

Towards a Swedish diachronic corpus

Intended content, structure and format of version 1.0

Eva Pettersson

Uppsala University • eva.pettersson@lingfil.uu.se

Lars Borin

Språkbanken Text, University of Gothenburg • lars.borin@svenska.gu.se

CONTENTS

1	Introduction	1
2	Content and structure	2
2.1	Principles for inclusion in the first version	3
2.2	Text selection	4
2.2.1	Genres represented in all time periods	4
2.2.2	Opportunistic data collection from 1800 onwards	5
2.2.3	COHA-like subcorpus	5
2.2.4	Social media subcorpus	7
3	Annotation	8
4	Format	11
5	Metadata	14
6	User Interface	17
7	Summary and conclusions	18
	References	19

1

INTRODUCTION

THE CLARIN research infrastructure aims to make digital language resources available to researchers from all disciplines, with a special focus on the humanities and social sciences. As part of the activities in the Swedish CLARIN node, *Swe-Clarín*,¹ we aim to develop a freely accessible Swedish diachronic corpus. We strongly believe that the existence of such a resource would be very valuable to facilitate large-scale research on Swedish language change, and to enable studies comparing the historical development of Swedish with that of other languages for which diachronic corpora exist.

In previous work, we investigated the structure and contents of available diachronic and historical corpora for a range of languages (Pettersson and Borin 2019a), as well as the textual resources available for the Swedish language, for different time periods and for different genres (Pettersson and Borin 2019b). Based on the findings from these studies, the current report gives an overview of the intended structure and contents of the upcoming first version of the Swedish diachronic corpus. In Section 2, we start by discussing the principles for inclusion of texts in the first version of the corpus, regarding genres, time periods and limitations in text quantities. The levels of linguistic annotation are introduced in Section 3, whereas Section 4 presents the corpus formats. Metadata information to be included is proposed in Section 5, and the features of the user interface are discussed in Section 6. Finally, a summary with some conclusions is presented in Section 7.

This report reflects work in progress, and most likely the specifications will change as the actual work on compiling the corpus gets underway. New versions of this report will be released as the need arises.

¹<https://sweclarin.se/>

2

CONTENT AND STRUCTURE

CONTENTWISE, THE LONG-TERM goal of the Swedish diachronic corpus is to provide a collection of texts with as much material as possible reflecting the full extent of the history of the Swedish language, ideally allowing investigation of both internally motivated and contact-induced language change. For the very first version of the corpus, though, we need to delimit its contents in a reasonable way. Thus, the following general constraints are introduced for the first version:

1. Only include texts that are already available in a digital format. Thus, we do not apply OCR, HTR or manual transcription of manuscripts for the first version of the corpus.
2. Only include texts written (primarily) in Swedish. Thus, we do not include texts written within the Swedish borders, but in another language, in the first version of the corpus.²

Concerning textual content, we will follow four main principles for inclusion in the first version of the corpus:

1. Include texts from genres represented in all historical stages of Swedish.³
2. Follow an opportunistic approach for texts from 1800 onwards.
3. Create a subpart of the corpus following the same design principles as

²Due to code-switching, some text passages may still contain text written in another language, e.g. Latin or (Low) German.

³Here and later, when we write “historical stages of Swedish”, we of course refer only to data in the form of digital text, which excludes spoken language in the case of all but the most recent historical period, and in practice we will restrict ourselves to dealing with written language, which only very rarely will involve transcribed speech.

the COHA corpus (Davies 2012), in order to facilitate comparative studies.

4. Create a subpart of the corpus containing social media text, enabling research on the development over time for this particular sublanguage.

In Section 2.1, these four principles are discussed in more detail. In Section 2.2, the actual selection of texts, based on these principles, is presented and discussed.

2.1 PRINCIPLES FOR INCLUSION IN THE FIRST VERSION

As reported in Pettersson and Borin (2019b), we have distinguished five genres for which there is text available for all the traditionally defined time periods of the Swedish language development, from Old Swedish (1225–1526) over Early Modern Swedish (1526–1732) and Late Modern Swedish (1732–1900) to Contemporary Swedish (1900 onwards):

1. religious text
2. secular prose
3. court records
4. laws and regulations
5. academic and scientific text

A good starting point for the first version of the Swedish diachronic corpus could thus be to include the texts available for these five genres. This way the corpus will, already from the beginning, contain material for all historical stages of Swedish, except for Runic Swedish (800–1225). We have decided not to include Runic Swedish in the very first version of the diachronic corpus, mainly for four reasons. Firstly, Runic writings differ from other Swedish texts in that it uses a different alphabet, which would require the implementation of separate search and browsing facilities for what even at the outset would amount to a relatively small proportion of the total data. Secondly, since the majority of the extant Runic texts were carved into stone, they are typically very short, mainly describing that someone has died. Thirdly, the Runic inscriptions arguably reflect a Common Nordic rather than a specific Swedish (Danish, Norwegian) language variety, which makes their inclusion – at least in the first version of the diachronic corpus – less of a priority. Lastly, nearly all Runic writings are already collected and accessible in a transliterated

4 Towards a Swedish diachronic corpus

and standardized form in *Samnordisk runtextdatabas* (Scandinavian Runic-text Database, SRD).⁴

In addition to including texts from the five genres that are available for all time periods, we intend to follow a more opportunistic approach for texts from 1800 onwards, and include other genres as well, due to the increased research interest and amounts of text available for this time period. For the same period, we also plan to create a subpart of the corpus following a design similar to the one in the *Corpus of Historical American English* (COHA) (Davies 2012), to facilitate comparative studies both within Swedish, and for Swedish as compared to English. As reported in Pettersson and Borin (2019a), the COHA corpus is divided into decades, where each decade contains text from four genres: fiction, popular magazines, newspapers, and non-fiction books. Furthermore, for each decade, there is a division into roughly half fiction and half non-fiction texts.

For Late Modern Swedish, there is a substantial amount of newspaper text available, as well as periodicals, secular prose and scientific text, that fit this structure well. These genres are also well represented in Contemporary Swedish, even though newspaper text and prose from the 20th century onwards is often restricted by copyright, meaning that the sentences may only be presented in a random order. For both Late and Contemporary Swedish, governmental texts are also well represented, and could be added to the non-fiction part of the corpus.

Another genre that has gained a lot of research interest in recent years, also for the study of language change (e.g. Eisenstein et al. 2014), is social media text. We therefore also plan for a subpart of the Swedish diachronic corpus containing social media text from 1998 onwards. This subcorpus will enable research on the development over time of the language used in (Swedish) social media text.

2.2 TEXT SELECTION

2.2.1 Genres represented in all time periods

Table 1 presents the amounts of text and the time periods covered for the text selection to be included in the first version of the Swedish diachronic corpus, based on the five genres that are represented in all time periods. The numbers in the table refer to the texts listed in Pettersson and Borin (2019b), excluding texts from *Svenska fornskriftsällskapet* (since these are only available in

⁴<https://www.nordiska.uu.se/forskn/samnord.htm/>

printed books) and from *Project Runeberg* (due to a lack of download possibilities). It could also be noted that the genres are sometimes quite broadly defined. This could be exemplified by the genre ‘Academic and Scientific Text’, where the texts from the Old Swedish time period mainly contain home remedies and astrological treatises that would not be considered science in present-day society. Furthermore, the academic texts from the Early Modern period are protocols from the Academic Consistory of Uppsala University, which are not scientific texts in the strict sense.

As could be expected, the numbers show that we generally have access to more data for younger time periods. For example, there are approximately 2 million words of law text for the Late Modern time period and 8 million words for Contemporary Swedish, but only about 500,000 words for Old Swedish and 100,000 words for Early Modern Swedish. An exception is the genre containing religious texts, where there is more data available for the older time periods. Interestingly, for several genres, the Late Modern period offers a considerably lower volume of digitized text than both older and younger time periods. This is true for religious text as well as for court records and scientific text.

2.2.2 Opportunistic data collection from 1800 onwards

In addition to including text from the five genres represented in all time periods (religious text, secular prose, court records, laws and regulations, and academic and scientific text), we aim for an opportunistic approach from 1800 onwards. Table 2 lists the amounts of text and the years covered for the additional genres that are represented from 1800 onwards. The numbers in the table refer to the texts listed in [Pettersson and Borin \(2019b\)](#), excluding texts that are not (yet) available in a digital format, in accordance with the principle to only include already digitised texts in the first version of the corpus (see further Section 2). Due to copyright issues, some of the texts from the 20th century onwards, such as newspaper text and blog text, may only be represented with a random sentence order.

As seen from the table, the largest amounts of text constitute either formal text, such as periodicals, newspaper text, and governmental text, or user-generated text, such as Wikipedia text, blogs and chats. The smallest amounts of text are from informal genres typically written by hand, such as personal stories, diaries and letters.

2.2.3 COHA-like subcorpus

Following the structure of the COHA corpus for a subpart of the Swedish diachronic corpus, we aim for a division into decades, from the 1810s to the

6 Towards a Swedish diachronic corpus

Time Period	Text Type(s)	Years Covered	#words
Religious Text			
Old Swedish	Bible text, legends, revelations	1276–1550	1,632,646
Early Modern	Bible text, prayers	1526–1732	2,005,594
Late Modern	Bible text, religious prose	1758–1891	820,854
Contemporary	Bible text, religious prose, hymns	1917–1937	1,060,692
Secular Prose			
Old Swedish	Chronicles, poems, legends	1303–1457	237,686
Early Modern	Chronicles, books, plays	1529–1727	634,216
Late Modern	Books, plays, poems	1700s–1942	51,321,347
Contemporary	Books, plays	1900–1999	38,378,003
Court Records			
Old Swedish	Tänkeböcker, inheritance dispute	1381–1560	768,105
Early Modern	Tänkeböcker, court records	1540–1719	3,049,034
Late Modern	Court records	1707–1862	878,861
Contemporary	Verdicts	1981–2009	32,206,334
Laws and Regulations			
Old Swedish	Laws, regulations, by-laws	1203–1460	520,024
Early Modern	Church-related laws, regulations	1527–1686	114,004
Late Modern	National laws	1686–1809	2,147,684
Contemporary	Swedish Code of Statutes	1880–2012	8,058,400
Academic and Scientific Text			
Old Swedish	Medicine, home remedies, astrology	1350–1525	147,009
Early Modern	Protocols from Uppsala University	1624–1699	4,981,618
Late Modern	Royal Swedish Academy of Sciences	1740–1778	27,878
Contemporary	Yearbooks, academic texts	1931–2016	30,298,020

Table 1: Data collection for genres represented in all time periods.

2000s, with roughly 50% fiction and 50% non-fiction for each decade. This could be achieved using a subset of the opportunistically collected data described in the previous section. The exact number of words for each decade will be decided during the work of actually building the corpus, but in the case of Swedish there are a number of fiction novels and plays available for this time span, mainly through *Litteraturbanken*, *Dramawebben*, *Språkbanken Text* and *Project Gutenberg* (see further [Pettersson and Borin 2019b](#): §§ 3.4.2 and 3.5.2). For the non-fiction part, COHA includes popular magazines, newspapers and non-fiction books. For Swedish, there are periodicals (corresponding to “popular magazines” in COHA) available from 1810 onwards (see further [Pettersson and Borin 2019b](#): §§ 3.4.4 and 3.5.5). We also have access to newspaper text, even though the order of the sentences has been shuffled for the contemporary texts, in order to avoid copyright infringement (see further [Pettersson and Borin 2019b](#): §§ 3.4.5 and 3.5.6). It will however be hard to follow

Text Type	Years Covered	#words
Personal stories	1962	13,177
Diaries	1814–1829	590,000
Essays, school tests	2009–2013	593,540
Accounts, registers	1913–1960	969,019
Song texts	mainly 1890s	1,032,700
Letters	1858–1912	1,507,958
Periodicals	1810–2010	123,465,333
Wikipedia	2017	370,211,509
Newspapers	1770s–2017	1,668,825,105
Governmental texts	1867–2016	1,884,833,772
Blogs, chats	1998–2017	7,993,516,656

Table 2: Opportunistic data collection from 1800 onwards, excluding the genres already listed in Table 1.

the COHA structure entirely, since there is a lack of non-fiction books spanning the whole period. The period from 1931 onwards could be covered by academic and scientific texts, as described in [Pettersson and Borin \(2019b: § 3.5.12\)](#). There is however currently a lack of similar texts available for the time period before 1931. One suggestion would therefore be that the non-fiction part of the COHA-like subcorpus would initially contain periodicals and newspaper text only.

2.2.4 Social media subcorpus

To enable research on the development of the language used in social media over time, we intend to include a subpart of the Swedish diachronic corpus containing user-generated text. For this purpose, we have access to text from blogs and online discussion forums, comprising in total 7,993,516,656 words, produced during the time period 1998–2017, as presented in Table 2 in Section 2.2.2. If possible, it would be interesting to also include Twitter data. The terms of use for Twitter data are however unclear and need to be further investigated.

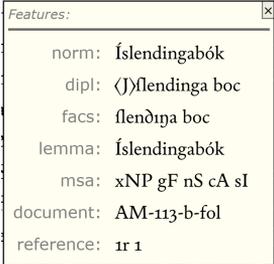
3

ANNOTATION

To enable advanced search queries and refined data analysis, we are aiming for a corpus with rich linguistic annotation, ideally including the following linguistics features:

- **spelling standardization**

Enables search for different spelling variants of the same word form, e.g., to search for *skriva* ‘to write’ to find spelling variants such as *skriva*, *skrifva*, *skriffa* etc. Note that there may be several levels of standardization, at least in the case of the oldest texts. For instance, in the manually digitized Old Nordic texts available through Menota (<https://menota.org/>) there are up to three renderings of manuscript text words: a facsimile rendering, a diplomatic rendering where abbreviations have been expanded but where some of the differences against the facsimile version are indicated by typography (italics, brackets, etc.), and a standardized (“normalised”) form (see Figure 1). Still other variants can be imagined, e.g. a ‘standardized diplomatic’ form, without the diplomatic typographical indications and with case differences eliminated.



Features:	
norm:	Íslendingabók
dipl:	⟨J⟩lendinga boc
facs:	flendriþa boc
lemma:	Íslendingabók
msa:	xNP gF nS cA sI
document:	AM-113-b-fol
reference:	1r 1

Figure 1: The first word(s) of *Íslendingabók* in the three transcription levels – norm(alized), dipl(omatic), facs(imile) – used by Menota.

- **lemmatisation**

Enables search for all inflectional forms of the same lemma, e.g., to search for *skriva* ‘to write’ to find all instances of *skriva* ‘to write’, *skriver* ‘writes’, *skrev* ‘wrote’ etc. Note that lemmas will have a standardized spelling as a matter of course, although this may not always coincide with the modern spelling, if the lemma refers to a standard dictionary, e.g., the Old Swedish dictionaries by [Söderwall \(1884\)](#) and [Schlyter \(1877\)](#), or the 19th-century dictionary by [Dalin \(1850–1853\)](#). The question whether the corpus itself should also contain the modern form or not in such cases is complex and cannot be resolved at this point.

- **part-of-speech tagging**

Enables search for sequences of words belonging to certain parts-of-speech, such as a pronoun followed by a verb. Also enables search for all instances of a certain word form analysed as belonging to a particular part-of-speech, e.g., to search for *kör* as a verb (‘drives’), excluding instances of *kör* analysed as a noun (‘choir’).

- **morphological analysis**

Enables search for inflectional features of words, such as nominal case forms and verbal tense forms, for parts of compounds, etc.

- **syntactic annotation**

Enables search for grammatical constructions, such as the position of adverbials in relation to other constituents in the sentence.

- **named entities**

Enables search for names of persons, organisations and geographical locations.

- **alignment**

If there are more than one version of a text from different time periods (or in different languages), corresponding passages can be explicitly *aligned*, which enables comparison of the ‘same’ linguistic items across time (or between languages). The prototypical example is furnished by bible translations, where there exists a standardized system for aligning versions down to verse level, and where word-level alignment is often undertaken in comparative and typological studies (e.g. [Buch et al. 2013](#); [Östling 2015, 2016](#)).

It remains to be decided what methods and tools to use for the linguistic markup, as well as the level of specificity in the morphological and syntactic analysis. One could think of a fully manual approach to one or more of the annotation layers, which would most likely render a reliable, high-quality lin-

guistic analysis. Considering the amounts of text in the corpus, this approach would however not be feasible, at least not for the corpus as a whole.

Another approach would be to use automatic tools for the different steps, such as the linguistic annotation tools integrated in Sparv (Borin et al. 2016) and/or the UDPipe pipeline for tokenisation, lemmatisation, part-of-speech tagging and dependency parsing (Straka and Straková 2017). Since these tools are trained on present-day Swedish data, this approach would probably achieve fairly reliable results for Contemporary Swedish texts. By using spelling standardization tools, we could also achieve comparable results for Early Modern and Late Modern Swedish using the same tools (see further Pettersson 2016). In the spelling standardization step, the original spelling is still kept as the main word form, but an additional annotation level is added, where the historical spelling is automatically translated to a more standardised spelling. This standardised spelling could then be used as input to the succeeding analysis tools, such as tagging and parsing, that are often sensitive to spelling. A drawback of using this method is however that in addition to handling spelling variation in the strict sense, the spelling standardization tools might also transform for example extinct or variable morphological suffixes, that could be of interest to the end user to search for in the morphological analysis.

Concerning Old Swedish texts, these are too dissimilar from Contemporary Swedish texts for tools trained on present-day Swedish data to be useful, even with spelling standardization as a preprocessing step. To add linguistic analysis to these texts, we either need to perform manual annotation of the data, or develop linguistic analysis tools trained on Old Swedish data, or the combination of both. It should also be noted that in the user questionnaire sent out to a number of potential users of the Swedish diachronic corpus, some of the researchers point out that high-quality manual annotation for part of the corpus is of importance (see further Pettersson and Borin 2019b: § 4).

In the very first version of the corpus, we will only include linguistic annotation if this information is already present in the data to be included. Adding linguistic annotation to other texts will be done for future versions of the corpus. It is however important to have the intended annotation in mind when planning for the corpus format, as is further discussed in the next section.

4

FORMAT

Regarding the format(s) of the corpus, two aspects need to be considered. First of all, we want a format for downloading and visualization that is easily readable and understandable to humans, preferably a format that is previously known and recognisable to the user. Secondly, we also need a format that is suited for automatic processing by computers, to enable search facilities and automatic analysis of the text.

With these criteria in mind, we intend to adapt a token-based, tab-separated format similar to the CoNLL-X format (Buchholz and Marsi 2006). The CoNLL format is a plaintext format, with one token – a word form or punctuation mark – on each line, followed by different layers of linguistic annotation separated by tab characters, and blank lines denoting sentence⁵ boundaries. We believe this format to be a well-known and often-used format that is easy to read and process by both humans and machines. Another advantage of this column-based format is that it could be imported in Microsoft Excel, a program reportedly used by researchers in the humanities for different analysis purposes. The particular flavor most likely to be adopted for the first version of the corpus is CoNNL-U Plus,⁶ which allows for the inclusion of arbitrary additional annotation categories in the form of added fields formally declared in the file header.

In accordance with the linguistic analysis steps proposed in the previous sec-

⁵The units referred to here as “sentences” may not always correspond to the modern notion of sentence. Especially in the Old Swedish texts, it may be very difficult even for a human annotator to decide about sentence boundaries (Bouma and Adesam 2013). The compilers of the HaCOSSA corpus solved this problem by abandoning the sentence as the maximal syntactic unit, opting instead for assigning this role to clauses where all the obligatory arguments of the main (lexical) verb have been included (including subordinate clauses) and any non-clausal dependents of the main lexical verb, but not, for instance, adverbial subordinate clauses or all relative clauses (Höder 2011).

⁶<https://universaldependencies.org/ext-format.html>

12 *Towards a Swedish diachronic corpus*

tion (Section 3), we suggest that the following columns should be present for each word in a text (where only the word form column needs to be assigned a value, and unassigned values are represented by an underscore):⁷

1. ID: Token index, where the value is an integer starting at 1 for each new sentence.
2. SDC:XID: ‘Native ID’ used in the resource, e.g., chapter+verse number used in a bible text
3. FORM: Word form or punctuation symbol as used by language tools (implies a lowercased standardized form, possibly in modern spelling).
4. SDC:F_FORM: Form corresponding to the Menota facsimile transcription level (see Section 3)
5. SDC:D_FORM: Form corresponding to the Menota diplomatic transcription level (see Section 3)
6. SDC:S_FORM: Standardized spelling of a word form (form corresponding to the Menota normalized transcription level; see Section 3)
7. LEMMA: Lemma (base form) of the word, as used in a standard dictionary or in modern spelling.
8. UPOS: Coarse-grained part-of-speech tag (from the Universal POS tag set: <https://universaldependencies.org/u/pos/index.html>).
9. XPOS: Fine-grained part-of-speech tag (including possible named-entity codes for proper nouns).
10. FEATS: Morphosyntactic feature specification
11. HEAD: Syntactic head of the current token, represented by 0, if the current token is the root of the sentence, or else by an ID value.
12. DEPREL: Dependency relation to the HEAD.
13. DEPS: In the CoNLL-U Plus format, the DEPREL is typically understood to be one of the universal dependency (UD) relations (see <https://universaldependencies.org/u/dep/index.html>). Some texts may come with dependency analyses already in place reflecting different formats, e.g. that used by PROIEL (Eckhoff et al. 2018). The CoNLL-U DEPS column may then be used to capture this information.

⁷As per CoNLL-U Plus conventions, project-specific columns are given a namespace prefix (“SDC:”). The order of columns as listed here is not final.

14. MISC: Miscellaneous information not belonging in any of the other columns.

This format presupposes tokenisation, resulting in a file with one word or punctuation mark on each line, and blank lines separating sentences. To prepare for automatic linguistic annotation using the UDPipe package (Straka and Straková 2017), we intend to use the Swedish UDPipe tokeniser that is part of the CoNLL 2018 Shared Task baseline models (Straka 2018).

Apart from the CoNLL-based format, we also intend to provide download possibilities for plaintext format (untokenised, ‘raw’ text) and a simplified XML format similar to the format currently used in *Språkbanken Text*.

5

METADATA

Metadata, containing information about a text, its contents and the context in which it was written, is a singularly important piece of information in any corpus. In the Swedish diachronic corpus, metadata information will be added at the top of each file. In the CoNLL files as well as in the plaintext files, each piece of metadata information will be given in a separate line, and be preceded by a hashtag sign ('#') and a unique, predefined label specifying the sort of information given in this line, such as 'author', 'title' etc. In the XML format, metadata will be given in a more TEI-like style, following the XML standards for structuring information hierarchically.

To cover commonly occurring metadata information as well as the requirements expressed by the researchers replying to the user questionnaire described in [Pettersson and Borin \(2019b\)](#), we plan to include the following metadata labels (where labels for which information is missing may be omitted):

- ID: unique ID for referencing this particular text
- author: name of the author (first name followed by surname)
- translator: name of the translator (first name followed by surname), in case of translations
- title
- subtitle
- originalTitle: source language title, in case of translations
- manuscriptDate: estimated dating of the manuscript on which the digital edition is based; may be a single year, a specific date (yyyy-mm-dd) or a time span
- originDate: estimated dating of the original manuscript; may be a single year, a specific date (yyyy-mm-dd) or a time span

- retrieveDate: the date when the digital edition was accessed (yyyy-mm-dd)
- sourceDescription: free text description of the textual content
- genre: religion, fiction, letter, charter, court record, law, regulation, account, register, diary, personal story, song text, periodical, governmental, academic, map, newspaper, essay, school, social media, non-fiction (could be extended)
- subgenre: biblical, legend, revelation, prayer, hymn, prose, play, chronicle, poem, tänkebok, by-law, medicine, astrology, home remedies, protocol, yearbook, science, social science, humanities (could be extended)
- location: geographical location in which the text was produced, as precise as possible (can be hierarchical, depending on the size of the location, e.g. village>parish>county)
- language: ISO 639-3 language code
- languageVariety: language variety, such as Finno-Swedish, skånska (Scanian Swedish) etc.
- codeswitching: ISO 639-3 language code(s) for language(s) occurring in the document, in addition to the main language
- originalLanguage: ISO 639-3 language code for the source language, in case of translations
- manuscript: name of the manuscript on which the digital edition is based
- manuscriptChapter
- manuscriptPages
- printer: name of the printer (first name followed by surname)
- printedVolume: name of the volume in which the manuscript has been printed
- printedIssue
- printedPages
- printedDate
- editor
- digitisationMethod: digitisation method (manually transcribed, OCR-scanned with manual post-correction or OCR-scanned without manual post-correction), or born-digital
- transcriptionPrinciples: transcription principles (diplomatic transcription, standardised spelling, abbreviation expansion etc.)
- transcriber: name of the transcriber (first name followed by surname)

16 *Towards a Swedish diachronic corpus*

- retrievedFrom: URL, organization or contact person from which the text has been accessed
- retrieveFormat: format in which the digital edition was retrieved, e.g. txt, docx or PDF
- words: number of words in the text
- sentences: number of sentences (or sentence-equivalents) in the text
- URL: URL reference to digital edition
- cite: reference to publication to be cited when using the text in research
- availability: licence statement (possibly with URL reference)
- misc: any additional information not covered by the above, in free text

6

USER INTERFACE

The long-term goal is to provide a web-based user interface offering search functions for words and phrases, as well as for lemmas, different spelling variants of the same word form, parts-of-speech, morphological forms, syntactic categories and named entities, as described further in Section 3. An important feature of the planned user interface is to offer functions facilitating comparisons – both absolute and statistical – over time of selected linguistic features (words, syntactic constructions, sequences of linguistic items, etc.).

However, in the first version of the corpus, the user-interface functionality will be limited to search facilities for words and phrases.

Apart from an online search interface, we will also offer the texts for download, in any of the three formats described in Section 4.

For both search and download, we will provide the user with the possibility to select his/her own subcorpus, based on metadata information such as author, time period, genre/subgenre, geographical location, language variety and/or principles for digitisation, transcription and annotation (notably: manual or automatic).

7

SUMMARY AND CONCLUSIONS

In this report, we have presented our plans for the very first version of a Swedish diachronic corpus, containing data from all time periods from Old Swedish to present-day Swedish. The main principle for the corpus structure is to include as many texts as possible, while at the same time trying to balance the contents both regarding text types and time periods included.

To enable comparative studies, both within Swedish and between Swedish and other languages, we also focus on a subpart of the corpus following a similar structure as the *Corpus of Historical American English* (COHA). Likewise, to meet the increasing interest in the language used in social media over time, another subpart of the corpus will contain texts from blogs and chats, produced from 1998 onwards.

To enable advanced search queries and analysis of the data, we plan to incorporate rich linguistic annotation in future versions of the Swedish diachronic corpus. As a consequence of these plans, the main corpus format will be a CoNLL-based format with columns prepared for spelling standardization, lemmas, morphological forms, syntactic analysis and named entity information.

Furthermore, a range of metadata labels are suggested, to cover both commonly occurring pieces of metadata information, and needs expressed by potential users of the corpus.

Finally, the user interface will initially be limited to searches based on words and phrases. Apart from web-based search, we will also provide download possibilities. For both search and download, the user will have the possibility to select his/her own subcorpus, based on metadata.

REFERENCES

- Borin, Lars, Markus Forsberg, Martin Hammarstedt, Dan Rosén, Roland Schäfer and Anne Schumacher. 2016. Sparv: Språkbanken's corpus annotation pipeline infrastructure. *Proceedings of SLTC*. The Sixth Swedish Language Technology Conference.
- Bouma, Gerlof and Yvonne Adesam. 2013. Experiments on sentence segmentation in Old Swedish editions. *Proceedings of the workshop on computational historical linguistics at NODALIDA 2013*, 11–26. Linköping: LiUEP.
- Buch, Armin, David Erschler, Gerhard Jäger and Andrei Lupas. 2013. Towards automated language classification: A clustering approach. Lars Borin and Anju Saxena (eds), *Approaches to measuring linguistic differences*, 303–327. Berlin: De Gruyter Mouton.
- Buchholz, Sabine and Erwin Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. *Proceedings of the 10th conference on computational natural language learning (CoNLL-X)*, 149–164. New York City: ACL.
- Dalin, Anders Fredrik. 1850–1853. *Ordbok öfver svenska språket. Vol. I–II*. Stockholm: Joh. Beckman.
- Davies, Mark. 2012. Expanding horizons in historical linguistics with the 400 million word Corpus of Historical American English. *Corpora* 7 (2): 121–157.
- Eckhoff, Hanne, Kristin Bech, Gerlof Bouma, Kristine Eide, Dag Haug, Odd Einar Haugen and Marius Jøhndal. 2018. The PROIEL treebank family: a standard for early attestations of Indo-European languages. *Language Resources and Evaluation* 52 (1): 29–65.
- Eisenstein, Jacob, Brendan O'Connor, Noah A. Smith and Eric P. Xing. 2014. Diffusion of lexical change in social media. *PLoS ONE* 9 (11): e113114.
- Höder, Steffen. 2011. Phrases and Clauses Tagging Manual for syntactic analyses of Old Nordic texts encoded as Menotic XML documents (PaC-Man). Version 2.0. Publication date 2011-05-11.

- Östling, Robert. 2015. Word order typology through multilingual word alignment. *Proceedings of ACL/IJCNLP 2015 (Volume 2: Short papers)*, 205–211. Beijing: ACL.
- Östling, Robert. 2016. Studying colexification through massively parallel corpora. Päivi Juvonen and Maria Koptjevskaja-Tamm (eds), *The lexical typology of semantic shifts*, 157–176. Berlin: De Gruyter Mouton.
- Pettersson, Eva. 2016. Spelling normalisation and linguistic analysis of historical text for information extraction. Ph.D. diss., Department of Linguistics and Philology, Uppsala University, Uppsala.
- Pettersson, Eva and Lars Borin. 2019a. Characteristics of diachronic and historical corpora: Features to consider in a Swedish diachronic corpus. Swe-Clarín Report Series (SCRS), no. SCR-01-2019. <https://sweclarin.se/sites/sweclarin.se/files/diachronic-corpora-sweclarin-v3.pdf>.
- Pettersson, Eva and Lars Borin. 2019b. Swedish diachronic texts: Resources and user needs to consider in a Swedish diachronic corpus. Swe-Clarín Report Series (SCRS), no. SCR-02-2019. <https://sweclarin.se/sites/sweclarin.se/files/swedish-historical-texts.pdf>.
- Schlyter, Carl Johan. 1877. *Samling af Sweriges gamla lagar. Corpus iuris Sueo-Gotorum antiqui – Bd 13: Ordbok till samlingen af Sweriges gamla lagar. Glossarium ad corpus iuris Sueo-Gotorum antiqui*. Stockholm: Gleerup.
- Söderwall, Knut Fredrik. 1884. *Ordbok öfver svenska medeltids-språket. Vol I–III*. Lund: Svenska fornskriftsällskapet.
- Straka, Milan. 2018. CoNLL 2018 Shared Task - UDPipe Baseline Models and Supplementary Materials. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.
- Straka, Milan and Jana Straková. 2017. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. *Proceedings of the CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies*, 88–99. Vancouver: ACL.