
VERBFORM Present Participle to VERB	0	0	0
MultiPOS-MultiPOS CCONJ SCONJ INCSC	70.42	70.42	70.42
MultiPOS-MultiPOS Functional Token INCSC	422.54	422.54	422.54

- Ratio of right dependency arcs: The proportion of right-directed arcs relative to the total number of arcs in the sentence.
- Ratio of left dependency arcs: The proportion of left-directed arcs relative to the total number of arcs in the sentence.

Syntactic complexity is also measured through six syntactic functions, quantified using N_c incidence scores. The syntactic relations analyzed include variation in modifiers (both pre-modifiers and post-modifiers), subordinate clauses, relative clauses, and prepositional complements.

- A modifier refers to any dependent whose dependency relation to its head is one of the following: adjectival modifier (`amod`), nominal modi-

fier (nmod), appositional modifier (appos), numeric modifier (nummod), adverbial modifier (advmod), or discourse element (discourse). A modifier occurring before its head is classified as a pre-modifier, while a post-modifier appears after its head.

Modifier Variation refers to the total number of modifiers, measured in relation to all syntactic functions.

- A subordinate refers to a dependency relation (DEPREL) annotated with one of the following labels: clausal subject (csubj), clausal complement (ccomp), open clausal complement (xcomp), adverbial clause modifier (advcl), clausal modifier of noun (acl), or relative adnominal clauses (acl:relcl) between the current token and any of its direct dependents.

- A relative clause refers to any dependent whose relation to its head includes the feature PronType=Rel, indicating that the Pronominal Type is a relative pronoun (PRON), determiner (DET), numeral, (NUM) or adverb (ADV). The N_c for the feature Rel Clause INCSC represents the number of tokens in relative clauses.

- A prepositional complement refers to a dependent whose dependency relation with its head is annotated as case. The N_c for the feature PREP Comp INCSC corresponds to the number of prepositional complements and other dependents subordinated under the heads of prepositional complements.

Table 19 provides a summary of the syntactic features of the abstract of this chapter.

4.6 *Viewing and exporting data*

The results from the analyzed texts are displayed directly under the *Statistics* and *Visualize* modules, providing insights into various linguistic characteristics of the selected texts. The *Statistics* module presents numerical data, such as word frequency, sentence length, and readability scores, offering a quantitative overview of the corpus. In contrast, the *Visualize* module transforms these data points into graphical representations, such as bar charts or graphs, making patterns and relationships more accessible and interpretable.

All data, including the annotated texts, statistical summaries, and visualizations, can be downloaded as a text file (.txt) or CSV file (.csv) for further analysis.

Table 19: Syntactic features. Scores are calculated on the chapter's abstract.

Name	Median	Mean	Total
Dependency length	8.00	8.00	8.00
Subordinate INCSC	225.35	225.35	225.35
Modifier variation	295.77	295.77	295.77
Pre-modifier INCSC	218.31	218.31	218.31
Post-modifier INCSC	77.46	77.46	77.46
Relative clause INCSC	133.80	133.80	133.80
Longest dependency length	8.00	8.00	8.00
Dependency arcs longer than 5	14.00	1400	35.00
Ratio of left dependency arcs	36.62	36.62	36.62
Prepositional complement INCSC	98.59	98.59	98.59
Ratio of right dependency arcs	63.38	63.38	63.38

5 *SWEGRAM in use*

SWEGRAM is a powerful tool for annotating and analyzing texts in both English and Swedish. It provides clear, easily interpretable linguistic annotations alongside statistical analyses derived from these annotations. As demonstrated in this chapter, SWEGRAM has broad applications across various fields, including corpus linguistics, text analysis, language education, and computational linguistics.

Researchers utilize SWEGRAM for quantitative linguistic analysis, extracting statistics on text length, word frequency, readability metrics, and syntactic complexity. It also supports text comparison and corpus creation, allowing users to analyze multiple texts with metadata to study genre variation, writing development, or historical language change. In educational settings, SWEGRAM is valuable for assessing students' writing skills, tracking syntactic development, and examining linguistic patterns in national exam essays. Additionally, the tool facilitates qualitative text analysis, enabling detailed investigations of specific grammatical structures and stylistic choices.

One of SWEGRAM's key functionalities is its ability to create corpora by allowing users to upload multiple text files, which are then processed through the SWEGRAM annotation pipeline. Users can attach metadata–such as author details (e.g., identity, gender, age, residence) and text attributes (e.g., text ID, year, genre, place of publication)—to enhance filtering and statistical analysis. A prominent example of SWEGRAM's application is its use in creating the Uppsala Corpus of Student Writings (Megyesi et al.

2016). This corpus comprises approximately 2,000 texts (about one million words) written by students for the national test in Swedish, spanning grades 3 to high school. The corpus includes both native Swedish speakers (L1) and learners of Swedish as a second language (L2). Each text was enriched with metadata, including the year and semester of production, text type (e.g., narrative, descriptive, investigative, argumentative), grade level, gender, and Swedish proficiency (L1 or L2). The full SWEGRAM annotation pipeline—covering tokenization, sentence segmentation, lemmatization, morphological analysis, part-of-speech tagging, and syntactic annotation following the Universal Dependencies standard—was applied. The Uppsala Corpus of Student Writings has been instrumental in quantitative empirical studies of student writing, examining linguistic variation across different groups based on gender, geographic location, age, and grade level in synchronic and diachronic perspectives. Designed as a monitor corpus, it supports longitudinal analyses of writing development over time.

SWEGRAM has been widely used in research on language and writing development, particularly in studies focusing on syntactic complexity across different age groups, educational levels, and text types. In a study by Anne Palmér, two essays from a national test in Swedish were compared in terms of lexical features, part-of-speech distribution, and syntax to demonstrate the potential applications of SWEGRAM (Näsman et al. 2017). Other studies have analyzed student texts from national exams in larger scale.

A study by Palmér & Hussenius (2022) examines the writing development of sixth-grade students, focusing on discursive writing and the linguistic tools they use. Using SWEGRAM for automated quantitative analysis, supplemented by minor qualitative assessments, the study analyzes 240 argumentative and narrative texts from Swedish national exams. Findings show that argumentative texts are shorter and have a more limited vocabulary than narrative ones, with a strong reliance on verbal structures. However, some students have started incorporating nominalization and denser informational content. Adverbs play a key role in structuring arguments, expressing evaluations, logical connections, and writer stance, while participles contribute to expressive and varied language, particularly in high-scoring texts. The study concludes that mastering verbal lexicogrammar is essential for writing development, with nominalization as the next step. The applied text measures effectively capture both genre and developmental aspects, providing insights for writing instruction and assessment.

A study by Bendegard et al. (2023) investigated productivity, syntactic complexity, and variation in third-grade student texts. The findings showed that while narrative texts were significantly longer than explanatory ones, they did not necessarily exhibit greater syntactic complexity. Instead, syntac-

tic variation differed by genre, suggesting that different text types activate distinct linguistic structures, which has implications for writing instruction and development. SWEGRAM was instrumental in computing key linguistic metrics such as total word count and average words per macro-syntagm.

Several master's and doctoral theses have also employed SWEGRAM to analyze linguistic patterns and writing development in Swedish. For example, Josefsson (2017) examined how SWEGRAM assigns part-of-speech tags to Swedish and Swedish as a second language student texts, particularly focusing on the classification of the word *så* as an adverb. In another study, Jönsson-Ahola (2024) conducted a comparative investigation of high school students' writing development between the fall and spring semesters of Swedish II. The analysis, based on 124 student texts, used SWEGRAM to compute LIX, OVIX, and full nominal ratios to measure linguistic complexity and syntactic variation. The study found that all quantitative measures increased between the two semesters, reflecting writing progression. A qualitative analysis further revealed distinct patterns of individual development, leading to discussions on how to assess and promote writing development effectively.

Beyond educational and research settings, SWEGRAM has been applied in workplace language development initiatives. For instance, the ESF-project *Kompetensutveckling i äldreomsorgen* (Competence Development in Elderly Care) led by Arbetsmarknadsförvaltningen (The Labour Market Administration), Stockholms stad (Stockholm City), aims to develop a model for language and professional competence development among elderly care staff. In this project, SWEGRAM has been used to assess participants' linguistic development. Due to time constraints, only text data were collected for later analysis. However, SWEGRAM's ability to measure language development through lexical variation and grammatical complexity has been recognized as a valuable tool for future evaluations. SWEGRAM has been particularly useful in this project because it enables teachers to analyze lexical variation automatically, eliminating the need for manual categorization. By calculating the distribution of different parts of speech, the tool facilitates rapid assessments of vocabulary changes over time and differences between participant texts within the same genre (personal communication with Ulf Sparredal, 2025).

SWEGRAM is expected to remain a helpful tool in linguistic research, writing development studies, and educational assessment. Its ability to process both small and large-scale text collections, generate comprehensive linguistic annotations, and support both quantitative and qualitative analyses makes it invaluable for researchers, educators, and language professionals. Whether used for corpus creation, writing assessment, workplace language development, or a combination of these, SWEGRAM provides powerful analytical capabilities that enhance a wide range of linguistic investigations.

6 *Released versions of SWEGRAM*

SWEGRAM is freely available in two versions: a web-based service accessible at http://swegram.ling.su.se and downloadable versions for local installation hosted on GitHub at https://github.com/bmegyesi/swegram-v2.

The web-based version operates without requiring an account or login, and no personal user information is stored. Uploaded texts remain on the server for a maximum of one week before being automatically deleted. As a result, users must re-upload their files if they wish to continue working on them at a later time.

In contrast, the downloadable versions allow users to save files directly to their own computers and are powered by Docker containers. One version features a command-line interface, while the other is a web-based version that mirrors the functionality of the online version. Both options have been thoroughly tested in Linux and macOS environments.

To run SWEGRAM locally, users need to install three additional packages: (i) Pandoc,[5] a universal document converter, (ii) EFSELAB (Östling 2018)[6] for morpho-syntactic annotation, and (iii) udpipe (Nivre et al. 2016)[7] for linguistic analysis. We recommend setting up a virtual environment to ensure smooth processing. Detailed installation and setup instructions are provided on the GitHub page.

The current online production environment for SWEGRAM is containerized and comprises three main components: the frontend user interfaces are built using a VUE project with an Nginx proxy; the backend API, responsible for text annotation and data processing, is powered by the FastAPI framework; and the latest MySQL image is used to store data and facilitates interactions with the backend API. Each container operates independently, allowing for separate maintenance and updates.

SWEGRAM is licensed under the Creative Commons CC BY-SA license, enabling users to freely use, share, copy, distribute, modify, and adapt the tool in various forms and for any purpose, including commercial use. Proper attribution to SWEGRAM is required for any use, distribution, or modification. Additionally, we kindly request that users reference this chapter and our previous publication (Megyesi et al. 2019) where applicable.

5 https://pandoc.org
6 https://github.com/robertostling/efselab
7 https://ufal.mff.cuni.cz/udpipe/1/install

7 Conclusion

This chapter provides a comprehensive overview of SWEGRAM, a tool designed for the automatic annotation and analysis of Swedish and English texts. SWEGRAM allows researchers in the humanities and social sciences to perform detailed linguistic analyses without requiring programming skills. The tool enables the upload of texts, which are then annotated for morphological and syntactic features, providing quantitative insights such as word frequencies, sentence structures, and readability metrics. Users can also visualize syntactic relations and create their own annotated corpora for further study. This chapter discusses the components, functionalities, and case studies involving SWEGRAM, highlighting its role in facilitating large-scale empirical linguistic research. The tool is available as both a web-based service and a downloadable version, making it accessible and flexible for various research needs.

Acknowledgments

We would like to extend our heartfelt gratitude to everyone who has contributed to the development of this tool. First and foremost, we are deeply thankful to Anne Palmér from the Department of Scandinavian Languages at Uppsala University for her insightful ideas and contributions to the initial version of SWEGRAM, as well as to Jesper Näsman for implementing the first online version of the tool. Our sincere thanks also go to Shifei Chen for upgrading the back-end, significantly improving the tool's processing speed. Additionally, we are grateful to the Master's program students in language technology at Uppsala University, who, over the years, have helped evaluate, debug, and provide invaluable user feedback. Lastly, our big thanks go to the reviewers and editors of the book for their thorough review and valuable feedback.

This work is part of Swe-CLARIN,[8] the Swedish branch of the European CLARIN infrastructure, and has received long-term financial support from the Swedish Research Council.

References

Af Geijerstam, Åsa. 2006. *Att skriva i naturorienterande ämnen i skolan* [Writing in science subjects in school]. Department of Linguistics & Philology,

8 https://sweclarin.se

Uppsala University. (Doctoral dissertation). https://uu.diva-portal.org/smash/record.jsf?pid=diva2%3A169321&dswid=4726.

Anthony, Laurence & Paul Baker. 2015. ProtAnt: A tool for analysing the prototypicality of texts. *International Journal of Corpus Linguistics* 20(3). 273–292. https://doi.org/10.1075/ijcl.20.3.01ant.

Bendegard, Saga, Josefin Lindgren & Anne Palmér. 2023. Produktivitet och syntaktisk komplexitet och variation i 9-åringars berättande och beskrivande texter [Productivity, syntactic complexity, and variation in 9-year-olds' narrative and descriptive texts]. In *Språk i praktiken: I en föränderlig värld* [Language in practice: In a changing world] (ASLA:s skriftserie 30), 39–59. ASLA, Svenska föreningen för tillämpad språkvetenskap. https://doi.org/10.17045/sthlmuni.24321526.

Bird, Steven, Ewan Klein & Edward Loper. 2009. *Natural language processing with Python: Analyzing text with the Natural Language Toolkit*. O'Reilly.

Björnsson, Carl Hugo. 1968. *Läsbarhet* [Readability]. Liber.

BNC Consortium. 2007. *British national corpus, XML edition*. Oxford Text Archive. http://www.natcorp.ox.ac.uk/corpus/index.xml?ID=consortium.

Borin, Lars, Markus Forsberg, Martin Hammarstedt, Dan Rosén, Roland Schäfer & Anne Schumacher. 2016. Sparv: Språkbanken's corpus annotation pipeline infrastructure. In *SLTC 2016. The sixth Swedish language technology conference, Umeå university, 17-18 november, 2016*. http://www8.cs.umu.se/~johanna/sltc2016/abstracts/SLTC_2016_paper_31.pdf.

Borin, Lars, Markus Forsberg & Johan Roxendal. 2012. Korp – the corpus infrastructure of Språkbanken. In *Proceedings of the 8th international conference on language resources and evaluation* (*lrec 2012*), 474–478. http://www.lrec-conf.org/proceedings/lrec2012/pdf/248_Paper.pdf.

Burnard, Lou. 2006. *Xaira: software for language analysis*. https://doi.org/10.4230/DagSemProc.06491.17.

Cap, Fabienne, Yvonne Adesam, Lars Ahrenberg, Lars Borin, Gerlof Bouma, Markus Forsberg, Viggo Kann, Robert Östling, Aaron Smith, Mats Wirén, et al. 2016. Sword: Towards cutting-edge Swedish word processing. In *SLTC 2016. The sixth Swedish language technology conference, Umeå, Sweden, 17-18 November, 2016*.

Coleman, Meri & Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology* 60. 283–284. https://api.semanticscholar.org/CorpusID:144250124.

Ejerhed, Eva, Gunnel Källgren, Ola Wennstedt & Magnus Åström. 1992. *The linguistic annotation system of the Stockholm-Umeå corpus project*. University of Umeå, Department of General Linguistics.

Flesch, Rudolph. 1948. A new readability yardstick. *Journal of Applied Psychology* 32(3)(221–233). https://doi.org/10.1037/h0057532.

Graesser, Arthur C, Danielle S McNamara, Max M Louwerse & Zhiqiang Cai. 2004. Coh-metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers* 36(2). 193–202. DOI: 10.3758/BF03195564.

Guiraud, Pierre. 1954. *Les caractères statistiques du vocabulaire: Essai de méthodologie* [The statistical characteristics of vocabulary: An essay on methodology]. Presses universitaires de France. https://books.google.se/books?id=XSUeAAAAIAAJ.

Gustafson-Capková, Sofia & Britt Hartmann. 2006. Manual of the Stockholm Umeå corpus version 2.0. *Unpublished Work*.

Hammarstedt, Martin, Anne Schumacher, Lars Borin & Markus Forsberg. 2022. *Sparv 5 user manual*. Tech. rep. Göteborg: University of Gothenburg, Sweden. https://hdl.handle.net/2077/73604.

Hoffmann, Sebastian, Stefan Evert, Nicholas Smith, David Lee & Ylva Berglund Prytz. 2008. *Corpus linguistics with bncweb – A practical guide*. Frankfurt am Main: Peter Lang. https://www.peterlang.com/document/1104732.

Honnibal, Matthew & Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. https://sentometrics-research.com/publication/72/.

Hultman, Tor G & Margareta Westman. 1977. *Gymnasistsvenska* [Upper secondary school Swedish]. Institutionen för nordiska språk, Lund.

Hutto, C.J. & Eric Gilbert. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media* 8(1). 216–225. DOI: 10.1609/icwsm.v8i1.14550.

Johansson, Victoria. 2009. *Developmental aspects of text production in writing and speech*, vol. 48. Department of Linguistics, Phonetics, Centre for Languages & Literature, Lund University. https://lucris.lub.lu.se/ws/files/5221582/1487260.pdf.

Jönsson-Ahola, Holger. 2024. *Den oundvikliga skriftspråksutvecklingen: en komparativ undersökning av gymnasieelevers skriftspråksutveckling mellan hösttermin och vårtermin under kursen svenska II* [The inevitable development of written language: A comparative study of upper secondary students' writing development between the fall and spring semester in the course Swedish II]. Bachelor's thesis, Stockholm University. https://su.diva-portal.org/smash/record.jsf?pid=diva2%3A1873436&dswid=-6759.

Josefsson, Eva. 2017. *"Så", ska det taggas som adverb? En granskning av hur annoteringsverktyget swegram ordklasstaggar elevtexter i svenska och svenska som andraspråk* ["Så", should it be tagged as an adverb? An examination of how the annotation tool SWEGRAM assigns part-of-speech tags to student texts in Swedish and Swedish as a second language]. Bachelor's thesis, Uppsala University.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý & Vít Suchomel. 2014. The Sketch Engine: Ten years on. *Lexicography* 1. 7–36. DOI: 10.1007/s40607-014-0009-9.

Kincaid, Peter J. & Leroy J. Delionbach. 1973. Validation of the automated readability index: A follow-up. *Human Factors* 15(1). 17–20. https://doi.org/10.1177/001872087301500103.

Kincaid, Peter J., Robert P. Jr. Fishburne, Richard L. Rogers & Brad S. Chissom. 1975. *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*. https://stars.library.ucf.edu/cgi/viewcontent.cgi?article=1055&context=istlibrary.

Laurence, Anthony. 2011. *Antconc : a learner and classroom friendly, multi-platform corpus analysis toolkit*. https://www.laurenceanthony.net/software/antconc/.

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard & David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for computational linguistics (acl) system demonstrations*, 55–60. DOI: 10.3115/v1/P14-5010.

McCallum, Andrew K. 2002. *MALLET: A machine learning for language toolkit*. http://mallet.cs.umass.edu.

McLaughlin, Harry G. 1969. Smog grading: A new readability formula. *Journal of Reading*. 639–646. http://www.jstor.org/stable/40011226.

Megyesi, Beáta, Jesper Näsman & Anne Palmér. 2016. The Uppsala corpus of student writings: Corpus creation, annotation, and analysis. In *Proceedings of the tenth international conference on language resources and evaluation (LREC'16)*, 3192–3199. Portorož, Slovenia: European Language Resources Association (ELRA). https://aclanthology.org/W17-0216/.

Megyesi, Beáta, Anne Palmér & Jesper Näsman. 2019. *Swegram: Annotering och analys av svenska texter* [SWEGRAM: Annotation and analysis of Swedish texts]. Department of Linguistics & Philology, Uppsala University, Sweden. https://uu.diva-portal.org/smash/record.jsf?pid=diva2%3A1326439&dswid=9822.

Megyesi, Beáta & Rex Ruan. 2024. *Swegram 2.0: Guidelines to annotation and analysis of English and Swedish texts*. Department of Linguistics, Stockholm

University, Sweden. https://www.diva-portal.org/smash/get/diva2:1908711/FULLTEXT01.pdf.

Melin, Lars & Sven Lange. 2000. *Att analysera text: Stilanalys med exempel* [Analyzing text: Stylistic analysis with examples]. Studentlitteratur.

Mitchell, Marcus P., Beatrice Santorini & Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19(2). 313–330. https://aclanthology.org/J93-2004.

Näsman, Jesper, Beáta Megyesi & Anne Palmér. 2017. SWEGRAM: A web-based tool for automatic annotation and analysis of Swedish texts. In *Proceedings of the 21st nordic conference on computational linguistics*, 132–141. Gothenburg, Sweden: Association for Computational Linguistics. https://aclanthology.org/W17-0216/.

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty & Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the tenth international conference on language resources and evaluation* (*LREC'16*), 1659–1666. Portorož, Slovenia: European Language Resources Association (ELRA). https://aclanthology.org/L16-1262/.

Östling, Robert. 2016. Shallow learning for sequence tagging. In *6th Swedish language technology conference* (*SLTC16*), *Umeå*, *Sweden*. https://people.cs.umu.se/johanna/sltc2016/abstracts/SLTC_2016_paper_10.pdf.

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

Palmér, Anne & Siri Hussenius. 2022. Vad säger de nationella proven i årskurs 6 om elevers pågående skrivutveckling? [What do the national tests in year 6 reveal about students' ongoing writing development?] In *Fjortonde nationella konferensen i svenska med didaktisk inriktning – didaktiska perspektiv på språk och litteratur i en globaliserad värld.* (*smdi 14*) [The fourteenth national conference on Swedish with a didactic focus – didactic perspectives on language and literature in a globalized World.] Malmö universitet.

Pettersson, Eva, Beáta Megyesi & Joakim Nivre. 2013. Normalisation of historical text using context-sensitive weighted Levenshtein distance and compound splitting. In *Proceedings of the 19th Nordic conference of computational linguistics* (*Nodalida 2013*), 163–179. https://aclanthology.org/W13-5617/.

Pilán, Ildikó. 2018. *Automatic proficiency level prediction for intelligent computer-assisted language learning*. University of Gothenburg, Sweden. http://hdl.handle.net/2077/55895.

Pilán, Ildikó & Elena Volodina. 2016. Classification of language proficiency levels in Swedish learners' texts. In *Proceedings of Swedish language technology conference*. https://people.cs.umu.se/johanna/sltc2016/abstracts/SLTC_2016_paper_7.pdf.

Schmidt, Thomas & Kai Wörner. 2014. Exmaralda. In Jacques Durand, Ulrike Gut & Gjert Kristoffersen (eds.), *Handbook on corpus phonology*, 402–419. Oxford University Press. http://ukcatalogue.oup.com/product/9780199571932.do.

Scott, Mike. 2016. *WordSmith Tools Version 7*. Stroud: Lexical Analysis Software. https://lexically.net/wordsmith/version7/.

Volodina, Elena & Sofie Johansson Kokkinakis. 2012. Introducing the Swedish Kelly-list, a new lexical e-resource for Swedish. In *Proceedings of the eight international conference on language resources and evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2012/pdf/264_Paper.pdf.

Wittenburg, Peter, Hennie Brugman, Albert Russel, Alex Klassmann & Han Sloetjes. 2006. ELAN: A professional framework for multimodality research. In Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odijk & Daniel Tapias (eds.), *Proceedings of the fifth International Conference on Language Resources and Evaluation (LREC'06)*. Genoa, Italy: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2006/pdf/153_pdf.pdf.

List of abbreviations

ARI	Automated Readability Index
BNC	British National Corpus
CEFR	Common European Framework of Reference for Language
CFEATS	Language-specific Features
CLI	Coleman Liau Index
CSV	Comma-Separated Values
DEPREL	Dependency Relation
HEAD	Syntactic Head
INCSC	Incidence Score
KELLY	KEywords for Language Learning for Young and Adults Alike
LIX	Läsbarhetsindex (readability index)
NLP	Natural Language Processing

NLTK	Natural Language Toolkit
NORM	Normalized Token
Ovix	Ordvariationsindex (word variation index)
PoS	Part-of-Speech
SMOG	the Simple Measure of Gobbledygook
SUC	Stockholm Umeå Corpus
TTR	Type-Token Ratio
UD	Universal Dependency
UFEATS	Universal Features
UPOS	Universal Part-of-Speech
XML	Extensible Markup Language
XPOS	Language-specific Part-of-Speech

Corresponding author

Beáta Megyesi
Department of Linguistics
Stockholm University
beata.megyesi@ling.su.se