Abstract: The recently compiled Swedish Diachronic Corpus offers access to a total of approximately 16 billion words, covering texts from the 13th century onwards. The corpus contains 14 main genres, with a number of subgenres, compiled from a wide range of sources, including corpus providers and libraries as well as individual researchers and private citizens. All texts in the corpus follow a consistent format, are extensively annotated with metadata, and freely available for download. We firmly believe that the existence of a Swedish diachronic corpus among the resources offered by CLARIN will open up avenues to new, interesting research questions within humanities research, and be a valuable resource for large-scale studies of the Swedish language throughout history – studies that have previously been impossible to conduct in a thorough and consistent manner. Thanks to its embedding in the CLARIN context it also carries the potential to enable broad historical studies from a comparative European perspective.

Keywords: diachronic corpus, historical corpora, corpus linguistics, digital humanities, language change

1 Introduction

History as an academic discipline is often understood as dealing with that part of our past which coincides with the existence of writing (whatever came before that is normally referred to as prehistory). The study of history from various points of view is central to the humanities and social sciences, which are the core research areas supported by the CLARIN research infrastructure. Thus, it is only natural

Acknowledgements: The research reported here has been enabled by the Swedish national research infrastructure SWE-CLARIN, supported (in equal parts) by an infrastructure grant from the Swedish Research Council (Språkbanken & SWE-CLARIN; contract 2017-00626) and by its 10 partner institutions.

*Corresponding author: Eva Pettersson, Uppsala University, Uppsala, Sweden, e-mail: eva.pettersson@lingfil.uu.se
Lars Borin, Språkbanken Text, University of Gothenburg, Gothenburg, Sweden, e-mail: lars.borin@svenska.gu.se

Open Access. © 2022 the author(s), published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License. https://doi.org/10.1515/9783110767377-022
that historical corpora figure prominently among the CLARIN resource families.¹ For similar reasons, diachronic language resources was one of five thematic activities defined by the Swedish CLARIN consortium in 2019, to be pursued jointly among consortium members in various constellations in order to develop the infrastructure.

This activity – coordinated by the Swedish CLARIN K-centre for Diachronic Language Resources (DiaRes)² – has consisted mainly in the compilation of the corpus described in this chapter.

Corpora containing comparable texts from several stages of the historical development of a language are a prerequisite for enabling large-scale studies of language change and linguistic phenomena occurring in texts from different time periods. Consequently, diachronic corpora of this kind are a very valuable resource for many disciplines in the humanities and social sciences, including digital humanities, historical linguistics, literature, history, and others.

Although historical language corpora have been compiled for a long time – most notably covering various periods in the history of English (e.g., Biber, Finegan, and Atkinson 1994; Kroch and Taylor 2000; Kroch, Santorini, and Delfs 2004; Taylor et al. 2003), but also other languages³ – diachronic corpora in the sense intended here are, by and large, a product of the last decade, in which we have seen the creation of the Corpus of Historical American English (COHA; Davies 2012), the Register in Diachronic German Science Corpus (RIDGES) (Odebrecht et al. 2017), the Icelandic Parsed Historical Corpus (IcePaHC) (Rögnvaldsson et al. 2012) and the BDCamões Collection of Portuguese Literary Documents (described by Silva et al. (2022) in this volume), to name a few. For Swedish, however, a diachronic corpus has so far been lacking.

This chapter presents the Swedish Diachronic Corpus, a freely available resource with a total of approximately 16 billion words, covering Swedish texts from the 13th century onwards. We start by giving a short overview of the conventionally recognised historical stages of Swedish in Section 2, before introducing the methods used and the considerations taken during the compilation of the Swedish Diachronic Corpus, for instance concerning text selection (Section 3). We then move on to describing the structure and contents of the resulting corpus

¹ For more information about the CLARIN resource families initiative, see Fišer, Lenardič, and Erjavec (2018), where the first batch of resource types is described. Historical corpora have since been included as part of the second batch: https://www.clarin.eu/resource-families.
² https://sweclarin.se/eng/centers/diares
³ For instance, a collaboration between the Universities in Gothenburg and Lund, Sweden, resulted in several million words of digitized Old and Early Modern Swedish texts, which were made available to researchers in the late 1990s. See https://project2.sol.lu.se/fornsvenska/.
in Section 4, and wrap up by discussing how this resource could be useful to researchers interested in studying the Swedish language over time, from different perspectives. Finally, in Section 6 conclusions are drawn and some ideas for future work are presented.

2 The historical stages of Swedish

To be able to construct a corpus of texts from different time periods throughout the history of Swedish, we need to first define the stages of the Swedish language development. Any sharp division into time periods could be questioned, since languages change gradually. One of the most common ways of describing the Swedish language evolution is outlined in for example Bergman (1995), where the history of the language is divided into five stages: Runic Swedish (ca. 800–1225), Old Swedish (ca. 1225–1526), Early Modern Swedish (ca. 1526–1732), Late Modern Swedish (ca. 1732–1900) and Contemporary Swedish (1900 onwards).

2.1 Runic Swedish (ca. 800–1225)

As implied by the name “Runic”, the texts from this time period are written using the runic alphabet. It could also be noted that the languages spoken and written in the different parts of Scandinavia during this time period are very similar, and often regarded as dialects of one language, referred to as Old Nordic.4

2.2 Old Swedish (ca. 1225–1526)

The Old Swedish period is often defined as starting around 1225; Västgötalagen (‘the Westrogothic Law’) is one of the most important documents from this period, as it is written in Latin script (as opposed to Runic Swedish). Old Swedish is characterized by influences from Latin and Greek, due to the establishment of (Catholic) Christianity, and from German, due to trading relations with the Hanseatic League. It also has a considerably more complex morphology than present-day Swedish, with a wider range of inflections for case, gender, and different verb forms.

4 This language is also called Old Norse and Old Scandinavian.
2.3 Early Modern Swedish (ca. 1526–1732)

Following the Old Swedish period, the (Early) Modern Swedish era is generally considered to have started with the publication of a Swedish translation of the New Testament in 1526, commissioned by King Gustav Vasa. This translation was widely disseminated, partly due to the emergence of the printing press, and partly due to a Swedish Church law from 1686, requiring clergymen to ensure that people knew important passages of the bible. This widespread use of one and the same text led to a more standardized orthography.

2.4 Late Modern Swedish (ca. 1732–1900)

The Modern Swedish period is sometimes divided into the Early Modern period and the Late Modern period, where the publication of the first issue of the periodical Then Swänska Argus in 1732 marks the beginning of the Late Modern era. Due to its genre, this text has a more informal style of writing than the bible texts.

An important milestone in the linguistic history of Swedish, taking place during the Late Modern period, is the foundation of the Swedish Academy in 1786, contributing to a comprehensive standardization of orthography, which was basically completed by the early 19th century. Many loan words enter the language from French.

2.5 Contemporary Swedish (ca. 1900–)

The Contemporary Swedish period starts around 1900, with two important events influencing the Swedish language: the author August Strindberg’s breakthrough novel Röda rummet (The Red Room) in 1879, and the spelling reform of 1906, since which the orthography of Swedish has in principle remained unchanged. Other characteristics of the Contemporary Swedish period are the abandonment of plural verb forms, the shift towards a more informal, colloquial language, and English loan words entering the language at a higher rate. These developments belong mainly to the period since the middle of the 20th century, which is often referred to as Present-Day Swedish.
3 Compiling the corpus

The corpus compilation work has had both a top-down and a bottom-up aspect, as described by Ljubešić et al. (2022) in this volume. It has been top-down in its conception, as a thematic activity of the Swedish CLARIN consortium, and in most of the activities described below in this section, while the bottom-up element has manifested itself primarily in how the actual content of the corpus – the texts – have become available for inclusion in it (see Section 4).

To secure a high-quality corpus, both content-wise and from a user perspective, three preparatory steps were taken prior to actually compiling the corpus: (i) a survey of existing historical and diachronic corpora for various languages; (ii) a survey of textual resources available for Swedish; and (iii) a user questionnaire.

3.1 Survey of diachronic resources for various languages

As a first step, we conducted a survey of existing historical and diachronic corpora for different languages, with a special focus on the structure and contents of these corpora. The goal of this study was to identify important aspects to be taken into consideration in the development of the Swedish Diachronic Corpus, and how to structure the corpus in order for it to be comparable to other diachronic (and historical) corpora. In this survey, we studied diachronic corpora for Czech, English, Faroese, French, Georgian, German, Hungarian, Icelandic, Portuguese, Slovene, and Spanish. One of the main sources for finding these corpora was the CLARIN Resource Family for Historical Corpora.5 See pettersson and Borin (2019a) for more details on this survey.

The first thing to note is that these corpora vary considerably in size, ranging from 53,000 words in the Faroese Parsed Historical Corpus (FarPaHC), to 400 million words in the Corpus of Historical American English (COHA), to many billions of words in the Google Books Ngram Corpus. The corpus size is highly dependent not only on available textual resources for the language in question, but also on the level of annotation included in the corpus, and the quality of this annotation. Typically, the smaller corpora in the study are carefully annotated by humans with features such as part-of-speech, lemma, morphological, and syntactic analysis, enabling the user to formulate more advanced search queries with high-quality results. Larger corpora, on the other hand, are crucial for extensive studies of linguistic change over time, but are generally either not annotated

---

5 https://www.clarin.eu/resource-families/historical-corpora
at all, or (semi-)automatically annotated. From this, we concluded that for the
Swedish Diachronic Corpus we want to provide as large a corpus as possible to
enable large-scale diachronic studies, with a representative part of this corpus
being manually annotated, offering advanced, high-quality search possibilities.

A related consideration is which subcorpora to include, with regard to time
periods (granularity) and genres (balance and representativeness). In order for
comparisons between time periods to be truly commensurable, the composition
of texts from different genres should be as equally distributed over the periods
compared as possible. Otherwise, results from studies supposedly investigating
language change on the basis of such datasets might indicate differences between
genres, rather than differences between certain time periods (unless special
methods for comparing datasets of different sizes are incorporated). At the same
time, the amount of existing historical texts is limited, especially for certain time
periods and genres. Furthermore, some genres do not exist in all time periods,
while other genres might prosper during one era, but be far less common during
another era. In line with our aim of compiling an open-ended corpus that will
grow over time, we therefore include as many texts as we can find in the corpus,
especially for the earlier time periods and the less commonly found genres. At a
later stage, we aim to create a subset of the corpus that will be more balanced,
and better suited to different kinds of comparative studies, as well as develop
auxiliary language tools to support such studies.

One example of a well-balanced corpus is COHA, covering the time period
1810–2009. This corpus contains four genres (fiction, popular magazines, newspa-
pers, and non-fiction books), and is divided into decades, with all four genres rep-
resented for (nearly) every decade. In the Swedish Diachronic Corpus however, we
want to cover a much longer time period, meaning that it might be harder to find
texts from the same genre for all time periods included. Newspapers, for example,
did not exist in the earliest time periods. One way to overcome this would be to
divide the texts into broader domains, such as fiction vs. non-fiction, or try to
find one or a few genres that exist for many (if not all) time periods covered, for
example, legal texts or church documents. For the corpus to be representative
of the language at a given point in time, several other genres should however be
included as well, to reflect the actual text production during that period, such as
charters and religious legends for earlier stages of the language and newspapers
and social media texts to represent present-day text production.

Although of course genres, likewise, may not always be stable over time with regard to their
linguistic characteristics. Fortunately, a large diachronic corpus such as this one will allow
scholars to investigate language change from many different viewpoints simultaneously, which
arguably should enhance the reliability of the conclusions.
A very important, yet quite often neglected, aspect of corpus compilation, is the metadata included for each text in the collection, which enables users to select relevant texts for their particular research questions, and to get to know the material from different aspects. For the corpora studied in the survey, information on title, author, and publication year is (almost) always provided (the few exceptions being very old texts, where these features may be unknown). Since historical texts may occur in many versions, metadata for older texts often also state which edition the text represents. Other metadata elements included in the corpora are genre, sub-genre, number of words/characters/bytes, edits done during transcription (including more formal edits like the representation of characters not included in the Unicode scheme and “correction” of line breaks inside words, etc., as well as more advanced edits, such as spelling standardization), publisher, editor, transcriber, annotator, volume, issue, language variety (Early New Modern, etc.), region in which the text was produced, availability (license etc.), extent of the sample (if not the full text), levels of linguistic annotation, and general notes.

Ideally, it would of course be desirable to include very detailed metadata, to facilitate for the users of the corpus to identify texts of interest to them. However, especially in the case of the oldest texts, even basic information such as author or publication year may be missing or unreliable, meaning that the level of metadata information available may vary greatly between different texts. Thus, for some texts it might only be possible to include limited and less reliable metadata information.

Finally, the format in which the texts, and the metadata associated with them, are stored and made downloadable, needs to be considered. For storage, it is important to use a format suitable not only for storing the actual text, but also for providing metadata information and various levels of linguistic annotation. For download purposes, it could be beneficial to provide a format that the user recognizes from other corpora, so that s/he does not have to learn and understand an entirely unknown format. The most common formats for storage and metadata information in the corpora studied in the survey are:

1. a plaintext format, possibly with headers containing metadata information (using a standardized terminology);
2. a tab-separated format (such as CoNLL) with slots reserved for certain pre-defined annotation elements;
3. a TEI (XML) format;\textsuperscript{7} or

\textsuperscript{7} Text Encoding Initiative; see https://tei-c.org/.
4. a table listing all the texts and their metadata contents, typically in a spreadsheet format (e.g., xls), or as an HTML table

3.2 Survey of existing Swedish resources

When planning the structure and contents of the Swedish Diachronic Corpus, it was crucial to have an idea of what types and amounts of text are available for different periods of the history of Swedish. In the second step, we therefore surveyed the digital textual resources available (or potentially available) for Swedish, for different time periods and for different genres. What is available, in what format, and what is needed to include the text(s) in the corpus? In this phase, we browsed available corpus repositories, and also reached out to researchers and relevant e-mail lists to inquire about material. In this way, we have managed to strike a good balance between publicly available corpus resources and private collections of texts. See Pettersson and Borin (2019b) for more details.

As could be expected, the volumes of text available increase the closer we get to the present day, with about 4.6 million words for the Old Swedish era, as compared to almost 10.7 billion words for contemporary Swedish, as illustrated in Table 1.

<table>
<thead>
<tr>
<th>Time Period</th>
<th>Number of words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Old Swedish</td>
<td>4,641,408</td>
</tr>
<tr>
<td>Early Modern Swedish</td>
<td>24,700,328</td>
</tr>
<tr>
<td>Late Modern Swedish</td>
<td>1,516,865,748</td>
</tr>
<tr>
<td>Contemporary Swedish</td>
<td>10,696,957,453</td>
</tr>
</tbody>
</table>

The survey also shows that there are a number of different types of text available, ranging from more formal laws, governmental texts, and scientific publications to secular prose, song lyrics and newspapers, to informal diaries and letters. Five genres (court records, laws and regulations, religious texts, scientific text, and secular prose) contain texts from all the targeted time periods, 8 “Available” in this context means “available in a more or less directly usable format from the sources listed in this section”. Much more material than this could be scraped off the internet, in particular for present-day language.
enabling comparative studies of the same text genre over the whole time span covered by the corpus.

3.3 User questionnaire

To get an idea of the needs and wishes of the primary target users for a Swedish diachronic corpus, we sent out a questionnaire comprising of eight questions to 15 linguists – specialists in Swedish historical linguistics – asking about their experience of working with historical corpora, and which features would be important in order for a Swedish diachronic corpus to be useful for them. In the following, we present the questions and a short summary of some of the answers from the 12 scholars who responded to the questionnaire:

1. Are there any research questions within your field where you think a Swedish diachronic corpus could be useful? If so, how?

   All the researchers who answered the questionnaire agree that a Swedish diachronic corpus would be very useful, or even essential, for their research. It is suggested that such a corpus could be used, for example (i) for quantitative hypothesis testing based on qualitative findings; (ii) for contrastive studies between phenomena occurring in several languages or language varieties (such as Swedish in Sweden as opposed to Fenno-Swedish); or (iii) for finding unpredictable patterns in a large and differentiated text material. The specific areas of research mentioned by the participants in the questionnaire as relevant for using a Swedish diachronic corpus are historical morphology (e.g., what stems certain derivative suffixes connect to in different time periods), historical phonology, and historical sociolinguistics, as well as studies on language change (such as lexicalization or semantic change over time), syntax, spelling, word order, word frequencies, stylistics, and variation in texts from different time periods and locations, and in texts written by people with different dialects.

2. Have you previously used any existing diachronic (or historical) corpus, for Swedish or for any other language? If so, which corpus did you use, and what did you think was good with the design and the contents of that corpus? And what could have been done differently to make the corpus more useful to you?

   All participants in the survey have experience of using historical corpora in their research, in one way or another. At the same time, several researchers point out that it is often hard to find the relevant texts for their research, since the texts are not gathered in one place, and that the lack of a common corpus format makes it time-consuming to learn new formats and to make comparisons between texts.
Other disadvantages mentioned are the incompleteness of available corpora, the uncertainty about the quality of the transcription and the annotation, and that there is often no way of knowing how much editing the text has undergone as compared to the original. In addition, it is often hard to search in corpora that have been OCR-scanned without manual post-correction, due to bad OCR quality. There is also a need for better graphical user interfaces, such as the one in Korp (Borin, Forsberg, and Roxendal 2012) or CQP (Evert 2019).

As advantageous corpus features, several researchers mention the capability to download texts to process them on their own computer, rather than being limited to a search interface. Moreover, the capability to view a search word (or phrase) in its context, using concordances, is also strongly desired by many users, enabling a quick qualitative assessment of the semantic, collocational, or morphological relevance of the word or phrase in its context. A clear chronological structure of the texts is also vital for being able to select appropriate texts.

3. What time period should be covered by a Swedish diachronic corpus?

The general answer to this question is “as much as possible”. All participants agree that the period from Old Swedish (1200s) and onwards should be included. Furthermore, most researchers think that it would be good to include Runic Swedish as well, whereas some argue that it is already available through Samnordisk runtextdatabas (Scandinavian Runic-text Database), though this is not linguistically annotated, and that Runic Swedish is a bit too far from Swedish as we know it today to make it useful for conducting comparative studies that include this particular time period.

4. How do you think the corpus should be structured with respect to the time periods covered? Should it for example be divided into decades, 50-year periods or 100-year periods? Or rather periods defined within historical linguistics, such as Old Swedish, Early Modern Swedish, etc.?

The majority of the participants in the questionnaire emphasize that the best thing would be for the user to be able to define his/her own subcorpora, to avoid being stuck with predefined time periods that might not suit particular research questions and interests. However, if the corpus should be divided into predefined time periods, periods based on a certain number of years are generally preferred over linguistically motivated periods, since this yields a more fine-grained division.

\[9\] https://spraakbanken.gu.se/korp/
\[10\] https://www.nordiska.uu.se/forskn/samnord.htm/
5. Are there any particular text types/genres that you think should be included in the corpus?

The general answer to this question is “the more the better”, and if large amounts of text are available, it is up to the user to create his/her own balanced subcorpus, if needed. At the same time, one user remarks that it is important that there is some kind of balance in the corpus, so that no particular genre is strongly overrepresented, for example because that particular text type is easier to find. A balance of texts distributed over different time periods is also requested. On the other hand, it is important that the corpus reflects the genre development over time, so that the design of the corpus is not limited to the genres present for the earlier stages of the language.

An interesting observation about this question is that the users are specifically interested in texts that are hard to find, such as informal texts written by “ordinary” people. Particular text types mentioned are letters, diaries, texts written in different dialects, and texts that represent the spoken language in one way or another, for example drama. Tänkeböckerna (medieval court records) are also called for.

6. How would you spontaneously interpret the term “Swedish” in the expression “Swedish diachronic corpus”? Should, for example, all texts written within the borders of Sweden (past or present?) be included, regardless of whether they are written in Swedish or, for example, Finnish or Latin? Or should only texts written in Swedish be included? Should the corpus also include texts written in Swedish outside the borders of Sweden, which could then include, for example, texts in Fenno-Swedish or American Swedish?

There is a broad consensus among the researchers that “Swedish” in the context of the Swedish diachronic corpus should be defined as “texts written in the Swedish language”, including language varieties such as Fenno-Swedish, American Swedish, and different Swedish dialects. It is, however, pointed out that it could be useful to mark these texts as such, in order for the users to be able to select or deselect specified language varieties.

7. Are there any specific parameters that you would like to be able to search for in the corpus?

Some participants in the questionnaire think that a (word-based and phrase-based) lexical search is enough, stating that the most important factor for their research is to have access to large amounts of text from different time periods, and that they do not trust annotation that has been added automatically, due to annotation errors, especially for older texts. For these researchers, quality is more important than quantity, and there are suggestions that we should only annotate smaller parts of the corpus, but do it manually, or make only a coarse annotation,
such as part-of-speech tagging, which has a higher chance of being correct. Alternatively, we could put effort into improving the OCR quality instead, since this is often a troublesome issue.

Most researchers are, however, interested in (automatic) linguistic annotation, which enables them to formulate more advanced search queries. Linguistic annotation types specifically mentioned as interesting are lemmatization and/or truncation of words, morphology, part-of-speech tagging, phrase structures and syntactic categories. Personal names and place names could also be useful in studies of sociolects and geographical location. For semantics, the meaning of the words is important, in order to distinguish between colexified senses.

8. What metadata categories should be included in the corpus

As might be expected, many researchers emphasize that as much metadata as possible is desirable. Features mentioned as particularly interesting are author, year, genre, and geographical location. It is also suggested that it would be good to classify texts as being written in different language varieties, such as Fenno-Swedish, American Swedish, etc., but it might sometimes be hard to make such a classification, due to uncertainty and different opinions on how to classify a specific text in this aspect.

Other metadata elements mentioned are volume, issue, and the name of the editor. For manuscripts that are copies of older texts, the name of the scribe who copied the later manuscript is also important. Even the printer could be of importance, since some printers had their own orthographical norm. For some research questions, the age of the author is relevant, too, and if the text is digitally available as fulltext, there should be a link. Finally, there is a suggestion that we should add a short presentation of the text, to give the user an idea of the contents of the text.

4 Contents of the corpus

Based on the findings from the three preparatory steps described in the previous section, we decided on the structure and contents of a first version of the Swedish Diachronic Corpus. In the following, we describe the principles for inclusion of texts (regarding time periods, genres, and text providers), the format of the corpus, the levels of linguistic annotation, and the metadata elements attached to each text in the corpus.
4.1 Time periods

In the first version of the Swedish Diachronic Corpus, we have chosen to include texts from the 13th century onwards, thus excluding Runic Swedish (see Section 2). The reasons for this are threefold. First, nearly all existing Nordic Runic texts are already accessible through Samnordisk runtextdatabas (Scandinavian Runic-text Database),¹¹ a database containing approximately 6,500 inscriptions, with the texts represented in a transliterated and standardized form, and with a translation in English. This means that researchers working with Runic Swedish may find all texts of interest collected there. Secondly, excluding Runic Swedish means that all texts in the corpus are written using the same alphabet, facilitating comparative studies as well as formatting and annotation issues. Thirdly, the runic inscriptions are quite different content-wise from texts written during other time periods. Starting with Old Swedish, we see text genres that occur in several (or all) time periods, opening up for interesting comparisons and research questions.

At the time of writing, the oldest texts are from the 13th century, and the most recent texts are from 2017. It is, however, worth noticing that the Swedish Diachronic Corpus is designed to be an open-ended corpus,¹² meaning that the contents of the corpus are not static, but new texts will be added over time, to extend the breadth and depth of the corpus.

4.2 Genres

Labelling a text as belonging to a certain genre is an important feature in the corpus, enabling the user to select texts according to his/her interest. This labeling task is, however, not always as trivial as it may seem. First of all, how specific should the genre categories be? Should, for example, prayers and hymns be genres of their own, or should these text types simply be categorized as religious texts? This was also commented on by one of the researchers answering the user questionnaire (see Section 3.3), who pointed out that classifying texts as “religious texts” would not be a good idea, since there is quite a difference in text type between religious legends and, for example, prayers.

Secondly, how do we choose a genre when there are several genres that could be considered for a specific text? A hymn, for example, could be labelled either as a religious text or as song lyrics, depending on which characteristics of the text

¹¹ https://www.nordiska.uu.se/forskn/samnord.htm/
¹² In other words, a kind of monitor corpus, although not entirely prototypical as such.
that are considered most important. Likewise, a medical journal could either be
classified as a periodical or as a scientific text.

In the Swedish Diachronic Corpus, we have chosen to classify the texts into a
smaller number of main genres, with additional subgenres to account for differ-
ences between text types within a specific genre. This way, a religious legend will
be grouped with the main genre “religion” and the subgenre “religious prose”.

Furthermore, many of the texts included in the Swedish Diachronic Corpus
were provided by people or organizations that had already classified the texts as
belonging to a certain genre. One example is Fornsvenska textbanken (Delsing
2002), containing a collection of machine-readable editions of Old Swedish and
Early Modern Swedish texts, covering the time period 1162–1758. These texts are
subclassified into seven genres: laws, diplomas and court records, medicine,
secular prose, religious prose, verse, and accounts. We have followed their genre
classification when including these texts in the Swedish Diachronic Corpus, with
some minor changes (such as making “medicine” a subgenre of “scientific text”,
for example).

The resulting corpus contains 14 main genres, with a number of subgenres, as
listed below (subgenres in parentheses):

– Court (court records, parish meetings, judgments)
– Governmental (acts, bills, investigations, memorials, protocols, reports, etc.)
– Informal (diaries, folklore, personal stories)
– Laws and regulations (by-laws, church regulations, laws)
– Letters and charters (letters, charters)
– Lyrics (song lyrics)
– Newspapers
– Political pamphlets
– Periodicals (culture, medicine, politics and economy, popular science,
women and society)
– Religious texts (bible texts, hymns, postils, prayers, religious prose)
– Scientific and academic text (agriculture, astrology, humanities, medicine,
natural science, protocols, social sciences)
– Secular prose (biographies, children’s literature, drama, essays, fiction, folk-
lore, humour, non-fiction, novels, poetry, proverbs, short stories, speeches)
– Student writings (social science)
– User-generated text (blogs, chats, Wikipedia)

Five of the main genres (court records, laws and regulations, religious texts, sci-
entific text, and secular prose) are represented in all the time periods targeted by
the corpus, enabling comparative studies of the same text genre over the whole
time span.
4.3 Text providers and corpus size

One aim in the corpus compilation process has been to collect texts from many different sources, both from corpus providers and from more private collections. To achieve this, we implemented an approach of both searching on the platforms of known corpus providers and also sending out requests to researchers in the field of historical linguistics, as well as email lists targeted at a digital humanities audience. The result is a collection of texts in the Swedish Diachronic Corpus, retrieved from:

1. **established corpus providers**, such as Dramawebben (The Drama Web),13 Fornsvenska textbanken (Delsing 2002), Litteraturbanken (Swedish Literature Bank),14 Språkbanken Text (text division of the National Swedish Language Bank),15 Project Gutenberg16 and Project Runeberg;17
2. **libraries**, such as Göteborgs universitetsbibliotek (Gothenburg University Library),18 Kungliga biblioteket (National Library of Sweden),19 and Uppsala universitetsbibliotek (Uppsala University Library);20
3. **archives**, such as Folklivsarkivet (the Folklife Archives)21 and Riksarkivet (Swedish National Archives);22
4. **local history societies** interested in historical texts, such as Jämtlands läns fornskriftsällskap;23
5. **research projects**, such as the Gender and Work project (Ågren et al. 2011);
6. **individual researchers**, such as Anna Wallberg Gustafsson at Lund University, who has contributed political pamphlets, and Professor Harry Lönnroth at University of Jyväskylä, who has contributed court records;
7. **private citizens**, such as Guno Haskå, volunteer at Demografisk Databas Södra Sverige (Demographical Database for Southern Sweden), who has contributed court records.

---

13 https://litteraturbanken.se/dramawebben
14 https://litteraturbanken.se/om/english.html
15 https://sprakaarkbanken.gu.se/
16 https://www.gutenberg.org/
17 http://runeberg.org
18 https://gupea.ub.gu.se/
19 https://www.kb.se/kb-in-english.html
20 https://ub.uu.se/
21 https://www.folklivsarkivet.lu.se/en/
22 https://riksarkivet.se/startpage
23 http://www.fornskrift.se/
In the first version of the corpus, released in 2020, we have limited the texts considered for inclusion to texts that are already digitized. This version of the corpus comprises approximately 16 billion words.

### 4.4 Format

The format(s) used for corpus storage and user access is of utmost importance for the usefulness of the corpus. Our aim is to be able to provide the corpus in one or more formats that (i) are easy to use and understand; (ii) are standardized and used in other corpora as well; (iii) can hold both metadata information and linguistic annotation in an intuitive and transparent way; and (iv) have the potential to be easily integrated in existing search interfaces, etc. To meet these requirements, we decided on three formats:

1. a plain text format, with a metadata header at the top of each file (see Section 4.6);
2. a tab-separated, CoNLL-U Plus format, with slots reserved for a predefined set of linguistic annotation elements (see Section 4.5) and a metadata header at the top of each file;
3. an XML format (not implemented in the first version of the corpus).

### 4.5 Annotation

Regarding annotation, we aim to include a wide range of linguistic features for each text. In order not to delay the appearance of the first version of the corpus, released in 2020, we have decided to include only linguistic markup already present in the source text, thus not adding any new annotation. This presents a potential obstacle to harmonization, for example for search and comparison across corpus components. However, most of the texts are, at present, either not annotated at all or annotated with largely compatible part-of-speech tagsets and dependency syntax. In particular, as opposed to English, for example, there are no historical corpora with legacy phrase-structure annotation for Swedish, meaning that aiming to add (minimal) UPOS and UD annotations to all the datasets makes a lot of sense.

For representing the linguistic annotation, we use the tab-separated CoNLL-U Plus format.\(^{24}\)

---

\(^{24}\) https://universaldependencies.org/ext-format.html
Table 2: Annotation in the Swedish Diachronic Corpus.

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1  ID</td>
<td>Token index (integer, starting at 1 for each new sentence)</td>
</tr>
<tr>
<td>2  SDC:XID</td>
<td>‘Native ID’ used in the resource (e.g., chapter+verse number used in a bible text)</td>
</tr>
<tr>
<td>3  FORM</td>
<td>Word form or punctuation symbol as used by language tools (lower-cased standardized form, possibly in modern spelling)</td>
</tr>
<tr>
<td>4  SDC:F_FORM</td>
<td>Form corresponding to the Menota facsimile transcription level</td>
</tr>
<tr>
<td>5  SDC:D_FORM</td>
<td>Form corresponding to the Menota diplomatic transcription level</td>
</tr>
<tr>
<td>6  SDC:S_FORM</td>
<td>Standardised spelling of a word form (form corresponding to the Menota normalized transcription level)</td>
</tr>
<tr>
<td>7  LEMMA</td>
<td>Lemma (base form) of the word, as used in a standard dictionary or in modern spelling</td>
</tr>
<tr>
<td>8  UPOS</td>
<td>Coarse-grained part-of-speech tag (from the Universal POS tag set: <a href="https://universaldependencies.org/u/pos/index.html">https://universaldependencies.org/u/pos/index.html</a>)</td>
</tr>
<tr>
<td>9  XPOS</td>
<td>Fine-grained part-of-speech tag (including possible named-entity codes for proper nouns)</td>
</tr>
<tr>
<td>10 FEATS</td>
<td>Morphosyntactic feature specification</td>
</tr>
<tr>
<td>11 HEAD</td>
<td>Syntactic head of the current token (represented by 0 if current token is the root of the sentence, or else by an ID value)</td>
</tr>
<tr>
<td>12 DEPREL</td>
<td>Dependency relation to the HEAD</td>
</tr>
<tr>
<td>13 DEPS</td>
<td>In the CoNLL-U Plus format, the DEPREL is typically understood to be one of the universal dependency (UD) relations (see <a href="https://universaldependencies.org/u/dep/index.html">https://universaldependencies.org/u/dep/index.html</a>). Some texts may come with dependency analyses already in place reflecting different formats, e.g. that used by PROIEL (Eckhoff et al. 2018). The CoNNL-U DEPS column may then be used to capture this information.</td>
</tr>
<tr>
<td>14 MISC</td>
<td>Miscellaneous information not belonging in any of the other columns</td>
</tr>
</tbody>
</table>

Table 2 shows the columns that should be present for each word in a text in the diachronic corpus (where only the token index and the word form columns need to be assigned a value, and unassigned values are represented by an underscore).²⁵

²⁵ As per CoNNL-U Plus conventions, project-specific columns are given a namespace prefix (“SDC:”).
4.6 Metadata

To enhance searchability and to enable the user to select only the parts of the corpus that are relevant to his/her specific research interests, we have developed a scheme of 42 metadata features, including author, title, date (original date, manuscript date and publication date), genre and subgenre, language variety, printer, digitization method, transcription principles, level of linguistic annotation, source, license, number of words and sentences, and more. What metadata elements to include was decided incrementally. First, a set of initial metadata elements were chosen, based on the elements suggested by the Text Encoding Initiative, combined with the authors' own experience of corpus work and corpus use. These were then revised and extended based on input from the user questionnaire (Section 3.3) and the metadata elements suggested by the text providers. Table 3 shows the resulting set of metadata elements present in the Swedish Diachronic Corpus.

Table 3: Metadata elements in the Swedish Diachronic Corpus.

<table>
<thead>
<tr>
<th>Metadata element</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>unique ID for referencing this particular text</td>
</tr>
<tr>
<td>author</td>
<td>author's name; first name followed by surname</td>
</tr>
<tr>
<td>authorBorn</td>
<td>author's date of birth; single year (yyyy) or date (yyyy-mm-dd)</td>
</tr>
<tr>
<td>pseudonym</td>
<td>pseudonym used for this text; first name followed by surname</td>
</tr>
<tr>
<td>translator</td>
<td>translator's name; first name followed by surname</td>
</tr>
<tr>
<td>title</td>
<td>title of the text</td>
</tr>
<tr>
<td>subtitle</td>
<td>subtitle of the text</td>
</tr>
<tr>
<td>originalTitle</td>
<td>source language title (in case of translations)</td>
</tr>
<tr>
<td>manuscriptDate</td>
<td>date of the manuscript on which the digital edition is based; single year (yyyy), specific date (yyyy-mm-dd) or time span</td>
</tr>
<tr>
<td>originDate</td>
<td>date of the original manuscript (may be different from the manuscript on which the digital edition is based); single year (yyyy), specific date (yyyy-mm-dd) or time span</td>
</tr>
<tr>
<td>retrieveDate</td>
<td>date when the digital edition was accessed (yyyy-mm-dd)</td>
</tr>
<tr>
<td>sourceDescription</td>
<td>free text description of the textual content</td>
</tr>
<tr>
<td>genre</td>
<td>main genre of the text (e.g., “religion” or “secular prose”)</td>
</tr>
<tr>
<td>subgenre</td>
<td>subgenre of the text (e.g., 'bible text' or 'poetry')</td>
</tr>
<tr>
<td>location</td>
<td>geographical location in which the text was produced</td>
</tr>
</tbody>
</table>

26 https://tei-c.org
### Table 3 (continued)

<table>
<thead>
<tr>
<th>Metadata element</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>language</td>
<td>ISO 639-3 code for the main language in the text</td>
</tr>
<tr>
<td>languageVariety</td>
<td>language variety (e.g., Fenno-Swedish)</td>
</tr>
<tr>
<td>codeswitching</td>
<td>ISO 639-3 code(s) for language(s) occurring in the document, in addition to the main language</td>
</tr>
<tr>
<td>originalLanguage</td>
<td>ISO 639-3 code for source language (in case of translations)</td>
</tr>
<tr>
<td>manuscript</td>
<td>name of the manuscript on which the digital edition is based</td>
</tr>
<tr>
<td>manuscriptChapter</td>
<td>manuscript chapter(s) on which the digital edition is based; single chapter or span of chapters</td>
</tr>
<tr>
<td>manuscriptPages</td>
<td>manuscript page(s) on which the digital edition is based; single page or page span</td>
</tr>
<tr>
<td>printer</td>
<td>name of the printer; first name followed by surname</td>
</tr>
<tr>
<td>printedVolume</td>
<td>name or number of volume in which the manuscript is printed</td>
</tr>
<tr>
<td>printedIssue</td>
<td>name or number of issue in which the manuscript is printed</td>
</tr>
<tr>
<td>printedPages</td>
<td>page(s) in the volume/issue containing the actual text; single page or page span</td>
</tr>
<tr>
<td>printedDate</td>
<td>publication date for printed version of the text; single year (yyyy) or date (yyyy-mm-dd)</td>
</tr>
<tr>
<td>editor</td>
<td>name of the editor; first name followed by surname</td>
</tr>
<tr>
<td>publisher</td>
<td>name of the publisher (person or company)</td>
</tr>
<tr>
<td>digitisationMethod</td>
<td>digitization method: manually transcribed, OCR-scanned with manual post-correction, OCR-scanned without manual post-correction, or born-digital</td>
</tr>
<tr>
<td>transcriptionPrinciples</td>
<td>transcription principles: diplomatic transcription, standardised spelling, abbreviation expansion, etc.</td>
</tr>
<tr>
<td>transcriber</td>
<td>transcriber’s name; first name followed by surname</td>
</tr>
<tr>
<td>retrievedFrom</td>
<td>URL, organization or person from which the text was retrieved</td>
</tr>
<tr>
<td>retrieveFormat</td>
<td>format in which the text was retrieved, e.g., txt, docx, or PDF</td>
</tr>
<tr>
<td>annotation</td>
<td>levels of linguistic annotation added to the text</td>
</tr>
<tr>
<td>annotationMethod</td>
<td>annotation method: manual, automatic, or semi-automatic</td>
</tr>
<tr>
<td>words</td>
<td>number of words in the text</td>
</tr>
<tr>
<td>sentences</td>
<td>number of sentences in the text</td>
</tr>
<tr>
<td>sentenceOrder</td>
<td>order of the sentences: original or shuffled (due to copyright)</td>
</tr>
<tr>
<td>URL</td>
<td>URL reference to digital edition</td>
</tr>
<tr>
<td>cite</td>
<td>reference to publication to be cited when using the text</td>
</tr>
<tr>
<td>availability</td>
<td>license statement (possibly with URL reference)</td>
</tr>
</tbody>
</table>
4.7 Download and search

An important aspect of the Swedish Diachronic Corpus, is that the user should be able to download and use it, without license restrictions. Hence, all the texts in the corpus are freely accessible and downloadable from the project website. To deal with copyright issues, the order of the sentences has been shuffled in some of the modern texts. This way, researchers may still study phenomena occurring within sentences, even though we are not allowed to share the running text. The metadata element sentenceOrder clearly marks whether the original sentence order is preserved in a text or not, enabling the user to disregard texts with a randomized sentence order.

As seen from Figure 1, the website displays one entry for each main genre. Clicking on a genre displays the subgenres associated with the main genre, along with information on the time period covered, the total number of words, a short readme file, and two download links: one for the plaintext version of the texts and one for the CoNLL-U Plus version (with slots for linguistic annotation; see Section 4.5). It is also possible to sort the columns by genre, time period, or number of words.

For search queries, we intend to integrate the Swedish Diachronic Corpus in Korp, since this search interface contains a majority of the features requested by the users in the user questionnaire (see Section 3.3).

5 Uses and usefulness of the corpus

As is also pointed out by Silva et al. (2022) in this volume (concerning the BDCamões Collection of Portuguese Literary Documents), a diachronic corpus spanning several centuries of text, with a variety of genres and authors, will exhibit a wide range of different orthographic and syntactic traditions, and open up to exciting new areas of research. Furthermore, in the user questionnaire sent out to a number of researchers in Swedish historical linguistics as part of the preparations for compiling the Swedish Diachronic Corpus, we asked (among other things) for their opinions on the usefulness of a Swedish diachronic corpus. As described in Section 3.3, all the researchers who answered the questionnaire agreed that such a corpus would be very useful, or even essential, for their research. Specific areas of research pointed out by the users were:

27 https://cl.lingfil.uu.se/svediakorp/
28 https://spraakbanken.gu.se/korp/
quantitative hypothesis testing based on qualitative findings;
contrastive studies of phenomena occurring in several languages or language varieties (such as Swedish in Sweden as opposed to Fenno-Swedish);
finding hitherto unstudied patterns in a large and differentiated text material;
historical morphology (e.g., what stems certain derivational suffixes combine with at different time periods);
historical phonology and historical sociolinguistics;
studies of language change (such as lexical or semantic change);
syntax, spelling, word order, word frequencies, stylistics and variation in texts from different time periods and locations, and in texts written by people with different dialects.

After the release of the first version of the Swedish Diachronic Corpus, in October 2020, it has also been reported that the corpus is used for teaching in courses on the history of the Swedish language, and for developing a BERT language model for named entity recognition in historical Swedish texts.
6 Conclusion and future work

In this chapter, we have presented the Swedish Diachronic Corpus, a corpus of approximately 16 billion words, spanning from the Old Swedish period (ca. 1225–1526), over the Early Modern Swedish period (ca. 1526–1732) and the Late Modern Swedish period (ca. 1732–1900) to Contemporary Swedish texts (1900 onwards). The corpus contains a mix of texts from many different sources; established corpus providers as well as libraries and archives, local historical societies, individual researchers, and private citizens. The texts are classified into 14 main genres, with a number of subgenres, and a set of 42 metadata elements gives the user as clear a picture as possible about each text. Furthermore, all texts in the corpus are freely available for download and use.

The corpus is intended as an open-ended (monitor) corpus, meaning that new texts will be added to the corpus over time, for example by digitization of texts from time periods for which only smaller amounts of data are currently available, or by adding data for the most recent years.

Apart from supplementing the corpus with additional texts, there are a number of features that we would like to work on for future, updated versions of the corpus. First, the texts are currently only available for download. We plan to also offer a search option, by integrating the corpus into the Korp interface. In connection with this, and to broaden the search options, we also intend to add linguistic annotation to more texts in the corpus, including lemma (to enable users to search for a word in all its inflectional forms), part-of-speech, morphology, and syntax. Named entities – identifying names of persons and locations – may also be of interest. In addition, for the older texts (as well as for some of the most recent social media texts), spelling standardization and diachronic linkage of lexical entries would be useful for enabling the user to search for a word regardless of its form in a particular textual material from a particular time in history.

For the download function, we plan to add XML as one of the download formats (something which is also beneficial for the integration into Korp). Furthermore, we want to make it possible for the user to define his/her own time intervals for downloading files, instead of being limited to the zip files currently available on the project website.

Future work also includes defining a sub-corpus within the Swedish Diachronic Corpus, with balanced sets of texts for different time periods, with regard to the amount of text as well as the genres included. In this project, we plan

29 https://spraakbanken.gu.se/korp/
to follow the structure of existing diachronic corpora for other languages, for example the COHA corpus (Davies 2012).

Swedish is one of several closely related Nordic standard languages with long written histories, for all of which digitized historical texts are available to varying extents. An obvious extension of the Swedish Diachronic Corpus would be to include it in a larger Nordic diachronic corpus, where the linguistic closeness of the languages (which increases the further back in time we go) will readily allow, for example, reliable annotation transfer among the languages. For more on this topic, see the contribution by Ljubešić et al. (2022), where a similar scenario is presented for (Western) South Slavic languages, for which the time depth to their common proto-language is about the same as in the case of the Nordic languages.

To sum up, we firmly believe that the existence of a Swedish diachronic corpus among the resources offered by CLARIN will open up avenues to new, interesting research questions within humanities research. It goes without saying that it will be a valuable resource for large-scale studies on the Swedish language throughout history – studies that have previously been impossible to conduct in a thorough and consistent manner – and thanks to its embedding in the CLARIN context it also carries the potential to enable broad historical studies in a comparative European perspective. For instance, the Sweden-Swedish data used in the comparative Swedish-Finnish investigation reported on by Fridlund et al. (2022) is one of the components of the Swedish Diachronic Corpus, viz. the historical newspapers made available by the National Library of Sweden and Språkbanken Text. In the same way that modern corpora have thoroughly re-shaped lexicographic practice (see Petrauskaité et al. (2022); Rauset et al. (2022)), the Swedish Diachronic Corpus could inform historical dictionary projects such as the large Swedish Academy Dictionary (Svenska Akademien 1898–), and complement existing lexicons covering particular historical periods (see, e.g., Adesam et al. 2021).

30 For instance through the Medieval Nordic Text Archive (Menota) – https://menota.org/EN_forside.xhtml – or the Old Norse treebanks available through the CLARINO centre INESS: https://clarino.uib.no/iness/treebanks.
Bibliography


Ljubešić, Nikola, Tomaž Erjavec, Maja Miličević Petrović & Tanja Samardžić. 2022. Together we are stronger: Bootstrapping language technology infrastructure for South Slavic languages with CLARIN.SI. In Darja Fišer and Andreas Witt (eds.), CLARIN. The infrastructure for language resources. Berlin: De Gruyter.


