SWE-CLARIN

RAPPORTER FRÅN SWE-CLARIN
SWE-CLARIN REPORT SERIES
sweclarin.se/scrs
[SCR-02-2019]

Swedish diachronic texts

Resources and user needs to consider in a Swedish diachronic corpus

Eva Pettersson[1] and Lars Borin[2]

[1]Department of Linguistics and Philology, Uppsala University, eva.pettersson@lingfil.uu.se
[2]Språkbanken Text, University of Gothenburg, lars.borin@svenska.gu.se

Contents

1 Introduction 1

2 Corpus providers 3

2.1 Alvin 3

2.2 Dramawebben 3

2.3 Fornsvenska textbanken 4

2.4 The Gender and Work project (GaW) 5

2.5 HaCOSSA 5

2.6 Jämtlands läns fornskriftsällskap 6

2.7 Litteraturbanken 6

2.8 Medieval Nordic Text Archive (Menota) 7

2.9 Project Runeberg and Project Gutenberg 7

2.10 Samnordisk runtextdatabas 8

2.11 Språkbanken Text 9

2.12 Svenska fornskriftställskapet 9

2.13 Vetenskapssocieten i Lund 10

3 Corpora 11

3.1 Runic Swedish (800–1225) 12

3.2 Old Swedish (1225–1526) 13

3.2.1 Religious texts 13

3.2.2 Secular prose 15

3.2.3 Letters and charters 15

3.2.4 Scientific text (medicine) 16

3.2.5 Court records 17

3.2.6 Laws and regulations 18

3.2.7 Accounts and registers 18

3.2.8 Old Swedish: Summary 19

3.3 Early Modern Swedish (1526–1732) 21

3.3.1 Religious texts 21

3.3.2 Secular prose 22

3.3.3 Diaries and personal stories 22

3.3.4 Song texts 23

3.3.5 Periodicals 23

3.3.6 Letters and charters 23

3.3.7 Academic text 24

3.3.8 Court records 24

3.3.9 Laws and regulations 26

3.3.10 Governmental texts 26

3.3.11 Accounts and registers 26

3.3.12 Maps 27

3.3.13 Text fragments 28

3.3.14 Early Modern Swedish: Summary 28

3.4 Late Modern Swedish (1732–1900) 30

3.4.1 Religious texts 30

3.4.2 Secular prose 30

3.4.3 Diaries and personal stories 32

3.4.4 Periodicals 33

3.4.5 Newspaper text 33

3.4.6 Letters 34
3.4.7 Academic and scientific text 34
3.4.8 Court records 34
3.4.9 Laws and regulations 35
3.4.10 Governmental texts 36
3.4.11 Late Modern Swedish: Summary 36
3.5 Contemporary Swedish (1900–) 38
3.5.1 Religious texts 38
3.5.2 Secular prose 38
3.5.3 Diaries and personal stories 39
3.5.4 Song texts 39
3.5.5 Periodicals 39
3.5.6 Newspaper text 40
3.5.7 Letters 40
3.5.8 Academic and scientific text 40
3.5.9 Court records 41
3.5.10 Laws and regulations 42
3.5.11 Governmental texts 42
3.5.12 Essays and school-related texts 42
3.5.13 Accounts and registers 42
3.5.14 Social media text and Wikipedia 43
3.5.15 Contemporary Swedish: Summary 43

4 User questionnaire 46

4.1 Questionnaire responses: summary 47
4.1.1 Research questions for a diachronic corpus 47
4.1.2 Previously used corpora 48
4.1.3 Time periods covered 50

4.1.4 Division into time periods 51
4.1.5 Text types covered 51
4.1.6 The definition of Swedish 52
4.1.7 Search parametres 53
4.1.8 Metadata information 54
4.2 Questionnaire responses: Discussion and implications 55

5 Summary and conclusions 57
5.1 Future work 59

References 62

1 Introduction

The CLARIN research infrastructure[1] (short for "Common Language Resources and Technology 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-Clarin*,[2] 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 comparative studies of the Swedish language development as compared to other languages for which diachronic corpora exist.

In a preceding companion report, we investigated the structure and contents of diachronic and historical corpora available for other languages, with the aim to identify important aspects to be taken into consideration in the development of a Swedish diachronic corpus, and how the corpus could be structured in order for it to be comparable to other diachronic and historical corpora (Pettersson and Borin 2019). When planning for the structure and contents of the Swedish diachronic corpus, it is also crucial to have an idea of what text types and amounts of text that are available for different time periods throughout the history of the Swedish language. In the current report, we therefore continue our work towards a Swedish diachronic corpus by examining textual resources available for the Swedish language, for different time periods and for different genres. In addition, we want to take the needs and wishes of the primary target users into consideration when building the corpus. For this reason, we sent out a questionnaire to a number of researchers in the humanities with a special interest in historical linguistics and the development of the Swedish language, asking for their experience of historical corpora, and what features that are important in order for a Swedish diachronic corpus to be useful for them.

[1] https://www.clarin.eu/

[2] https://sweclarin.se/

The report is structured as follows: In Section 2, we present corpus providers of special interest to our goals, and in Section 3, we give an overview of available textual resources for Swedish, based on the traditional division into time periods, i.e. Runic Swedish (appr. 800–1225), Old Swedish (appr. 1225–1526), Early Modern Swedish (appr. 1526–1732), Late Modern Swedish (appr. 1732–1900), and Contemporary Swedish (appr. 1900–) (Bergman 1995). The answers to the user questionnaire are reported in Section 4. Finally, the findings are summarized and conclusions are drawn in Section 5, including some directions for future work.

2 Corpus providers

In this section, we give a short presentation of corpus providers that are of special interest to the development of a Swedish diachronic corpus.

2.1 Alvin

Alvin[3] is a national platform for long-term preservation of and accessibility to digitised collections and digital cultural heritage material from Swedish cultural heritage institutions. It is hosted by the Uppsala University Library, in collaboration with the Gothenburg University Library and Lund University Library. Several other cultural heritage institutions are also members of Alvin.

The Alvin database contains a wide range of genres, and is constantly growing, as new material is digitised by the libraries. Most of the digitised material in Alvin is free to download and use, but material from the 1920s and onwards may be copyright-protected. Unfortunately for our purposes, most of the texts in Alvin are only available as images, and there is currently no indexing in the database to tell which documents that are also available in a transcribed format.

More information on the Alvin platform can be found here:

- Alvin webpage: http://info.alvin-portal.org/

2.2 Dramawebben

Dramawebben[4] (The Drama Web) aims to make texts representative of the Swedish history of theatre available to the general public. The corpus includes first edition dramas for adults and for children, distributed over widely different genres, such as comedy, tragedy, drama, farce, and fairy plays. In total,

[3] http://info.alvin-portal.org/

[4] https://litteraturbanken.se/dramawebben

Dramawebben distributes 456 dramas, written by 49 male and 78 female authors, born in the time period 1579–1894. About 200 of these plays have been OCRed.

Dramawebben is part of Litteraturbanken (see further Section 2.7) since 2018. Part of the corpus (currently 790,456 tokens) is also available for download via Språkbanken Text (see further Section 2.11), and searchable through the Korp interface.

More information on Dramawebben can be found here:

- Dramawebben webpage: https://litteraturbanken.se/dramawebben/om
- Dramawebben at Språkbanken Text: https://spraakbanken.gu.se/swe/resurs/drama

2.3 Fornsvenska textbanken

Fornsvenska textbanken (Delsing 2002) is a collection of machine-readable editions of Old Swedish and Early Modern Swedish texts, covering the time period 1162–1758. The collection currently contains approximately 1.2 million words, distributed over seven genres: laws, diplomas and court records ("tänkeböcker"), medicine, secular prose, religious prose, verse, and accounts. The texts are not collected to be a balanced selection of texts, and the website also states that the texts are distributed without quality assurance.[5]

The texts are displayed on the project website, and may also be downloaded as RTF files. At the website, metadata information is given for each text, including title, year, edition, genre, and information on any edits that have been done during transcription, as compared to the original manuscript. No licensing information is provided, but the same texts are available in XML format as part of Språkbanken Text's historical corpora under a Creative Commons attribution licence (see Section 2.11).[6]

More information on Fornsvenska textbanken can be found here:

- Fornsvenska textbanken webpage: http://project2.sol.lu.se/fornsvenska/
- Delsing, Lars-Olof. 2002. Fornsvenska textbanken. Svante Lagman, Stig Örjan Olsson and Viivika Voodla (eds), *Nordistica tartuensia 7*, 149–156. Tallinn: Pangloss.

[5] http://project2.sol.lu.se/fornsvenska/

[6] http://creativecommons.org/licenses/by/4.0/

2.4 The Gender and Work project (GaW)

Within the *Gender and Work* project (GaW) at the Department of History, Uppsala University, historians are interested in what men and women did to support themselves in the time period 1550–1800. Part of their research consists in searching for text passages revealing this information in different historical sources, and storing the information in the *Gender and Work database* (Ågren et al. 2011).[7] In connection with this, several historical sources have been OCRed and/or manually transcribed, with a special focus on court records.

More information on the Gender and Work project can be found here:

- GaW webpage: https://gaw.hist.uu.se/
- GaW database: http://gaw.ddb.umu.se:8080/gaw-query/query/index.xhtml
- Ågren, Maria, Rosemarie Fiebranz, Erik Lindberg and Jonas Lindström. 2011. Making verbs count. The research project 'Gender and Work' and its methodology. *Scandinavian Economic History Review* 59 (3): 271–291.

2.5 HaCOSSA

The *Hamburg Corpus of Old Swedish with Syntactic Annotations* (HaCOSSA) (Höder 2011a) is a morphologically and syntactically annotated corpus of 13 texts (both whole texts and excerpts) from the Late Old Swedish period (approximately 1375–1550), with a total of 128,204 words. The genres included in the corpus are religious and secular prose, law texts, non-fiction literature (geography, theology, history, natural science), and diplomas.

The texts are provided in XML format, following the standards of TEI P5 and MENOTA 2.0. The corpus is annotated with morphological categories, syntactic functions, clause types, clause linking strategies, complex verbs, direct speech, and code-switching, using the PaCMan 2.0 annotation scheme (Höder 2011b). However, only preverbal constituents and subjects are provided with full annotation.

The HaCOSSA corpus texts are freely available under a CC-BY license,[8] and a

[7] https://gaw.hist.uu.se/

[8] The publisher of the scientific manuscript editions from which the HaCOSSA texts have been digitized, *Svenska fornskriftsällskapet* (see further Section 2.12), has generously released these texts under a CC-BY license.

fully POS-tagged[9] and UD-annotated[10] version is under preparation at Språkbanken Text.

More information on the HaCOSSA corpus can be found here:

- HaCOSSA webpage:
  https://corpora.uni-hamburg.de/hzsk/en/islandora/object/text-corpus%3Ahacossa
- Höder, Steffen. 2011a. The Hamburg Corpus of Old Swedish with Syntactic Annotation (HaCOSSA). Archived in Hamburger Zentrum für Sprachkorpora. Version 1.0. Publication date 2011-06-30. http://hdl.handle.net/11022/0000-0000-9D16-7.
- Höder, Steffen. 2011b. Phrases and Clauses Tagging Manual for syntactic analyses of Old Nordic texts encoded as Menotic XML documents (PaCMan). Version 2.0. Publication date 2011-05-11.

2.6 Jämtlands läns fornskriftsällskap

Jämtlands läns fornskriftsällskap[11] is a society aiming to contribute to the exploration of the history of the Swedish counties Jämtland and Härjedalen, by publishing historical records and other unprinted source material. The members have transcribed a number of texts, some of which are printed in books available for purchase from their website, whereas other texts are freely available for download in readable PDF format. This work is done in cooperation with *Landsarkivet i Östersund* and *Föreningsarkivet i Jämtlands län*.

More information on Jämtlands läns fornskriftsällskap can be found here:

- Jämtlands läns fornskriftsällskap webpage:
  http://www.fornskrift.se/

2.7 Litteraturbanken

Litteraturbanken (The Swedish Literature Bank)[12] aims to make reliable digital versions of Swedish classics, and other texts that are of importance to the Swedish literary heritage, freely available to the general public as well as to students, teachers and scholars.

The number of books available through Litteraturbanken is constantly grow-

[9] With the *Universal Part of Speech Tags*: https://universaldependencies.org/u/pos/

[10] That is, a syntactic dependency analysis expressed using *Universal Dependencies* (v. 2): https://universaldependencies.org/

[11] http://www.fornskrift.se/

[12] https://litteraturbanken.se/om/english.html

ing. At the time of writing, Litteraturbanken contains 3,806 books, distributed over 1,983 different authors, comprising a total of 154,314,090 words. 1,133 of these books are available in plain text and ePub format, based on printed first editions or on later scholarly editions, whereas the rest of the texts are available in facsimile format, based on uncorrected OCR output. Around a hundred of the ePub books are protected by copyright, whereas the rest are freely accessible and downloadable.

So far, there has been a special focus on digitising the collected works of August Strindberg (1849–1912), Carl Jonas Love Almqvist (1793–1866) and Selma Lagerlöf (1858–1940). Currently, Litteraturbanken is working on making first editions of Swedish fiction books printed between 1870 and 1900 available in XML, plain text, PDF and ePub format; a task that is planned to be finished during 2020.

More information on Litteraturbanken can be found here:

- Litteraturbanken webpage: https://litteraturbanken.se/om/english.html
- Litteraturbanken at Språkbanken Text: https://spraakbanken.gu.se/swe/resurs/lb

2.8 Medieval Nordic Text Archive (Menota)

The *Medieval Nordic Text Archive* (Menota) contains a collection of Medieval Nordic texts, mainly manuscripts from the 1200s to the 1500s. In total, the archive contains 1.6 million words, distributed over 43 texts, written in Old Icelandic, Old Norwegian and Old Swedish. The texts are searchable via a web interface, and (all but one) downloadable in XML format with a Creative Commons share-alike licence.[13]

More information on the Medieval Nordic Text Archive can be found here:

- Menota webpage: http://www.menota.org/forside.xhtml

2.9 Project Runeberg and Project Gutenberg

Project Runeberg[14] is a non-profit organisation aiming to publish free electronic editions of Nordic literature. Due to copyright issues for present-day books, Runeberg mainly provides access to older books. The Runeberg website currently contains 4,000 works in Swedish from the time period 1495–2001.

[13] https://creativecommons.org/licenses/by-sa/4.0/

[14] http://runeberg.org

One disadvantage is however that only a few of the texts are downloadable in plain text format, and that the OCR quality varies, with a typically lower quality for older books. The project website lists both more thorough download possibilities and improved OCR quality as ongoing work.

Project Gutenberg[15] is similiar to Project Runeberg, but publishes text for many different languages. They offer more than 58,000 free eBooks, readable online and also downloadable in ePub format or in plain text format. For Swedish, there are 235 books available, written by 146 authors, born in the 1700s or 1800s, with a majority born in the 1800s.

More information on Project Runeberg and Project Gutenberg can be found here:

- Project Runeberg webpage: http://runeberg.org
- Project Gutenberg webpage: https://www.gutenberg.org/

2.10 Samnordisk runtextdatabas

Samnordisk runtextdatabas (Scandinavian Runic-text Database, SRD)[16] is a database of Runic text developed within a project with the same name at the Department of Scandinavian Languages, Uppsala University. The aim of the project is to include all existing Nordic Runic texts, within and outside the Nordic countries, and to make them accessible to researchers in a machine-readable format. The database contains approximately 6,500 inscriptions, with the texts represented in a transliterated and standardized form, and with a translation to English. Each inscription is also associated with information such as time period, finding area, excerpt source, type of object, image links and location coordinates.

The contents of the database are released under a Database Contents License (DbCL v1.0),[17] and may be downloaded from the project website: https://www.nordiska.uu.se/forskn/samnord.htm/?languageId=1.

More information on Samnordisk runtextdatabas can be found here:

- SRD webpage: https://www.nordiska.uu.se/forskn/samnord.htm/

[15] https://www.gutenberg.org/

[16] https://www.nordiska.uu.se/forskn/samnord.htm/

[17] https://www.opendatacommons.org/licenses/dbcl/1-0/index.html

2.11 Språkbanken Text

Språkbanken Text (the Text division of the National Swedish Language Bank) at the University of Gothenburg makes a large number of historical and contemporary Swedish corpora available for online search through the Korp corpus infrastructure (Borin, Forsberg and Roxendal 2012),[18] and also as downloadable datasets in XML format, comprising the texts plus any linguistic annotation and metadata added by the Korp corpus import pipeline.[19] The downloadable corpora are released under a Creative Commons attribution licence,[20] and the format is a light "XML-ification" of a CoNLL-like tabular format, where metadata information is given as attributes of the "text" element of this format. For copyright-protected corpora, the sentences are provided in random order.

The texts represent a heterogeneous mixture of genres and periods (Adesam et al. 2016), from the Old Swedish texts of Fornsvenska textbanken (see further Section 2.3), over a collection of medieval letters digitised by the Swedish National Archives (*Svenskt diplomatarium*, see further Section 3.2.3) and historical novels provided by *Litteraturbanken* (see further Section 2.7), to a large body of old newspapers, the so-called *Kubhist* corpus (see further Section 3.4.5).

More information on Språkbanken Text's historical corpora can be found here:

- Korp search interface, historical mode:
  https://spraakbanken.gu.se/korp/?mode=all_hist#?lang=en
- Språkbanken Text's resource download pages:
  https://spraakbanken.gu.se/eng/resources
- Adesam, Yvonne, Malin Ahlberg, Peter Andersson, Lars Borin, Gerlof Bouma and Markus Forsberg. 2016. Språkteknologi för svenska språket genom tiderna. *Studier i svensk språkhistoria 13*, 65–87. Umeå: Umeå University.

2.12 Svenska fornskriftställskapet

Svenska fornskriftsällskapet[21] was founded in 1843, and aims to publish domestic manuscripts in Swedish and Latin from the Medieval period up to the 16th century, as well as scientific works dealing with such manuscripts. Lit-

[18] https://spraakbanken.gu.se/korp/?mode=all_hist#?lang=en

[19] https://spraakbanken.gu.se/eng/resources

[20] http://creativecommons.org/licenses/by/4.0/

[21] 'The Swedish Old Manuscript Society'; http://www.svenskafornskriftsallskapet.se/

teraturbanken (see further Section 2.7) is currently working on digitising all the material from Svenska fornskriftsällskapet. According to their own assessment, the OCR quality is reasonably high for younger texts and for texts written using the Latin alphabet, whereas the quality is quite low for older texts, especially texts containing Old Swedish characters not in use in present-day Swedish, such as *æ, ø, þ* and *ð*.

More information on Svenska fornskriftsällskapet can be found here:

- Svenska fornskriftsällskapet webpage:
  http://www.svenskafornskriftsallskapet.se/

2.13 VETENSKAPSSOCIETEN I LUND

Vetenskapssocieteten i Lund (The Science Society in Lund) is currently working on digitising and publishing open access the script series *Skånsk senmedeltid och renässans* ('Late Medieval and Renaissance time in Skåne').[22] This series started in 1947, and contains source publications, surveys and petitions concerning the cultural and social life of the Skåne countryside during the Danish period, that is approximately from the middle of the 15th century to the peace in Roskilde in 1658. The society aims to digitise everything that has been published in the series by 2020, but there are already some publications available, as described further in Section 3.

This work is led by an editorial committee consisting of representatives from history, art history, archeology, linguistics, cultural geography, history of ideas, legal history and archival science. Prioritised areas are political and social history, ecclesiastical relationships, school and teaching, art history and business. The work in progress is especially concerned with the issuance of primary sources, with a special focus on texts of financial nature, such as accountings and land registers.

More information on Vetenskapssocieteten i Lund can be found here:

- Vetenskapssocieteten i Lund webpage:
  https://projekt.ht.lu.se/vetenskapssocieteten

[22] https://www.ht.lu.se/serie/85/

3 Corpora

In the following, we present textual resources available for different time periods throughout the history of the Swedish language. Any sharp division into time periods for the development of the Swedish language could be questioned, since language development usually happens gradually. In the following sections, for convenience, we do however present the resources based on the traditional time periods described in for example Bergman (1995):

- Section 3.1: Runic Swedish (appr. 800–1225)
- Section 3.2: Old Swedish (appr. 1225–1526)
- Section 3.3: Early Modern Swedish (appr. 1526–1732)
- Section 3.4: Late Modern Swedish (appr. 1732–1900)
- Section 3.5: Contemporary Swedish (appr. 1900–)

Before presenting the actual texts, there are four points of discussion in the description of the textual resources that should be mentioned.

First of all, we try to give an estimation of the size of the resources presented, preferably in the form of word counts. It should be noted though, that these word counts are highly approximate. For resources where the text provider presents statistics on word counts, we use these numbers. There may however be differences in the way the words are counted between text providers, for example regarding whether or not punctuation signs should be counted as words. In the cases where punctuation signs are counted on a par with words, the term used is normally "token" rather than "word". When no information on word count is given by the text provider, and if we have access to the text in question, we have used the shell command 'wc -w'[23] for counting the words in the file. In such cases, we have cleaned the files from any XML tags, metadata

[23] https://ss64.com/bash/wc.html

markup etc that could be easily detected, before doing the word count, but the text may still contain page numbering, comments from transcribers etc that were not detected in the cleanup phase. The order of magnitude of the word counts should be correct, however.

The second point of discussion relates to the dating of a text. In some cases it is unclear at what point in time a certain text was written, especially for the older periods. Furthermore, many manuscripts are handwritten copies of earlier texts, where the contents may differ to a lesser or greater extent from the original manuscript. Some of these texts might have been dated based on the assumed date of the original manuscript, whereas other texts have been dated based on the assumed date of the actual manuscript at hand.

Thirdly, we have divided the texts into genres, with the aim of exploring differences and similarities in the composition of text types available for the time periods to be included in the Swedish diachronic corpus. There are many ways to classify texts into genres, looking at texts from different points of view and with varying degrees of specificity. We have tried to keep the classification rather coarse, and to follow the categorisation made by the text providers, when applicable. In borderline cases, where it could be argued for a text to be classified as belonging to several genres, we have still only categorised it as belonging to one of these genres.

Finally, there are cases where a text collection contains texts extending over the borders between two time periods. If it is not trivial to separate the texts in such a collection into the two different time periods, we present the textual resource under the heading of the period with the longest time span stated for this collection.

3.1 Runic Swedish (800–1225)

For the Runic Swedish period, *Samnordisk runtextdatabas* (Scandinavian Runic-text Database)[24] provides access to nearly all existing Nordic Runic texts (see further Section 2.10). The database contains approximately 6,500 inscriptions, with the texts represented in a transliterated and standardized form, and with a translation to English. The contents of the database are released under a Database Contents License (DbCL v1.0),[25] and may be downloaded from the project website: https://www.nordiska.uu.se/forskn/samnord.htm/?languageId=1. Ta-

[24] https://www.nordiska.uu.se/forskn/samnord.htm/

[25] https://www.opendatacommons.org/licenses/dbcl/1-0/index.html

ble 1 lists the contents of the Scandinavian Runic-text Database, as given by the Skaldic Project.[26]

Region	**#Inscriptions**
Britain	124
Denmark	1065
Faroe Islands	9
Greenland	102
Ireland	16
Iceland	49
Norway	1653
Sweden: Bohuslän	19
Sweden: Dalarna	5
Sweden: Gotland	409
Sweden: Gästrikland	23
Sweden: Hälsingland	21
Sweden: Jämtland	5
Sweden: Lappland	1
Sweden: Medelpad	18
Sweden: Närke	39
Sweden: Småland	193
Sweden: Södermanland	455
Sweden: Uppland	1484
Sweden: Västergötland	314
Sweden: Värmland	8
Sweden: Västmanland	36
Sweden: Östergötland	459
Sweden: Öland	184
Other regions	45
Total	**6,736**

Table 1: Runic inscriptions included in *Samnordisk runtextdatabas*.

3.2 Old Swedish (1225–1526)

The textual resources presented for the Old Swedish period could be divided into the following seven genres: (i) religious texts; (ii) secular prose; (iii) letters and charters; (iv) scientific text (medicine); (v) court records; (vi) laws and regulations; and (vii) accounts and registers.

3.2.1 Religious texts

A common topic in older texts is the religious theme. For the Old Swedish period, there are purely biblical texts available, as well as religious legends telling the stories about the lives of saints and other important persons in Christianity, and other texts related to Christianity. Table 2 lists Old Swedish religious texts

[26] https://skaldic.abdn.ac.uk/db.php, https://skaldic.abdn.ac.uk/db.php?if=srdb&table=srdb

provided by *Fornsvenska textbanken* (in RTF format), *Svenska fornskriftsällskapet* (as printed books) and *HaCOSSA* (in XML format).

Title	Time Period	#words
Fornsvenska legendariet (several editions)[1]	1276–1308	125,025
Revelations of Heliga Birgitta (several editions)[1,2,3]	1340–1370	531,533
Acts of the Apostles[1]	1385	11,822
Helga manna leverne[1]	1385	30,014
Järteckensboken[1]	1385	29,125
Ängelns diktamen (Sermo Angelicus)[1]	1385	14,143
Gospel of Nicodemus (apocryphal)[1]	1300s	10,825
Pentateuch (two editions)[1]	1300s	145,150
Book of Genesis (from the Pentateuch)[2]	1300s	32,714
Tre heliga konungar[1]	1400–1450	620
Revelations of Saint Bonaventura[1,3]	1420	69,070
Själens tröst (explaining the Ten Commandments)[1]	1420s	145,478
Dagens sju tidegärder[1]	middle 1400s	3,860
Tungulus (Tundalus)[1,3]	1457	2,940
Själens Kloster (Claustrum Animæ)[1]	1460	25,476
Revelations of Heliga Mechtild[1,2]	1469	7,261
Writings of Sankt Bernhard[1]	1480–1500	unknown
Gudeliga snilles väckare[1,3]	1480–1520	unknown
The Book of Esther[1]	1484	6,860
The Book of Judith (apocryphal)[1]	1484	8,480
The Book of Ruth[1]	1484	2,190
Lucidarius[1]	1487	28,173
Joh. Gersons bok Om djefvulens frestelse[3]	1495	unknown
Stimulus Amoris[1]	1498–1502	26,650
Sermones sacri Svecice (sermons)[2,3]	1400s	5,694
Speculum virginum[3]	1400s	unknown
En Wadstena-nunnas bönbok (book of prayers)[1]	end of 1400s	6,960
Sjusovare-sagan[1]	1500–1550	1,865
Wars Herra Pino bok (table readings)[3]	1502	98 texts
Vår herres under[1]	1502	1,100
Vår herres födelse[1]	1502	925
Vår herres barndom[1]	1502	1,527
Jungfru Marie örtagård (hymns, prayers and songs)[3]	1510	unknown
Gersons Lärdom huru man skall dö[3]	1514	unknown
Swedish postils[1,3]	unknown	109,130
Legends about specific persons[1]	unknown	280,750
Total	**1276–1550**	**>1,632,646**

[1] Provided by Fornsvenska textbanken.
[2] Provided by HaCOSSA.
[3] Provided by Svenska fornskriftsällskapet.

Table 2: Old Swedish religious texts.

3.2.2 Secular prose

Due to religion being such a big part of life in the Old Swedish period, texts belonging to the category ‘secular prose’ are sometimes hard to distinguish from the category ‘religious fiction’ for this time period. In Table 3, we follow the classification into secular prose made by *Fornsvenska textbanken*, *HaCOSSA* and *Svenska fornskriftsällskapet* respectively. It could also be noted that a characteristic way of telling stories in the Old Swedish period was to tell them as poems, meaning that several stories listed in the table are written in verse.

Title	**Time Period**	**#words**
Svenska medeltidens rim-krönikor[3]	1229–1520	unknown
Euphemia poem: *Ivan Lejonriddaren*[1,3]	1303	41,741
Euphemia poem: *Fredrik av Normandie*[1,3]	1308	18,980
Euphemia poem: *Flores och Blanzeflor*[1,3]	1311	14,040
Erikskrönikan (Eric Chronicle)[1,3]	1330s	28,280
Konungastyrelsen (how to rule a country)[1,3]	1300s (transcr. 1632)	21,600
Konung Alexander[3]	1380s	unknown
Karl Magnus (Karlamagnús saga)[1,3]	1380–1400	10,930
Sju vise mästare (three editions)[1]	1400s	16,240
Aff Danmarks Konungum (chronicle)[1]	1430s	570
Joan Präst af India land[1,2]	1450	1,460
Herr abboten[1,2]	1457	565
Prosaiska krönikan[1]	1450–1457	4,730
Didrik af Bern (two editions)[1,3]	1450s	53,880
Namnlös och Valentin[1,3]	1450s	15,220
Fru Elins bok[3]	1476	unknown
Den fornsvenska dikten om ett gyllene år[3]	1503	unknown
Old Swedish sayings[1]	unknown	9,450
Total	**1229–1520**	**>237,686**

[1] Provided by Fornsvenska textbanken.
[2] Provided by HaCOSSA.
[3] Provided by Svenska fornskriftsällskapet.

Table 3: Old Swedish secular prose.

3.2.3 Letters and charters

Svenskt Diplomatariums huvudkartotek (SDHK)[27] at the Swedish National Archives, is a database containing information on Medieval letters (charters) connected to Sweden and Swedish conditions, written in the time period 817–1540. The database contains information about approximately 44,000 letters, with possibilites to search for person names, place names, dates etc.

[27] https://sok.riksarkivet.se/sdhk

Svenskt Diplomatarium (Diplomatarium Suecanum)[28] is a printed series of these charters, published by the Swedish National Archives. The publication series started in 1829 by Johan Gustaf Liljegren, and is still ongoing work. So far, all charters written before 1381 have been published, as well as the letters written in 1401 to 1420. The texts are published in their original language (usually Latin or Old Swedish), together with a summary in contemporary Swedish, as well as comments and notes on both linguistic aspects and the actual contents of the text, as explained by the National Archives:

> The intention is to make the medieval texts, often difficult to read, as accessible as possible. This means that they are provided with an explanatory summary of the contents in modern Swedish and that all the abbreviations in the originals are written in full. Additionally, the text editions are accompanied by historical commentaries as well as text critical notes. (https://riksarkivet.se/diplomatarium-suecanum, accessed 2019-04-12)

In *Språkbanken Text*, the Swedish charters from the time period 1208–1512 are published in XML format, with the order of the sentences scrambled. This set of charters contain a total of 967,228 words. In addition, Språkbanken Text also provides access to letters written by King Gustav Vasa in the time period 1521–1525, containing 253,642 words in total.

Moreover, *Svenska fornskriftsällskapet* provides access to 47 charters from the period 1368–1375, documenting the properties of Vadstena Abbey, and to a collection of letters to and from Bishop Hans Brask, written in the years 1522–1527.

3.2.4 Scientific text (medicine)

Table 4 lists books related to medicine, home remedies and astrology, provided by *Fornsvenska textbanken*.

Title	**Time Period**	**#words**
Hästläkedom (equine medicine)	1350–1520	1,363
Medicine books	1400–1525	95,886
Strödda läkedomar	1440–1460	708
Astrology	1450–1550	7,543
Bondakonst (by Peder Månsson)	1515–1520	41,509
Total	**1350–1550**	**147,009**

Table 4: Old Swedish texts about medicine, available through *Fornsvenska textbanken*.

[28] https://riksarkivet.se/diplomatarium-suecanum

3.2.5 Court records

Tänkeböcker are judicial protocols from municipal courts in the Swedish cities during the Middle Ages and up to the 17th century. In addition to the actual court records, these books also contain notes on other events of legal and even political nature. The oldest books are often written in German.

Table 5 presents tänkeböcker from the Old Swedish period, as well as a protocol on an inheritance dispute (between Erik Eriksson and Ture Turesson), available in a digitised format. The book from Kalmar, and the inheritance dispute, are available in RTF format via *Fornsvenska textbanken*. The Kalmar book contains text passages both in Low German and in Swedish.

Svenska fornskriftsällskapet provides tänkeböcker from Arboga (1451–1569) and Uppland (1490–1494). These are currently being OCR-scanned and digitised via *Litteraturbanken* (The Swedish Literature Bank, see further Section 2.7).

The books from Stockholm are available in a readable PDF format via the Gothenburg University Library (GUPEA).[29] The library is still working on OCR scanning and making available more tänkeböcker from Stockholm. So far, tänkeböcker written in the time period 1474–1591 are available, containing a total of approximately 1.7 million words.

Title	**Time Period**	**#words**
Kalmar stads tänkebok[1]	1381–1560	23,400
Inheritance dispute[1]	1451–1480	12,832
Arboga stads tänkeböcker[3]	1451–1569	unknown
Stockholms stads tänkeböcker[2]	1474–1500	500,624
Upplands lagmansdombok[3]	1490–1494	unknown
Stockholms stads tänkeböcker[2]	1504–1524	231,249
Total	**1381–1569**	**>768,105**

[1] Provided by Fornsvenska textbanken.
[2] Provided by GUPEA.
[3] Provided by Svenska fornskriftsällskapet.

Table 5: Old Swedish court records text.

[29] https://gupea.ub.gu.se/

3.2.6 Laws and regulations

Table 6 presents Old Swedish laws and regulations available in RTF format via *Fornsvenska textbanken*, in XML format via *HaCOSSA* or as printed books via *Svenska fornskriftsällskapet*.

Title	**Time Period**	**#words**
Skånelagen[1]	1203–1212	19,100
Äldre Västgötalagen[1]	1220s	15,010
Alsnö stadga[1]	1279	1,572
Biskop Brynjolfs stadga[1]	1281	686
Skenninge stadga[1]	1284	1,191
Östgötalagen (several editions)[1]	1280s	78,710
Yngre Västgötalagen (several editions)[1]	1280s	46,414
Äldre Västmannalagen[1]	1280s	16,260
Upplandslagen (several editions)[1,2]	1297	95,480
Yngre Västmannalagen[1]	1318–1347	18,700
Södermannalagen (several editions)[1]	1327	47,180
Skara stadga[1]	1335	371
Skenninge stadga[1]	1335	1,147
Uppsala stadga[1]	1344	2,238
Bjärköarätten[1]	1345	5,790
Tälje stadga[1]	1345	2,208
Kopparbergsprivilegierna[1]	1347	1,350
Gutalagen[1]	1350	12,540
Hälsingelagen[1]	1350–1400	17,900
Norbergs stadga[1]	1354	1,036
Smålandslagens Kyrkobalk[1]	1300s	3,150
Magnus Erikssons landslag[1]	1350s	45,330
Magnus Erikssons stadslag[1]	1350s	21,050
Vårfrupänningen (maintenance of Vadstena Abbey)[2]	1360–1460	3,318
Kung Kristoffers landslag[1]	1442	55,553
By-law for the economy of Vadstena Abbey[2]	1443	3,207
Election of Confessor Generalis at Vadstena Abbey[2]	1450	3,533
Guild regulations[3]	unknown	unknown
Total	**1203–1460**	**>520,024**

[1] Provided by Fornsvenska textbanken.
[2] Provided by HaCOSSA.
[3] Provided by Svenska fornskriftsällskapet.

Table 6: Old Swedish laws and regulations.

3.2.7 Accounts and registers

Another type of document that appears early in history is accounts and registers. *Vetenskapssocieteten i Lund* (The Science Society in Lund) is currently working on digitising and publishing open access the script series *Skånsk senmedeltid och renässans* (‘Late Medieval and Renaissance time in Skåne’), see further Section 2.13. In connection with this, two accounts from the Old Swedish period have been made available:

- *Palteboken*, real estate register of Lund’s archetypal estate, from the year 1515, containing approximately 97,733 words
- *1522 års uppbördsjordebok*, real estate register from the year 1522, containing approximately 17,335 words

In addition, *Svenska fornskriftsällskapet* provides access to the following accounts and registers:

- Accounts of Greger Mattson, 15th century
- Accounts of Folke Gregersson, 1496–1501
- Accounts of Britta Hansdotter, 1507–1512
- Real estate registers (*jordeböcker*) from Vadstena Abbey, 1500–1502

3.2.8 Old Swedish: Summary

Figure 1 visualizes the estimated number of words per genre for the Old Swedish resources presented throughout this section. Note that the number of words is highly approximate, since texts for which we do not know the word count are excluded from the chart. Moreover, there may be different definitions of what counts as a word for different texts, and metadata information may be included in the word count for some texts, see further the discussion in the introduction to Section 3.

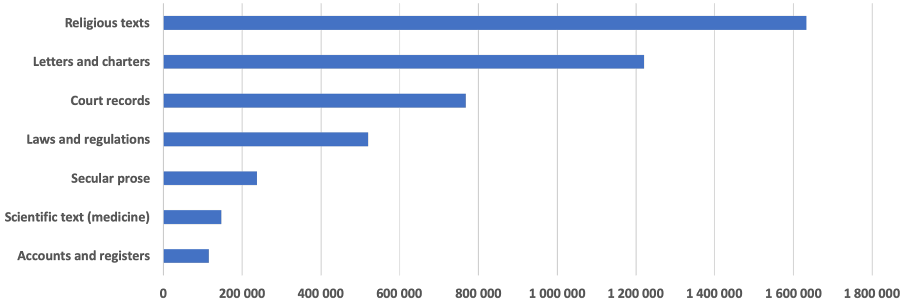

Figure 1: Approximate number of words per genre in the Old Swedish resources.

This being said, we can distinguish seven different genres for the texts from the Old Swedish period: (i) religious texts; (ii) secular prose; (iii) letters and charters; (iv) scientific text (medicine); (v) court records; (vi) laws and regulations; and (vii) accounts and registers.

As could be expected, texts related to religion are a common text resource from this period, where the total number of words amounts to approximately 1.6 million words. Apart from religiously oriented texts, documents concerning legal issues, i.e. court records, laws and regulations, sum up to a total of approximately 1.3 million words. Other rather frequently occurring texts from this time period are (formal) letters and charters, comprising approximately 1.2 million words, whereas the genres of secular prose, scientific text and accounts and registers contain less than 300,000 words each.

Figure 2 shows the approximate number of words per century for the Old Swedish period. As seen from the figure, most texts are from the 1300s and 1400s, with a fairly even distribution between these two centuries, while there are fewer texts available from the oldest century and the youngest century. It should however be pointed out that the Old Swedish period is defined as starting in the year of 1225 and ending in the year of 1526, meaning that the 13th century is only represented by a 75-year period, and the 16th century by a 27-year period, as opposed to 100-year periods for the 1300s and 1400s. Even when this is taken into consideration, it is nevertheless clear that there are smaller amounts of text available for the 13th century, since this time period makes up approximately 25% of the total Old Swedish period, but is only represented by 12% of the available texts. On the other hand, the 16th century makes up approximately 9% of the total period, and is represented by 20% of the texts.

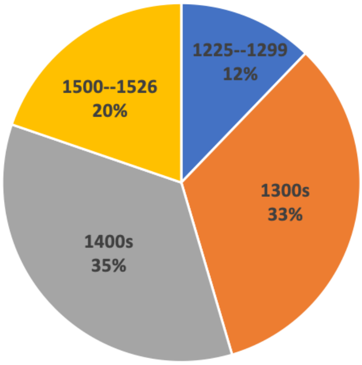

Figure 2: Approximate number of words per century for the Old Swedish resources.

3.3 Early Modern Swedish (1526–1732)

The textual resources presented for the Early Modern Swedish period could be divided into the following twelve genres: (i) religious texts; (ii) secular prose; (iii) diaries and personal stories; (iv) song texts; (v) periodicals; (vi) letters and charters; (vii) academic text; (viii) court records; (ix) laws and regulations; (x) governmental texts; (xi) accounts and registers; and (xii) maps.

3.3.1 Religious texts

Table 7 lists religious texts from the Early Modern Swedish time period, provided by *Fornsvenska textbanken*, *Litteraturbanken*, *Project Runeberg*, *Språkbanken Text* and *Svenska fornskriftsällskapet*.

Title	Time Period	#Words
Acts of the Apostles[1]	1526	11,101
Book of Revelation[1]	1526	10,364
Gospel According to Mark[1]	1526	14,187
Gustav Vasas bibel: The New Testament[4]	1541	196,356
Gustav Vasas bibel: The Book of Genesis[4]	1541	39,295
Jungfru Marie psaltare (Rosary prayer)[4]	1534	unknown
Book of hymns (Olaus Petri)[3]	1536	84 pages
Book of Proverbs[1]	1536	15,460
Dauidz psaltare (Book of Psalms)[5]	1536	unknown
Jesu Syrach book (Wisdom of Sirach)[5]	1536	unknown
Salomos wijsheet (Wisdom of Solomon)[5]	1536	unknown
Book of Genesis[1]	1541	33,956
Gospel According to Mark[1]	1541	13,837
Gospel According to Luke[1]	1541	24,842
Book of Revelation[1]	1541	10,590
War sermon by Johannes Rudbeck[3]	1626	unknown
En sann christendom by Johann Arndt[1]	1647	2,906
The Bible[2]	1703	1,396,656
The Bible: The New Testament[4]	1703	164,286
The Bible: The Book of Genesis[4]	1703	33,479
En sann christendom by Johann Arndt[1]	1732	1,989
Total	**1526–1732**	**>2,000,000**

[1] Provided by Fornsvenska textbanken.
[2] Provided by Litteraturbanken.
[3] Provided by Project Runeberg.
[4] Provided by Språkbanken Text.
[5] Provided by Svenska fornskriftsällskapet.

Table 7: Early Modern Swedish religious texts.

3.3.2 Secular prose

Within the secular prose genre for the Early Modern time period, *Fornsvenska textbanken* provides access to three historical chronicles from the 1500s, written by Olaus Petri, Peder Swart and Per Brahe, as well as the war novel *Historia Trojana* and two humorous books: *Prosastycken på svenska* and *Mål-roo eller Roo-mål*. In addition, *Dramawebben* offers six plays from the 1600s in a readable PDF format, and *Litteraturbanken* has eight books from the time period 1611–1732 in downloadable ePub format. Furthermore, *Språkbanken Text* provides access to two handbooks on Swedish agriculture, written in 1727: *Engelska Åker-Mannen* and *En Grundelig Kundskap Om Swenska Åkerbruket*. Moreover, *Project Runeberg* has scanned 32 books from the period 1607–1725. These are however not available in a reliable digital text format, see further Section 2.9. Table 8 presents these works in more detail.

Title	**Time Period**	**#Words**
Historia Trojana[2]	1529	44,380
Olaus Petris krönika[2]	1530s	108,040
Peder Swarts krönika[2]	1560	51,940
Per Brahes krönika[2]	1585	26,380
Miscellaneous books[4]	1607–1725	unknown
Miscellaneous plays[1]	1611–1649	78,741
Miscellaneous books[3]	1611–1732	183,508
Mål-roo eller Roo-mål by Samuel Columbus[2]	1675	20,260
Prosastycken på svenska by Johan Runius[2]	1710	30,200
Handbooks on agriculture[5]	1727	90,767
Total	**1529–1732**	**>634,216**

[1] Provided by Dramawebben.
[2] Provided by Fornsvenska textbanken.
[3] Provided by Litteraturbanken.
[4] Provided by Project Runeberg.
[5] Provided by Språkbanken Text.

Table 8: Early Modern Swedish secular prose.

3.3.3 Diaries and personal stories

For the Early Modern period, we observe the emergence of diaries and personal stories; a genre that is not always easy to differentiate from the genre of secular prose. It could however be useful to distinguish this specific text type from the more general secular prose genre, since personal stories are often written in a more informal, spoken-like style.

The *Gender and Work* project (see further Section 3.3.8) provides access to a diary written by a *länsman* 'sheriff'/'policeman' in 1671, containing approximately 6,000 words. This diary has been OCR-scanned, without manual post-correction, and according to the researchers in the project, the OCR quality is rather low. Furthermore, *Fornsvenska textbanken* provides access to several diaries and personal stories, as presented in Table 9.

Title	**Time Period**	**#Words**
Notes by Carl Carlsson Gyllenhielm[1]	1640	53,020
Beskrivning över min vandringstid by Agneta Horn[1]	1657	40,460
Stratonice by Urban Hiärne (autobiography)[1]	1665	11,330
Nils Mattson Kiöpings resor (travelogue)[1]	1674	32,700
Diary by policeman[2]	1671	6,000
Diary by Haqvin Spegel[1]	1680	32,720
Total	**1640–1680**	**176,230**

[1] Provided by Fornsvenska textbanken.
[2] Provided by the Gender and Work project.

Table 9: Early Modern Swedish diaries and personal stories.

3.3.4 Song texts

Project Runeberg provides access to a 64-page book of songs from 1682: *Fyratijo små Wijsor/ til Swänska Språketz öfningh*, scanned by *Litteraturbanken* (The Swedish Literature Bank), containing approximately 6,356 words. This figure is however highly approximate, due to poor OCR quality.

3.3.5 Periodicals

In the Early Modern period, periodicals start to emerge, such as *Hermes Gothicus* (1624) and *Dædalius Hyperboreus* (1716, approximately 65,620 words), provided by *Project Runeberg*.

3.3.6 Letters and charters

Table 10 lists digitally available Swedish letters from the Early Modern time period. A large collection of letters from this era is *Bref och Skrifvelser af och till Carl von Linné* (letters by and to Carl Linnaeus), a resource that is freely available through *Projekt Runeberg*.[30] The collection includes about 5,000 letters to public authorities, to *Kungliga Vetenskapssocieteten i Uppsala* (Royal Society of Sciences, Uppsala), to *Kungliga vetenskapsakademien* (Royal Swedish Academy of Sciences), to and from Swedish people, and to and from non-Swedish people (in Latin). The book containing these letters comprises approximately 112,895 words, but this also includes footnotes by

[30] http://runeberg.org/linnebref/

the editors of the book. Apart from representing the letter genre, these texts also include a substantial amount of 18th century scientific language.

Furthermore, *Jämtlands läns fornskriftsällskap* has transcribed *Härjedalsbrev*, a collection of approximately 170 letters and diplomas concerning the Swedish county Härjedalen, written in the time period 1531–1645, with a total of approximately 64,436 words. There are also a couple of Early Modern letters available through *Fornsvenska textbanken*, as well as letters with news and gossips from courtier Johan Ekeblad to his brother Claes Ekeblad, written in the time period 1639–1655, and provided by *Språkbanken Text*.

Title	**Time Period**	**#Words**
Härjedalsbrev[2]	1531–1645	64,436
Letters by Anna Vasa (Swedish princess)[1]	1591–1612	8,060
Letters from Johan Ekeblad[4]	1639–1655	52,253
Letters by Jon Stålhammar[1]	1700–1708	44,780
Letters by and to Carl Linnaeus[3]	1700s	112,895
Total	**1531–1708**	**282,424**

[1] Provided by Fornsvenska textbanken.
[2] Provided by Jämtlands läns fornskriftsällskap.
[3] Provided by Project Runeberg.
[4] Provided by Språkbanken Text.

Table 10: Early Modern Swedish letters and charters.

3.3.7 Academic text

The Uppsala University library has digitised and made available under an open-access license protocols from the Academic Consistory of Uppsala University, from the time period 1624–1699, comprising a total of approximately 4,981,618 words.[31]

3.3.8 Court records

Court records are a common text source for the Early Modern Swedish period. Table 11 lists Early Modern court records texts provided by the *Gender and Work* project, the Gothenburg University Library (GUPEA), *Jämtlands läns fornskriftsällskap*, *Språkbanken Text*, *Svenska fornskriftsällskapet*, *Vetenskapssocieteten i Lund*, and Guno Haskå,[32] volunteer at *Demografisk Databas Södra Sverige* (Demographical Database for Southern Sweden).

The texts from the Gender and Work project are provided as plain text files,

[31] http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-128732
[32] http://haska.se/

whereas the texts from GUPEA, Jämtlands läns fornskriftsällskap and Vetenskapssocieteten i Lund are in a readable PDF format. The texts from Guno Haskå are provided in Microsoft Word format, and the text from Språkbanken Text is in XML format, whereas the document from Svenska fornskriftsällskapet is currently being digitised (unknown format).

Title	**Time Period**	**#Words**
Enköpings stads tänkeböcker[5]	1540–1595	unknown
Stockholms stads tänkeböcker[2]	1544–1591	1,043,466
Östra Härads i Njudung[1]	1602–1605	34,956
Linköping rådhusrätt och magistrat[1]	1609–1616	34,330
Vendels socken[1]	1615–1645	59,948
Kristianstads rådstugubok[6]	1616–1637	195,103
Stockholms stads tänkebok[4]	1626	121,366
Per Larssons dombok[1]	1638	11,439
Alsens tingslags domböcker[3]	1649–1679	41,674
Hammerdals tingslags domböcker[3]	1649–1690	98,244
Lits tingslags domböcker[3]	1649–1690	111,395
Offerdals tingslags domböcker[3]	1649–1690	90,622
Ragunda tingslags domböcker[3]	1649–1690	119,318
Svegs tingslags domböcker[3]	1649–1690	102,159
Undersåkers tingslags domböcker[3]	1649–1690	101,724
Revsunds tingslags domböcker[3]	1649–1700	205,794
Skånska generalguvernementet[7]	1675–1719	unknown
Ekenäs[1,8]	1678–1695	74,121
Norra Åsbo häradsrätt[6]	1680–1681	182,457
Norra Åsbo häradsrätt[7]	1686–1695	6,991
Hammerdals tingslags domböcker[3]	1691–1700	109,438
Offerdals tingslags domböcker[3]	1691-1700	129,390
Undersåkers tingslags domboksprotokoll[3]	1691–1700	105,004
Linköping rådhusrätt och magistrat[1]	1709	70,095
Total	**1540–1719**	**>3,049,034**

[1] Provided by the Gender and Work project.

[2] Provided by GUPEA.

[3] Provided by Jämtlands läns fornskriftsällskap.

[4] Provided by Språkbanken Text.

[5] Provided by Svenska fornskriftsällskapet.

[6] Provided by Vetenskapssocieteten i Lund.

[7] Provided by Guno Haskå.

[8] Transcribed by professor Harry Lönnroth (https://www.jyu.fi/hytk/fi/laitokset/kivi/henkilosto/henkilosto/lonnroth-harry)

Table 11: Early Modern Swedish court records texts.

3.3.9 Laws and regulations

Table 12 lists laws and regulations related to the church, provided in a transcribed plain text format by the Gender and Work project.

Title	Time Period	#Words
Westerås Recess (parliamentary resolution)	1527	12,193
Then Swenska Kyrkoordningen (Swedish Church Order)	1571	49,043
Upsala möte (decision from meeting in Uppsala)	1593	26,969
Kyrkolag (Church Law)	1686	25,799
Total	**1527–1686**	**114,004**

Table 12: Early Modern Swedish laws and regulations available through the *Gender and Work* project.

3.3.10 Governmental texts

Documents produced in the Swedish Riksdag during the time periods 1521–1617 (early parliamentary-like meetings), 1618–1866 (“Ståndsriksdagen”) and 1867–1970 (“Tvåkammarriksdagen”) have been digitised by *Riksdagsförvaltningen* (the Swedish Parliament administration)[33] and made available via *Kungliga biblioteket* (National Library of Sweden),[34] under a Creative Commons CC0 1.0 licence.[35]

For the Early Modern period, these documents include parliamentary acts from early governmental and governmental-like meetings from the period 1521–1617 and from 1719–1734, as well as parliamentary acts, protocols and registers from Ståndsriksdagen (the governmental structure constituted by nobles, clergies, burghers and peasants) from the time period 1627–1734. The documents are provided as readable PDF files, as well as in XML and plain text format. The total number of words for the Early Modern period amounts to 10,945,864.

3.3.11 Accounts and registers

Table 13 lists accounts and registers currently provided by *Vetenskapssocieteten i Lund* (The Science Society in Lund), as part of their publication series *Skånsk senmedeltid och renässans* (‘Late Medieval and Renaissance time in Skåne’),[36] as well as a transcription of church records from the Parish of Osby, written in 1647–1690, containing approximately 124,487 words, provided by

[33] https://www.riksdagen.se/sv/sa-funkar-riksdagen/riksdagsforvaltningen/

[34] https://data.kb.se/datasets/2017/09/riksdagstryck/

[35] https://creativecommons.org/publicdomain/zero/1.0/

[36] https://www.ht.lu.se/serie/85/

Demografisk Databas Södra Sverige (Demographical Database for Southern Sweden, DDSS).[37]

Title	Time Period	#Words
Revenue book of the monastery of Dalby[2]	1530–1531	18,753
Real estate registers, Cathedral Chapters of Lund[2]	1570–1650	70,366
Real estate registers (“jordebok”), Helsingborg[2]	1583–84	32,569
Residential registration for additional taxes, Helsingborg[2]	1584	26,443
Osby church records[1]	1647–1690	124,487
Decimantboken, real estate register, Southern Sweden[2]	1651	136,722
Total	**1530–1690**	**409,340**

[1] Provided by Demografisk Databas Södra Sverige.
[2] Provided by Vetenskapssocieteten i Lund.

Table 13: Early Modern Swedish accounts and registers.

In addition to the data presented in Table 13, the DDSS webpage also enables database search in birth and baptism records, marriage records, death and burial records, migration records, registers of ships, mercantile marine office records, and in a list of nicknames used in the parish of Örkened. Furthermore, they provide access to a list of all first names and surnames in the 961,959 registered births in the database (from the 1600s to the 1800s), with the modern spelling of the name connected to the different spelling variants of that name occurring in the database. For example, the name *Abraham* occurs 666 times in the database, with the following spellings: Aberaham (1), Abereham (1), Abrah (2), Abraham (606), Abrahamn (2), Abram (52), Abreham (1), and Araham (1).

3.3.12 Maps

Riksarkivet (Swedish National Archives) has transcribed approximately 18,000 maps from the 17th century, mainly from the 1640s, see further https://riksarkivet.se/geometriska. The maps from this time period contain a considerable amount of text, due to descriptions and explanations being part of the maps. At the time of writing, we do not have a total word count for all the 17th century maps, but we know that the 6,783 maps from the later part of the century (1680 onwards) contain 1,463,660 words. The maps are currently in HTML format, but are in the process of being converted to a TEI-compatible XML format. The text in these maps is searchable using the following url: http://jordebok.ra.se/v2sok.php.

[37] http://ddss.nu/

3.3.13 Text fragments

In addition to fulltext diaries, court records and church documents provided by the *Gender and Work* project (see further Sections 3.3.3, 3.3.8 and 3.3.9 respectively), the Gender and Work database also contains fragments of texts from several genres. Since the purpose of the Gender and Work database is to store information on what people did for a living in historical times, these text fragments are paragraphs of text containing this kind of information. The typical length of such a fragment is a couple of sentences. At the time of writing, the Gender and Work database contains 537,134 words from five different genres, where about two thirds of the material are from the 1600s, approximately a fourth from the 1700s, and about 10 percent from the 1500s:[38]

- court records (449,073 words)
- accounts (37,383 words)
- diaries (23,935 words)
- petitions (18,876 words)
- listings (7,867 words)

3.3.14 Early Modern Swedish: Summary

Figure 3 shows the estimated number of words per genre for the Early Modern resources presented in this report. As seen from the figure, we can distinguish twelve different text genres for the Early Modern Swedish period: (i) religious texts; (ii) secular prose; (iii) diaries and personal stories; (iv) song texts; (v) periodicals; (vi) letters and charters; (vii) academic text; (viii) court records; (ix) laws and regulations; (x) governmental texts; (xi) accounts and registers; and (xii) maps.

Comparing the textual resources presented for the Early Modern period to the resources from the Old Swedish period (see further Section 3.2), we see the emerge of six new genres: governmental texts, academic text, maps, diaries and personal stories, periodicals and song texts. In fact, the largest category by far is governmental texts, thanks to the considerable amount of parliamentary acts, protocols and registers (10,945,864 words in total) digitised by the Swedish Parliament administration and made available via the National Library of Sweden (see further Section 3.3.10). Furthermore, academic text constitutes the second largest genre in the Early Modern texts presented in this report (with a total of 4,981,618 words).

[38] https://gaw.hist.uu.se/vad-kan-jag-hitta-i-gaw/kallorna-i-gaw/, retrieved 28-03-2019

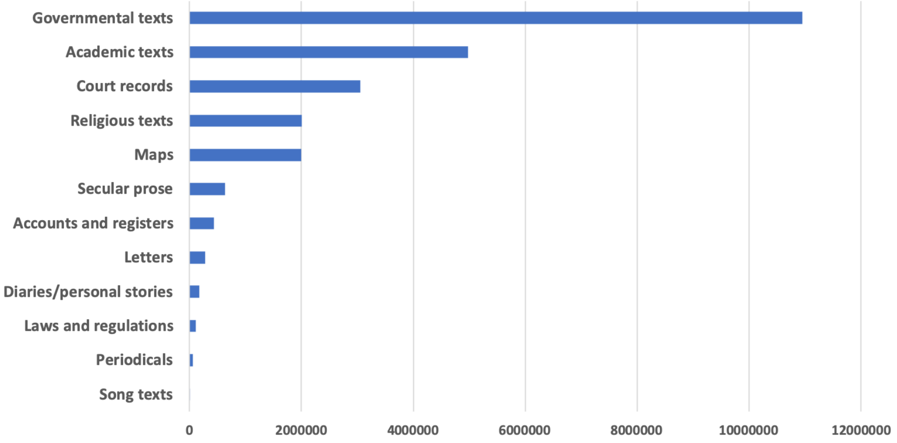

Figure 3: Approximate number of words per genre in the Early Modern Swedish resources.

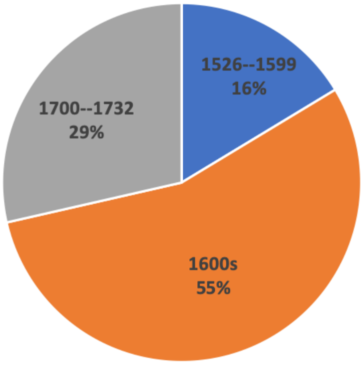

Figure 4: Approximate number of words per century for the Early Modern Swedish resources.

Similar to the Old Swedish period, we also see a substantial amount of court records text from the Early Modern period (3,042,043 words in total). Religious texts are however not as common as in the Old Swedish period. The religious texts from the Early Modern period are mainly biblical texts, whereas the Old Swedish period is characterized by religious legends and Christian edifying stories, in addition to texts from the bible. It should also be mentioned that the figure for the genre of secular prose is quite misleading, since the texts provided by Project Runeberg are not part of the estimate, due to downloading difficulties, prohibiting the calculation of word counts for these 32 books.

Figure 4 shows the distribution of words per century, for the Early Modern time period. We can see from the figure that the 16th century is somewhat underrepresented, since this period makes up approximately 36% of the total Early Modern period, but is represented by only 16% of the text material. In contrast, the 17th century makes up approximately 48% of the whole period, but holds about 55% of the text material, and the 18th century makes up approximately 16% of the Early Modern period, and is represented by about 29% of the total texts.

3.4 Late Modern Swedish (1732–1900)

The textual resources presented for the Late Modern Swedish period could be divided into the following ten genres: (i) religious texts; (ii) secular prose; (iii) diaries and personal stories; (iv) periodicals; (v) newspaper text; (vi) letters; (vii) academic and scientific text; (viii) court records; (ix) laws and regulations; and (x) governmental texts.

3.4.1 Religious texts

Table 14 lists Late Modern Swedish religious texts provided by *Fornsvenska textbanken* and *Språkbanken Text*.

Title	**Time Period**	**#Words**
Mag. Joh. Qvirsfelds himmelska örtagårds-sällskap[1]	1758	6,783
En sann christendom by Johann Arndt[1]	1855	393
The Bible[2]	1873	811,321
En sann christendom by Johann Arndt[1]	1891	2,357
Total	**1758–1891**	**820,854**

[1] Provided by Fornsvenska textbanken.
[2] Provided by Språkbanken Text.

Table 14: Late Modern Swedish religious texts.

3.4.2 Secular prose

For the Late Modern period, there are several works categorised as secular prose. As seen from Table 15, *Dramawebben*, *Litteraturbanken* and *Project Gutenberg* all give access to a number of books and plays in digitised format, where the documents from Dramawebben are provided in readable PDF format, while the texts from Litteraturbanken are downloadable as ePub files, and the documents from Project Gutenberg are in a plain text format. Since Litteraturbanken has had a special focus on digitising the collected works by

August Strindberg (1849–1912), Carl Jonas Love Almqvist (1793–1866) and Selma Lagerlöf (1858–1940), works written by these authors during the Late Modern period are given as separate items in the table. Also note that there is an overlap in the texts provided by these three organisations, meaning that the same text may occur in two or three of the text providers' collections, yielding slightly overlapping word counts.

Title	**Time Period**	**#Words**
Collected works by Carl Michael Bellman[6]	1700s	452,030
Miscellaneous books[4]	1700s–1899	13,315,768
Miscellaneous books[5]	1700s–1899	4,373,015
Miscellaneous plays[1]	1739–1899	1,600,620
Swenska sprätthöken (comedy drama)[2]	1743	32,400
Beskrifning öfwer Sweriges Lapmarker[2]	1747	17,990
Collected works by Carl Jonas Love Almqvist[4]	early 1800s	3,904,953
Swedish fiction (sentence order scrambled)[6]	1800–1900	16,272,174
56 Swedish novels (sentence order scrambled)[6]	1830–1942	4,347,047
Collected works by August Strindberg[4]	1879–1912	6,680,126
Novels by Selma Lagerlöf[4]	1891–1899	325,224
Tales and proverbs in dialect[3]	1890s	1,700 pages
Total	**1700s–1942**	**>51,321,347**

[1] Provided by Dramawebben.
[2] Provided by Fornsvenska textbanken.
[3] Provided by Institutet för språk och folkminnen.
[4] Provided by Litteraturbanken.
[5] Provided by Project Gutenberg.
[6] Provided by Språkbanken Text.

Table 15: Late Modern Swedish secular prose.

In addition to the books and plays in Dramawebben, Litteraturbanken and Project Gutenberg, *Språkbanken Text* provides access to the collected works of Carl Michael Bellman written in the 1700s, as well as a number of novels written by different authors during the 1800s and early 1900s. Due to copyright issues, the order of the sentences in the latter novels is however randomised. Furthermore, *Fornsvenska textbanken* offers both a comedy drama (*Swenska sprätthöken*) and a description of the nature in the northern part of Sweden. Finally, Erik Magnusson Petzell at *Institutet för språk och folkminnen* (the Swedish Institute for Language and Folklore),[39] is currently working on using Transkribus[40] for making historical, hand-written, Swedish dialectal texts searchable via a SAMPA-inspired format (Speech Assessment Methods Phonetic Alphabet).[41] Currently, he has approximately 500 pages of automati-

[39] http://www.sprakochfolkminnen.se/om-oss/om-webbplatsen/andra-sprak-an-svenska/english.html
[40] https://transkribus.eu/Transkribus/
[41] https://www.phon.ucl.ac.uk/home/sampa/

cally transcribed text, and 1,000 pages still to be transcribed, containing a collection of tales from the end of the 1890s. In addition, there are 200 pages of transcribed text containing proverbs and sayings from the county of Halland, covering the same time period.

Title	Time Period	#Words
Nils Mattson Kiöpings resor[1]	1740	28,478
Story from Ramsberg[3]	1748	351
Story from Ramsberg[3]	1759	409
Diary written by priest Muncktell[2]	1814–1829	590,000
Total	**1740–1829**	**619,238**

[1] Provided by Fornsvenska textbanken.
[2] Provided by the Gender and Work project.
[3] Provided by Riksarkivet.

Table 16: Late Modern Swedish diaries and personal stories.

3.4.3 Diaries and personal stories

Late Modern Swedish texts categorised as diaries or personal stories are presented in Table 16. These texts include a diary written by a priest named Muncktell during the time period 1814–1829, containing approximately 590,000 words, provided by the *Gender and Work* project in an OCR-scanned plain text format. According to the researchers in the Gender and Work project, this diary has quite good OCR quality.

In addition, Linda Oja at *Riksarkivet* (Swedish National Archives) has provided two texts with stories about things that happened in the town of Ramsberg in 1748 and 1759. These are manually transcribed and provided in Microsoft Word format. Furthermore, *Fornsvenska textbanken* provides access to the travelogue *Nils Mattson Kiöpings resor* in the edition from 1740.

Title	Time Period	#Words
Then Swänska Argus[2]	1732–1734	213,160
Miscellaneous periodicals from Runeberg[3]	1810–1933	5,358,564
Boijes magasin[1]	1818–1824	125,773
Ur Dagens Krönika[3]	1881–1890	1,995,149
Dagny[3]	1886-1913	8,124,256
Idun[3]	1887–1917	44,944,172
Total	**1732–1933**	**60,761,074**

[1] Provided by Alvin.
[2] Provided by Fornsvenska textbanken.
[3] Provided by Språkbanken Text.

Table 17: Late Modern Swedish periodicals.

3.4.4 Periodicals

Periodicals from the Late Modern period provided by *Alvin* (in readable PDF format), *Fornsvenska textbanken* (as RTF files), and *Språkbanken Text* (in XML format) are listed in Table 17.

3.4.5 Newspaper text

In the Late Modern period, newspapers emerged as a new genre. Table 18 lists the newspaper texts present in the current version of the *Kubhist* corpus (Kungliga bibliotekets historiska tidningar), a collection of Swedish newspapers covering the time period 1740–1926 (Adesam, Dannélls and Tahmasebi 2019). The current version of the corpus comprises approximately 1.1 billion tokens, whereas a second, soon-to-be-released, version will contain about 5.5 billion tokens.[42]

Title	Time Period	#Words
Götheborgs weckolista	1740s–1750s	606,495
Stockholmsposten	1770s–1830s	41,689,548
Post- och Inrikes Tidningar	1770s–1860s	195,490,243
Fahlu weckoblad	1780s–1820s	3,408,225
Aftonbladet	1830s–1860s	200,328,267
Jönköpingsbladet	1840s–1870s	38,795,168
Tidning för Wernersborgs stad och län	1840s–1890s	75,589,323
Folkets röst	1850s–1860s	24,604,568
Blekingsposten	1850s–1880s	39,547,561
Dalpilen	1850s–1920s	114,725,162
Gotlands tidning	1860s–1880s	17,281,641
Faluposten	1860s–1890s	17,449,706
Kalmar	1860s–1910s	203,511,038
Wermlands läns tidning	1870s	13,093,416
Bollnäs tidning	1870s–1880s	3,083,072
Lindesbergs allehanda	1870s–1880s	2,493,430
Wernamo tidning	1870s–1880s	5,402,616
Göteborgs weckoblad	1870s–1890s	15,862,339
Norra Skåne	1880s–1890s	31,089,053
Östergötlands veckoblad	1880s–1890s	12,591,454
Östgötaposten	1890s–1910s	20,802,723
Total	**1740s–1920s**	**1,077,445,048**

Table 18: Late Modern Swedish newspaper text available through the *Kubhist* corpus.

[42] Due to IPR restrictions, the new version of Kubhist will only contain material up until 1904.

The texts have been OCRed by the National Library of Sweden, without manual post-correction, and automatically annotated with part-of-speech and links to lexicon entries. Adesam, Dannélls and Tahmasebi (2019) conclude that the Kubhist corpus suffers from low OCR quality, but also give directions for measures to improve the OCR quality, and consequently the succeeding linguistic annotation as well, in order for the corpus to be more useful for digital humanities research.

3.4.6 Letters

Concerning Late Modern letters, *Litteraturbanken* has digitised a collection of letters written by author August Strindberg between 1858 and 1912. This collection of letters amounts to a total of 1,507,958 words, and is available in XML format via *Språkbanken Text*.

3.4.7 Academic and scientific text

The Uppsala University Library has ongoing work on digitising all Swedish dissertations from the time period 1602–1855, making them available as OCR-scanned PDF files. It is estimated that the library holds approximately 5,000 dissertations from this time period, out of which about 1,100 are written in Swedish (from 1730 onwards), and the rest are written in Latin. There are also cases of code-switching, where parts of a dissertation are written in Swedish and other parts of the same dissertation are written in Latin.

Since the digitisation project is still ongoing work, the total number of words in the dissertations is still unknown. Sample dissertations show a variation between around 2,000 words for an 1800s dissertation and 7,800 words for a 1700s dissertation. The majority of the dissertations are from the 1800s, where the lower word count is to be expected.

Furthermore, *Språkbanken Text* provides access to a collection of documents from *Kungliga vetenskapsakademien* (Royal Swedish Academy of Sciences) from the time period 1740–1778, with a total of 27,878 words.

3.4.8 Court records

Similar to the earlier time periods in the history of the Swedish language, court records are a rather common text source also for the Late Modern Swedish period. Table 19 lists Late Modern court records provided by the *Gender and Work* project (in plain text format), *Jämtlands läns fornskriftsällskap* (in readable PDF format), Linda Oja at *Riksarkivet* (the Swedish National Archives) and Guno Haskå,[43] volunteer at *Demografisk Databas Södra Sverige* (in Microsoft Word format).

Title	Time Period	#Words
Court records from Perstorp/Oderljunga[4]	1691–1762	unknown
Norra Åsbo häradsrätt[4]	1707–1716	7,038
Orsa Kyrkoarkiv[1]	1727–1794	92,834
Stora Malm[1]	1728–1741	46,453
Nås tingslags häradsrätt[3]	1734	1,050
Vendels socken[1]	1736–1737	55,780
Stora Malm[1]	1742–1760	59,402
Rödöns sockenstämmoprotokoll[2]	1742–1862	66,487
Skogsavvittringen i Jämtland[2]	1755–1758	132,810
Skogsavvittringen i Jämtland[2]	1759–1779	160,707
Stora Malm[1]	1761–1783	53,194
Skellefteå höstting[1]	1771	18,708
Skellefteå vårting[1]	1771	20,815
Stora Malm[1]	1784–1795	47,075
Stora Malm[1]	1796–1812	35,060
Court records from Perstorp/Oderljunga[4]	1808–1850	unknown
Häggenås sockenstämoprotokoll[2]	1821–1862	81,448
Total	**1691–1862**	**>878,861**

[1] Provided by the Gender and Work project.
[2] Provided by Jämtlands läns fornskriftsällskap.
[3] Provided by Riksarkivet.
[4] Provided by Guno Haskå.

Table 19: Late Modern Swedish court records texts.

3.4.9 Laws and regulations

Table 20 presents Late Modern Swedish laws and regulations provided by *Språkbanken Text* (in XML format).

Title	Time Period	#Words
Preparatory work for the law of 1734	1686–1734	1,603,126
1734 års lag	1734	98,120
Författningssamling (from church archives in Låssa)	1800	388,108
Regeringsformen (Instrument of Government)	1809	58,330
Total	**1686–1809**	**2,147,684**

Table 20: Late Modern Swedish laws and regulations available through *Språkbanken Text*.

[43] http://haska.se/

3.4.10 Governmental texts

Documents produced in the Swedish Riksdag during the time periods 1521–1617 (early parliamentary-like meetings), 1618–1866 (“Ståndsriksdagen”) and 1867–1970 (“Tvåkammarriksdagen”) have been digitised by *Riksdagsförvaltningen* (the Swedish Parliament administration)[44] and made available via *Kungliga biblioteket* (National Library of Sweden, KB),[45] under a Creative Commons CC0 1.0 licence.[46]

For the Late Modern period, these documents include parliamentary acts, protocols and registers from Ståndsriksdagen (the governmental structure constituted by nobles, clergies, burghers and peasants) from the time period 1738–1866, as well as protocols (with debates from the chamber), bills, proposals, reports, motions, committee reports, parliamentary letters and registers from Tvåkammarriksdagen (bicameral Riksdag). The documents are provided as readable PDF files, as well as in XML and plain text format. The total number of words for the Late Modern period amounts to 319,135,806.

3.4.11 Late Modern Swedish: Summary

Figure 5 shows the estimated number of words per genre for the Late Modern resources presented in this report. As seen from the figure, we can distinguish ten different text genres for the Late Modern Swedish period: (i) religious texts; (ii) secular prose; (iii) diaries and personal stories; (iv) periodicals; (v) newspaper text; (vi) letters; (vii) academic and scientific text; (viii) court records; (ix) laws and regulations; and (x) governmental texts.

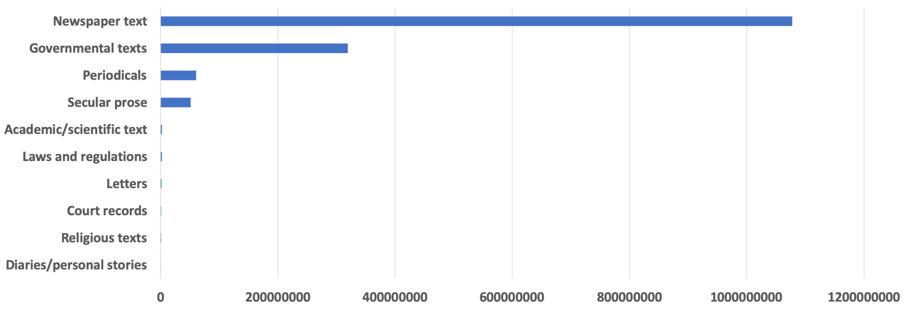

Figure 5: Approximate number of words per genre in the Late Modern Swedish resources.

[44] https://www.riksdagen.se/sv/sa-funkar-riksdagen/riksdagsforvaltningen/

[45] https://www.kb.se/kb-in-english.html

[46] https://creativecommons.org/publicdomain/zero/1.0/

Due to the large *Kubhist* corpus (1,077,445,048 words in total) of Late Modern Swedish newspaper text (see further Section 3.4.5) and the considerable amount of parliamentary acts, protocols and registers (237,206,770 words in total) digitised by the Swedish Parliament administration and made available via the National Library of Sweden (see further Section 3.3.10), the genres of newspaper text and governmental texts clearly dominate the contents of the textual resources available for the Late Modern Swedish time period. We also see a substantial amount of texts from the genres of periodicals and secular prose, whereas diaries/personal stories and religious text are the smallest categories for this era.

Figure 6 shows the distribution of words per 50-year period for the Late Modern time period. As is obvious from the figure, the closer in time we get, the more resources are available. Even though the 18th century period is longer than the other two periods, only 2% of the text material presented for the Late Modern period are written in the 1700s. It could be expected that there would be more texts available for the 19th century than for the 18th century, but maybe not on this scale. Part of the explanation for this lies in the fact that the two (by far) largest corpora (the Kubhist newspaper corpus and the governmental texts from the Swedish Parliament administration) both contain substantially more material from the 1800s than from the 1700s.

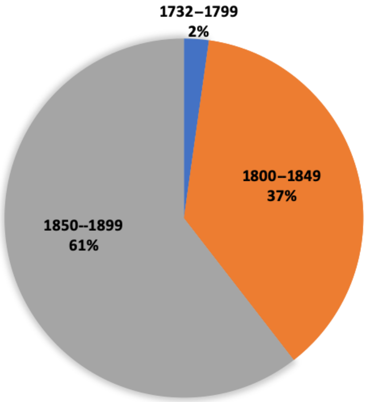

Figure 6: Approximate number of words per 50-year period for the Late Modern Swedish resources.

3.5 CONTEMPORARY SWEDISH (1900–)

The textual resources presented for the Contemporary Swedish period could be divided into the following fourteen genres: (i) religious texts; (ii) secular prose; (iii) diaries and personal stories; (iv) song texts; (v) periodicals; (vi) newspaper text; (vii) letters and charters; (viii) essays and school-related texts; (ix) academic and scientific text; (x) court records; (xi) laws and regulations; (xii) governmental texts; (xiii) accounts and registers; and (xiv) social media text and Wikipedia.

Title	Time Period	#Words
The Bible[2]	1917	894,720
En sann christendom by Johann Arndt[1]	1928	2,398
Psalmboken (Book of Hymns)[2]	1937	163,574
Total	**1917–1937**	**1,060,692**

[1] Provided by Fornsvenska textbanken.
[2] Provided by Språkbanken Text.

Table 21: Contemporary Swedish religious texts.

3.5.1 Religious texts

Contemporary Swedish religious texts provided by *Fornsvenska textbanken* and *Språkbanken Text* are presented in Table 21.

Title	Time Period	#Words
60 miscellaneous plays[1]	1900–1937	742,247
69 novels published by Albert Bonniers förlag[4]	1976–1977	6,578,675
60 novels published by Albert Bonniers förlag[4]	1980–1981	4,304,271
LäSBarT (easy-to-read books)[4]	1994–2008	1,129,083
23 books published by Norstedts[4]	1999	2,533,209
573 miscellaneous books[2]	1900s	19,435,810
83 miscellaneous book[3]	1900s	3,654,708
Total	**1900–2008**	**38,378,003**

[1] Provided by Dramawebben.
[2] Provided by Litteraturbanken.
[3] Provided by Project Gutenberg.
[4] Provided by Språkbanken Text (scrambled sentence order).

Table 22: Contemporary Swedish secular prose.

3.5.2 Secular prose

Table 22 presents Contemporary Swedish secular prose available through *Dramawebben* (in readable PDF format), *Litteraturbanken* (as downloadable ePub files), *Project Gutenberg* (in plain text format) and *Språkbanken Text* (in XML format). It should be noted that the order of the sentences in the texts from Språkbanken Text are scrambled, due to copyright issues. Similarly, around a hundred of the books provided by Litteraturbanken may not be distributed further, also due to copyright issues.

3.5.3 Diaries and personal stories

Jämtlands läns fornskriftsällskap has transcribed and made accessible (in readable PDF format) *Gertrud Daléns berättelser om jul, julkusar och faster Lotta* from 1962, containing approximately 13,177 words, where Gertrud Dalén writes about her childhood.

3.5.4 Song texts

Folklivsarkivet (the Folklife Archives) has a collection of 150 songs from the 1890s and onwards, see further https://www.folklivsarkivet.lu.se/en/the-scania-music-collections/song-collection/. 69 of these are hand-written song texts or song books that have been manually transcribed, and may be ordered from Folklivsarkivet in a readable PDF format, containing approximately 1,032,700 words. The texts may be distributed further, provided that ID information on the cover sheets is removed before publication.

3.5.5 Periodicals

Språkbanken Text provides access to a considerable amount of periodicals from the time period 1891–2010, covering a wide range of hobbies and interests, as listed in Table 23.

Title	**Time Period**	**#Words**
Svensk Tidskrift	1891–1940	7,202,567
Folkbiblioteksbladet	1904–1908	398,330
Morgonbris	1904–1924	3,551,943
Tiden	1909–1940	7,106,662
Rösträtt för kvinnor	1912–1919	2,202,776
Hertha	1914–1935	3,842,984
Biblioteksbladet	1916–1940	4,595,593
Kvinnornas tidning	1921–1925	5,468,918
Tidevarvet	1923–1936	6,813,909
Forskning & framsteg	1992–1996	743,831
Läkartidningen (sentence order scrambled)	1996–2006	21,042,021
Smittskydd (sentence order scrambled)	2002–2010	691,716
Total	**1891–2010**	**63,661,250**

Table 23: Contemporary Swedish periodicals available through *Språkbanken Text*.

3.5.6 Newspaper text

Contemporary Swedish newspaper texts provided by *Språkbanken Text* are presented in Table 24, where Press 65–98 contains news articles collected from the following papers: *Arbetet*, *Dagens Nyheter*, *Göteborgs Handels- och Sjöfartstidning*, *Göteborgs-Posten*, *Stockholmstidningen*, *Svenska Dagbladet*, and *Sydsvenska Dagbladet*. Due to copyright issues, the order of the sentences has been scrambled in all contemporary newspaper texts listed in the table.

Title	**Time Period**	**#Words**
Yearbooks from *Svenska dagbladet*	1923–1958	1,528,935
Press 65	1965	1,000,669
Press 76	1976	1,348,122
Dagens nyheter	1987	2,832,875
Göteborgsposten	1994	21,331,715
Press 95	1995	7,671,700
Press 96	1996	6,516,030
Press 97	1997	13,703,279
Press 98	1998	10,740,849
Göteborgsposten	2001–2013	250,672,582
Texts from newspaper websites	2001–2013	271,806,921
8 sidor (easy-to-read news)	2002–2017	2,832,875
Total	**1923–2017**	**591,986,552**

Table 24: Contemporary Swedish newspaper texts available through *Språkbanken Text*, with the order of the sentences scrambled due to copyright issues.

3.5.7 Letters

For Contemporary Swedish, *Jämtlands läns fornskriftsällskap* has published a collection of private letters written by composer Wilhelm Peterson-Berger to his friend Helga Englund in 1902–1925, with a total of 17,466 words.

3.5.8 Academic and scientific text

Table 25 presents academic and scientific texts available for the Contemporary Swedish time period via *Språkbanken Text*, *Tekniska museet* and *Vetenskapssocieteten i Lund*.

The texts from *Språkbanken Text* are academic texts provided in XML format, from the fields of humanities and social sciences, for the time period 1996–2012.

The texts from *Tekniska museet* (the Swedish National Museum of Science and Technology) are digitisations made within the research project *Digital Models. Techno-historical collections, digital humanities & narratives of industrialisation*, funded by the Royal Swedish Academy of Letters, History and Antiquities, between 2016 and 2019.[47] This project is a collaboration between the Swedish National Museum of Science and Technology and the Digital Humanities Hub, HUMlab at Umeå University, with the following aims stated on the project website:

> The project aims to explore the potential of digital technologies to reframe Swedish industrialisation and its stories about society, people and environments. Material from the museum's collections selected for digitisation (and research) are all related to different phases of Swedish industrialisation. (http://digitalamodeller.se/in-english/, accessed 2019-04-11)

In connection with this project, the Swedish National Museum of Science and Technology are providing access to OCR-scanned versions of all editions of the museum yearbook, Daedalus (1931–2015), released under a CC-BY licence. The material comprises approximately 15,000 pages of technical and industrial history, with a total of 4,099,101 words. The texts are provided both in a readable PDF format, and in XML format.

Title	**Time Period**	**#Words**
Daedalus museum yearbook[2]	1931–2015	4,099,101
Humanities texts[1]	1996–2012	14,454,573
Social science texts[1]	1997–2012	10,855,954
Yearbooks of humanities, theology & social science[3]	1999–2016	888,392
Total	**1931–2016**	**30,298,020**

[1] Provided by Språkbanken Text.
[2] Provided by Tekniska museet.
[3] Provided by Vetenskapssocieteten i Lund.

Table 25: Contemporary Swedish academic and scientific text.

Likewise, *Vetenskapssocieteten i Lund* (The Science Society in Lund) has since 1920 published yearbooks containing scientific papers and announcements related to humanities, theology and social sciences. They are currently working on digitising these yearbooks, and at the time of writing, the books from 1999–2001 and from 2003–2016 have been OCR-scanned and published open access in a readable PDF format,[48] with a total of approximately 888,392 words.

3.5.9 Court records

Concerning the genre of court records for the Contemporary Swedish time period, *Språkbanken Text* provides access to a collection of verdicts from the time period 1981–2009, in XML format, with a total of 32,206,334 words.

[47] http://digitalamodeller.se/in-english/
[48] https://www.ht.lu.se/serie/86/

3.5.10 Laws and regulations

Concerning laws and regulations, *Språkbanken Text* provides access to *Svensk författningssamling* (Swedish Code of Statutes) in XML-format, with a total of 8,058,400 words, produced during the time period 1880–2012.

3.5.11 Governmental texts

According to the Swedish principle of public access to official documents, all citizens have the right to access official documents. In line with this principle, the Swedish Riksdag is providing open access to its databases of calendar information planning, documents (such as committee reports, private members' motions, laws), information about members, results of votes and members' speeches in the Chamber.[49] *Språkbanken Text* provides access to these data in an XML format, covering the time period 1900–2016, with a total of 1,531,417,818 words. In addition, Språkbanken Text also provides access to Swedish party programs and election manifestos, the Swedish part of the European Parliament Proceedings Parallel Corpus,[50] and documents from authorities other than the Government ("förvaltningsmyndigheter"), as listed in Table 26.

Title	Time Period	#Words
Party programs and election manifestos	1887–2010	821,860
The Riksdag's open data	1900–2016	1,531,417,818
European Parliament Proceedings (Swedish part)	1996–2011	33,406,922
Förvaltningsmyndigheter (sentence order scrambled)	2011–2015	51,366
Total	**1887–2016**	**1,565,697,966**

Table 26: Contemporary Swedish governmental texts from *Språkbanken Text*.

3.5.12 Essays and school-related texts

Språkbanken Text provides access to a collection of essays from 2012–2013, with a total of 51,972 tokens, and a collection of sample sentences from school tests in social-science subjects from 2009, with a total of 541,568 words. For both corpora, the order of the sentences has been scrambled, due to copyright issues.

3.5.13 Accounts and registers

The *REAL register* is a historical register over notes in folklore and ethnology, provided by *Folklivsarkivet* (the Folklife Archives).[51] In particular, the

[49] https://data.riksdagen.se/in-english/, accessed 07-05-2019

[50] http://www.statmt.org/europarl/

[51] https://www.folklivsarkivet.lu.se/en/

REAL register is a transcription of *Realkatalogen*, a catalogue that systematizes knowledge about traditional Swedish rural culture. The register includes information from the manuscript archive collections from 1913 to around 1960. Most of these documents are only available in a non-readable PDF-format, but there is ongoing work on transcribing them. Currently, accession numbers 1–619 are transcribed and available in readable PDF format, either via the website[52] or by ordering them from the Folklife Archives. The records transcribed so far contain approximately 969,019 words.

3.5.14 Social media text and Wikipedia

A new text type evolving from the late 1990s and onwards is the genre of social media and Wikipedia text. Table 27 presents social media and Wikipedia texts provided by *Språkbanken Text* for the time period 1998–2017. These include miscellaneous blog texts, texts from two web forums (*Familjeliv* and *Flashback*), and Swedish Wikipedia (extracted 2017).[53] For the social media texts (unlike the Wikipedia corpus), the order of the sentences has been scrambled, due to copyright issues.

The texts from Familjeliv are divided into 23 different categories, based on the discussion theme, with a majority of themes related to family life, such as adoption, economy and law, family life, hobbies, household, pets, body and soul, planning to get pregnant, being pregnant, being a parent etc. Likewise, the Flashback texts are divided into 15 categories based on the discussion theme, for example cars, drugs, lifestyle, food, politics, sports and society.

Title	**Time Period**	**#Words**
Blog text	1998–2017	615,658,549
Familjeliv (chat forum)	2003–2017	4,253,869,738
Flashback (chat forum)	2000–2017	3,123,988,369
Swedish Wikipedia	2017	370,211,509
Total	**1998–2017**	**8,363,728,165**

Table 27: Contemporary Swedish social media and Wikipedia texts available through *Språkbanken Text*.

3.5.15 Contemporary Swedish: Summary

Figure 7 shows the estimated number of words per genre for the Late Modern resources presented in this report. As seen from the figure, we can distinguish fourteen different text genres for the Late Modern Swedish period:

[52] https://www.folklivsarkivet.lu.se/samlingar/intervjuer-och-fragelistsvar/

[53] Språkbanken Text also makes available Swedish tweets from the period 2013–2019 (slightly over 2 billion tokens) through the Korp online corpus search interface, but the data are not available for download, and for this reason we do not list them in Table 27.

(i) religious texts; (ii) secular prose; (iii) diaries and personal stories; (iv) song texts; (v) periodicals; (vi) newspaper text; (vii) letters and charters; (viii) essays and school-related texts; (ix) academic and scientific text; (x) court records; (xi) laws and regulations; (xii) governmental texts; (xiii) accounts and registers; and (xiv) social media text and Wikipedia.

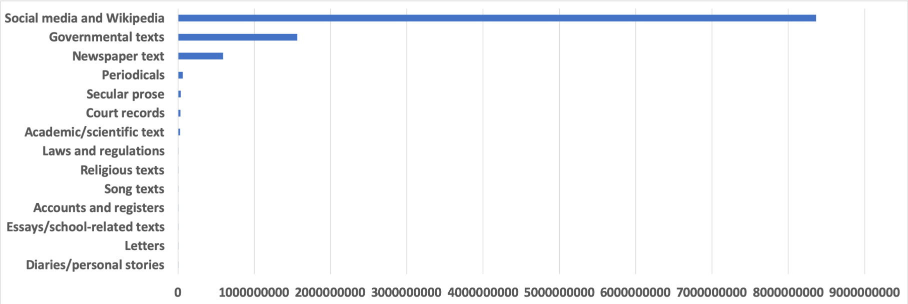

Figure 7: Approximate number of words per genre in the Contemporary Swedish resources.

The three clearly dominating genres are social media, governmental texts and newspaper text. Furthermore, for all these three genres, there is a substantial amount of text from the 2000s, as compared to the 1900s. As seen in Figure 8 (left), showing the distribution of words per 25-year period for the Contemporary Swedish time period, approximately 98% of the text material is from the 21st century.

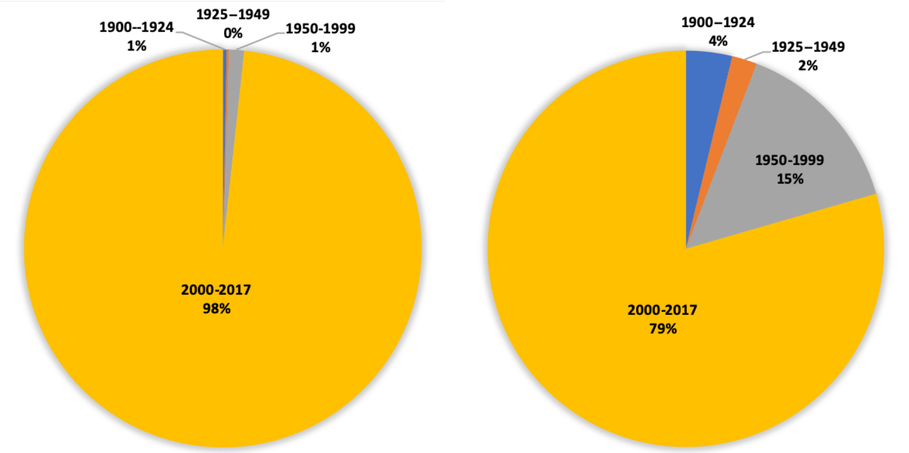

Figure 8: Approximate number of words per 25-year period for the Contemporary Swedish resources: total (left) and excluding social media text (right).

Figure 8 (right) shows the distribution of words per 25-year period for the Contemporary Swedish era, when the dominating social media texts are excluded. As seen from the figure, the 21st century texts are still clearly dominating, represented by 79% of the text material.

4 User questionnaire

In addition to the inventory of Swedish resources available for different time periods throughout the Swedish language development, we also wanted to get an idea of the needs and wishes of the primary target users for a Swedish diachronic corpus. We therefore sent out a questionnaire with eight questions to 15 members of our primary target group, i.e. historical linguists. The questions were the following:

1. Are there any research questions within your field where you think a Swedish diachronic corpus could be useful? If so, how?
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?
3. What time period do you think should be covered by a Swedish diachronic corpus?
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.? Motivate your answer.
5. Are there any particular text types/genres that you think should be included in the corpus (for example legal text, religious texts, fiction, newspaper text, etc.)?
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 (and how should then “Swedish” be distinguished from other Nordic varieties at certain points in time)? Should the corpus also include texts written in Swedish outside the borders of Sweden, which could then include for example texts in Finland Swedish or American Swedish? Or do you have any other alternative to how the term “Swedish” should be defined in this context?

Motivate your answer, and give examples to illustrate why you want to include or exclude certain texts, so that we can get a deeper understanding of the reasoning around these issues.

7. Are there any specific parametres that you would like to be able to search for in the corpus, for example words and phrases, names of persons or places, linguistic categories (such as part-of-speech, morphology, phrase category or syntactic function) or semantic features of any kind?
8. What metadata categories should be included in the corpus (year, author, genre, location, issue, etc.)?

4.1 QUESTIONNAIRE RESPONSES: SUMMARY

In this section following, we present the answers given by the 12 researchers who responded to the questionnaire.

4.1.1 Research questions for a diachronic corpus

[*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 was 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 variants (such as Swedish in Sweden as opposed to Fenno-Swedish), or (iii) for finding unpredictable patterns in a large and differentiated text material. It was also pointed out that the possibility to do a quick search in large corpora is a very positive development within humanities research, and that corpus studies are highly efficient, provided that it is possible to formulate precise enough search questions. One researcher mentioned that due to the lack of diachronic corpora for Swedish, it is often necessary for the researcher to compile his/her own corpus for different studies related to historical linguistics, and that it is then hard to find relevant texts, especially annotated texts.

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 lexicalisation 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.

In addition, it is desirable that the texts in the corpus should be both searchable and downloadable, and that it should be possible for the researcher to create a subcorpus of the particular texts that are of interest for the research question at hand.

4.1.2 Previously used corpora

[*Have you previously used an 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?*]

The participants in the questionnaire mention that they have experience of using the following historical/diachronic corpora in their research:

- *Fornsvenska textbanken* (Delsing 2002) and the corresponding *Studér Middelalder på Nettet* for Old Danish[54]
- the *Gender and Work* database[55] (Ågren et al. 2011)
- *Hamburg Corpus of Old Swedish with Syntactic Annotations (HaCOSSA)* (Höder 2011a)
- the Korp interface in *Språkbanken Text* (Borin, Forsberg and Roxendal 2012),[56] in particular for the Finland Swedish part and for newspaper text
- *Litteraturbanken* (The Swedish Literature Bank)[57]
- the *Kubhist* corpus (Adesam, Dannélls and Tahmasebi 2019)
- *Medieval Nordic Text Archive* (Menota)[58]

[54] https://dsl.dk/projekter/studer-middelalder-pa-nettet

[55] https://gaw.hist.uu.se/

[56] https://spraakbanken.gu.se/korp/?mode=all_hist#?lang=en

[57] https://litteraturbanken.se/

[58] http://www.menota.org/forside.xhtml

- the *Nordic Dialect Corpus* (NorDiaCorp)[59] (Johannessen et al. 2014)
- *Samnordisk runtextdatabas* (Scandinavian Runic-text Database)[60]
- *Scaldic Poetry of the Scandinavian Middle Ages*[61]
- newspaper texts from *Kungliga biblioteket* (National Library of Sweden)[62] and from the National Library of Finland[63]

The greatest advantage of diachronic and historical corpora in general is that they exist, since this has opened up for new research questions. One function that several researchers mention as a positive feature is the possibility to download texts to be able to process them on their own computer. Moreover, the possibility 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.

General disadvantages mentioned by the users are the lack of a common corpus format standard, the incompleteness of available corpora, and the insecurity about the quality of the transcription and the annotation. Some users point out that quality is often as important as quantity, and that the annotation needs to be close to perfect in order for the researcher to trust it. On the other hand, several participants in the questionnaire point out that many available corpora are not large enough for the research question at hand, and/or not representative enough in terms of text types and genres, translated and non-translated texts, literary and non-literary texts, texts written by people from different social backgrounds etc. There is also a need for better graphical user interfaces, such as the one in Korp or CQP (Evert 2019). One user promotes a corpus tool named *Conc*, which has the advantage that it shows the search word in three different windows simultaneously: (i) in a concordance, (ii) in an alphabetically sorted list with frequencies, and (iii) in a larger context (a few lines).

In some cases, it is perceived as worrying when there is no way to know how edited the text has been as compared to the original. It is also mentioned that it is often hard to search in corpora that have been OCR-scanned without manual

[59] Strictly speaking, NorDiaCorp is a diachronic/historical corpus only wrt the Norwegian part, which includes some subsequently added older material from 1950–1980, while the bulk of the data are from the 21st century.

[60] https://www.nordiska.uu.se/forskn/samnord.htm/

[61] https://skaldic.abdn.ac.uk/db.php

[62] https://www.kb.se/hitta-och-bestall/hitta-i-samlingarna/svenska-dagstidningar.html

[63] https://digi.kansalliskirjasto.fi/search?formats=NEWSPAPER

post-correction, due to bad OCR quality. One corpus specifically addressed concerning this problem is the *Kubhist* corpus.

Several participants in the questionnaire use the Korp interface for their research, and there are some suggestions for improvement. One user declares that even though the Korp interface contains large bodies of text to search in, the sample is not balanced, and there is a lack of clear information about what the different subcorpora contain. Furthermore, the metadata information is hard to understand and the search functions are tricky. Sometimes it is hard to get an overview of the results, if the corpora used as a basis for the search are large. One user suggests a function for creating a random sample of the hits for a search question in these cases, for example by showing every 10th hit, or only one hit from each subcorpus. Another drawback mentioned for the Korp interface, is that it is only possible to search in predefined text collections, whereas it would be desirable to be able to search in single texts as well, and in collections of text selected by the user.

Concerning *Fornsvenska textbanken*, one advantage mentioned is the possibility to download the texts. A disadvantage is that there is no possibility to search the corpus as a whole, only text by text. Furthermore, it is not possible to search for words in their standardized spelling.

Advantages specifically mentioned for *Samnordisk runtextdatabas* are (i) the possibility to perform search based on standardized word forms, (ii) the possibility to search based on different regions, (iii) the possibility for the user to compile his/her own sample of data to search within, (iv) that it provides translations, and (v) that it provides metadata. The texts would however profit from annotation, such as part-of-speech tagging.

4.1.3 Time periods covered

[*What time period do you think should be covered by a Swedish diachronic corpus?*]

The general answer to the question on what time period the Swedish diachronic corpus should cover 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* (not linguistically annotated, though), and that Runic Swedish is a bit too far from Swedish as we know it today. On the other hand, since Sweden has the most runic inscriptions in the world, there is also a considerable international interest in getting access to the runic inscriptions. Furthermore, almost all known inscriptions are collected in *Samnordisk runtextdatabas*, meaning that it will be fairly easy to collect the inscriptions and include them in the diachronic corpus.

It is also argued that since many researchers today use Korp, their research questions are governed by the material present in Korp, so the composition of texts in the diachronic corpus should preferably be discussed in a broad group of researchers, to include texts and time periods that are relevant to as many researchers as possible.

4.1.4 Division into time periods

[*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.? Motivate your answer.*]

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. Furthermore, the linguistically motivated time periods are perceived as somewhat artificial. It is also noted that political events in different parts of the Swedish language area may mean that the linguistically motivated time periods should be defined differently in for example Finland than in Sweden. In addition, if the user is allowed to create his/her own subcorpora based on any time interval, s/he could recreate the linguistically motivated time periods if needed, whereas the opposite is not true.

For predefined time periods based on a certain number of years, it is pointed out that for older stages of the language, fewer texts are available, and the dating is more uncertain, especially since the manuscript is often a younger copy of the original text. This means that the division into time periods might need to be more coarse for Old Swedish than for Contemporary Swedish, where a division into 10-year periods or 25-year periods is preferred by many researchers.

4.1.5 Text types covered

[*Are there any particular text types/genres that you think should be included in the corpus (for example legal text, religious texts, fiction, newspaper text, etc.)?*]

The general answer to the question on what text types/genres to include in the corpus 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 there isn’t one genre that is strongly overrepresented, for example because that particular text type is easier to find. Also a balance in texts distributed over different time periods is 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 older stages of the language.

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* (old form of court records, see further Section 3.2.5) are also called for.

In addition, it is noticed that it is not trivial to decide the division into genres. An example given by one of the participants in the questionnaire is the genre “religious texts”, which typically includes very vivid legends as well as exegetic notes, which are very different in both content and linguistic structure.

4.1.6 The definition of Swedish

[*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 (and how should then “Swedish” be distinguished from other Nordic varieties at certain points in time)? Should the corpus also include texts written in Swedish outside the borders of Sweden, which could then include for example texts in Finland Swedish or American Swedish? Or do you have any other alternative to how the term “Swedish” should be defined in this context?*

Motivate your answer, and give examples to illustrate why you want to include or exclude certain texts, so that we can get a deeper understanding of the reasoning around these issues.]

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 variants such as Finland 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 different language variants.

The majority of the participants in the questionnaire would not include texts mainly written in another language, such as Latin or German, even if the text was written within the Swedish borders and with Swedish people as the intended readers. The aim of the diachronic corpus should be to make it possible for the researchers to study the development of the Swedish language, and including texts written in other languages would distort the corpus and make it harder to search it properly. Texts mainly written in Swedish, but with code-switching to other languages in parts of the text, should however be included.

In some cases it might be hard to determine for an older text if it is actually written in Swedish or in for example Danish or Norwegian. One suggestion for these tricky cases is to include these texts in the corpus, possibly with metadata indicating this uncertainty.

4.1.7 Search parametres

[*Are there any specific parametres that you would like to be able to search for in the corpus, for example words and phrases, names of persons or places, linguistic categories (such as part-of-speech, morphology, phrase category or syntactic function) or semantic features of any kind?*]

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

Most researchers are however interested in (automatic) linguistic annotation, to be able to perform more advanced search queries. Lemmatisation and/or truncation of words are mentioned as a way of facilitating search queries for all inflectional forms of the same lemma, and for different spellings of the same word form. Morphology is also mentioned as a very important feature in the search, especially for older texts, and it is pointed out that the list of derivational and inflectional morphemes is limited, making it possible to detect them automatically. One user also calls for a more precise morphological annotation, with the possibility to distinguish between for example past participle forms and the supine form. It would also be desirable with annotations of phenomena in for example Old Swedish that are not part of the present-day Swedish language, such as extinct case forms and inflectional forms of the verb for signalling number and person.

Other linguistic features mentioned as interesting are part-of-speech tagging, as well as phrase structures and syntactic categories. One researcher points out the importance of being able to search not only for existing parts of a sentence, but also for missing constituents in the sentence, such as omitted subjects, subordinating conjunctions, temporal auxiliaries or infinitive markers.

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, to distinguish between polysemous words. It would also be desirable to be able to sort the results alphabetically or chronologically.

4.1.8 Metadata information

[*What metadata categories should be included in the corpus (year, author, genre, location, issue, etc.)?*]

As could 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. Concerning genre, it is also pointed out that the genre classification might be tricky, since different researchers may want to classify the same text as belonging to divergent categories, so it is important to have an intuitive and well-defined genre classification. A similar problem arises for language varietes. It would be good to classify texts as being written in Finland Swedish, skånska (Scanian Swedish), American Swedish etc, but it might sometimes be hard to do such a classification, due to uncertainty and different opinions on how to classify a specific text in this aspect.

Apart from the specific issue of a text, it could also be relevant to state the name of the editor. For manuscripts being copies of older texts, the name of the person writing the younger manuscript is also important. Even the printer could be of importance, since some printers (for example Salvius in the 1700s) had their own orthographical norm. For some research questions, the age of the author is relevant too.

If the text is part of a larger collection, such as the collected works of a particular author, it would be desirable to date the specific text within this collection. For search results in the form of for example concordances, it would also be good to point the user to the exact page in the text where this text passage occurs. Likewise, if the text is digitally available as fulltext, it would be desirable to add a URL to that webpage.

Finally, there is a suggestion to add a short presentation of the text, to give the user an idea of the contents of the text. In many editions of older text, there is a quite long preface describing the text and the particular edition of the text. A shorter description would give the user a much faster and easier way to determine if the text is relevant or not for his/her research.

4.2 Questionnaire responses: Discussion and implications

The responses from the Swedish historical linguists to our questionnaire provide valuable background information in our planning for a Swedish diachronic corpus.

At the same time the answers must be approached with some caution, since we note that while we learn from them that corpus data are used by Swedish historical linguists in their research, it also became clear that this community – at least if we assume the respondents to our questionnaire to be both representative and numerous enough – is less experienced when it comes to utilizing historical corpora and corpus-linguistic methodology than seems to be the case among historical linguists working on English, as witnessed by a respectable number of publications crucially utilizing historical and diachronic corpora of English; see the references in Jenset and McGillivray (2017), and also the *Language Change Database*[64] (Kesäniemi et al. 2018).

An important contributing factor is of course the large differences in existence and availability of historical and diachronic corpora between Swedish and English. Hopefully, with the planned Swedish diachronic corpus we will be able to narrow this resource gap considerably.

We see this relative inexperience of corpus methods reflected in some of the responses, where respondents conceptually seem to equate corpora with traditional data sources, such as original manuscripts and their scientific editions, i.e., basically a close-reading scenario, but on the computer screen instead of parchment or paper. In other words, in the Swedish case, the use of corpus data seems to be largely qualitative, and few – if any – Swedish historical linguists show any awareness of the methodological requirements defining the "quantitative historical linguistics" described by Jenset and McGillivray (2017).

[64] http://www.helsinki.fi/lcd/

Perhaps at least in part determined by this view on the nature and role of corpora, the responses to the question about how "Swedish" should be interpreted in Section 4.1.6 are interesting and instructive. With a few notable exceptions, they seem to reflect a traditional language-internally focused view on language change, and not for instance an approach where individual languages are seen as always embedded in and interacting with larger (linguistic and extralinguistic) contexts, as pursued, e.g., by students of *linguistic ecology* (Haugen 1972/1971; Ansaldo and Lim 2017) or *contact linguistics* (Hickey 2010). Internationally, we have long seen an increasing awareness among historical and comparative linguists of the influence of "extralinguistic" factors on language change. Including other languages than Swedish in the envisaged diachronic corpus will hopefully facilitate investigations into the role of language contact in the historical development of Swedish.[65] Obviously, users of the corpus must be able to select – include or exclude – such texts through corpus metadata filtering, a point made by several of the respondents.

Summing up, we believe that it is important in building a Swedish diachronic corpus both to address the issues raised by our respondents from the perspective of a more traditional Swedish historical linguistics methodology and to be forward-looking at the same time, so that the availability of very large amounts of diachronic texts together with increasingly sophisticated methods for automatic linguistic annotation can add real value to and open new avenues for historical linguistic inquiry (see, e.g., Jenset and McGillivray 2017; de Marneffe and Potts 2017).

[65] Exactly for this reason, we have from the outset explicitly referred to the future outcome of our endeavor as *Svensk diakronisk korpus* 'Swedish diachronic corpus' (where *svensk* 'Swedish' is ambiguous in exactly the same way in Swedish and English), and not, e.g., *Diakronisk korpus för svenska* 'Diachronic corpus of Swedish', thereby leaving the question open about the desired language composition of such a corpus.

5 Summary and conclusions

In this report, with the aim of making an inventory of available resources that could be suitable for inclusion in an upcoming Swedish diachronic corpus, we have presented textual resources available for different time periods throughout the history of the Swedish language development, from Runic Swedish (appr. 800–1225) over Old Swedish (appr. 1225–1526), Early Modern Swedish (appr. 1526–1732) and Late Modern Swedish (appr. 1732–1900) up to and including Contemporary Swedish (appr. 1900 onwards). The report also includes the results of a questionnaire sent out to the primary target group for such a corpus, i.e. language historians, with questions about their experience of historical and diachronic corpora, and what specific user needs and wishes they have related to a Swedish diachronic corpus.

Table 28 shows the approximate number of words in the text material presented for the different time periods studied. As seen from the table, there is a substantial increase in the amounts of text available the closer we get to present-day times, with about 4.6 million words for the Old Swedish era, as compared to almost 10.7 billion words for contemporary Swedish.

Time Period	Number of words
Old Swedish	4,641,408
Early Modern Swedish	24,700,328
Late Modern Swedish	1,516,865,748
Contemporary Swedish	10,696,957,453

Table 28: Approximate number of words in the text material available for different time periods in the development of the Swedish language.

This brings up the question of how to make the corpus balanced, since there are so large differences in the size of the material available for different time periods. The opportunistic way would be to include as much texts as possible in the corpus, and leave it up to the user to select appropriate subcorpora from the corpus as a whole, to suit their research. This would also be in line with the answers given by the users in the questionnaire, where many researchers expressed a need for large amounts of text, and that it should be possible for the user to easily select or deselect any texts from the corpus, to make the data fit their specific research questions. It could also be pointed out that the potential users of the corpus form a quite heterogeneous group with diverse research interests. Some users might for example be interested in studying texts from the 19th century only, and then as much data as possible from this time period would be desirable, regardless of what could be found for earlier time periods.

On the other hand, it could be useful to have one or more predefined subcorpora, that are balanced both regarding size and text types included in the sample. Such predefined subcorpora would make it easier for researchers to compare their results to results achieved by other researchers, on exactly the same corpus. It would further facilitate comparative studies involving other languages for which such balanced corpora exist, such as for example the *Corpus of Historical American English* (Davies 2012), the *GerManC* corpus for German (Scheible et al. 2011; Durrell et al. 2012) or the *Icelandic Parsed Historical Corpus* (Rögnvaldsson et al. 2012).

Concerning text types, table 29 summarises the text genres represented in the textual resources for the different time periods in our study (excluding Runic Swedish, that is hard to classify into genres).

Genre	Old	Early Modern	Late Modern	Contemporary
Religious texts	x	x	x	x
Secular prose	x	x	x	x
Diaries and personal stories	-	x	x	x
Song texts	-	x	-	x
Periodicals	-	x	x	x
Newspapers	-	-	x	x
Letters and charters	x	x	-	x
Academic and scientific text	x	x	x	x
Court records	x	x	x	x
Laws and regulations	x	x	x	x
Governmental texts	-	x	x	x
Essays and school-related texts	-	-	-	x
Accounts and registers	x	x	-	x
Maps	-	x	-	-
Social media text and Wikipedia	-	-	-	x

Table 29: Genres included in the text material available for different time periods in the development of the Swedish language.

As argued in Section 3, dividing texts into genres is not trivial, and there are alternative ways of doing this division. Following the classification made throughout this report however, we can see that there are five genres that are represented in all time periods: (i) religious texts, (ii) secular prose, (iii) academic and scientific text (including medicine), (iv) court records, and (v) laws and regulations. It would thus be possible to create a subpart of the diachronic corpus that is balanced in genres for all time periods, based on these five text types.

A text type that was specifically mentioned as desirable to study by the users participating in the questionnaire is informal texts written by "ordinary" people, such as letters, diaries, texts written in different dialects and texts that represent the spoken language in one way or another, for example drama. The genre overview in the table shows that diaries and personal stories are available for all time periods except the Old Swedish era, whereas letters are available for all periods but the Later Modern period. It should however be noted that letters include not only personal letters but also more formal letters, written to the authorities. This is especially true for the Old Swedish period, where the 'letter' category mainly comprises *Svenskt Diplomatarium* (Diplomatarium Suecanum),[66] containing Medieval charters. The 'drama' category is however included in the 'secular prose' genre, that is present for all time periods. Another text type mentioned by the users is *Tänkeböckerna* (old form of court records, see further Section 3.2.5). These are included in our overview of available resources, categorised as 'court records'.

5.1 Future work

From the user questionnaire, it is obvious that a Swedish diachronic corpus would be very useful for many researchers in (especially) historical linguistics. The next step in our work would therefore be to, based on insights from this report and the previous report in Pettersson and Borin (2019), start building a first version of the Swedish diachronic corpus.

In the long view, we aim for a large diachronic corpus, including predefined subcorpora based on balance and representativeness for different time periods, as well as the possibility for users to define their own subcorpora based on the texts in the corpus as a whole. A yet unsolved issue concerns the time span of the corpus as a whole, as well as the time span of these predefined time periods. Most researchers in the user questionnaire agree that Old Swedish should be included in the diachronic corpus, but whether or not Runic Swedish should be included is still an open question. However, since the Runic material is rather limited, and already digitised, it would be fairly easy to include, and the researcher accessing the corpus could choose whether s/he wants to include these specific data in their search or not. Concerning the time span of the pre-

[66] https://riksarkivet.se/diplomatarium-suecanum

defined time periods, due to the limited amount of text for older time periods as compared to younger time periods, it could be an option to divide for example Old Swedish into 50-year periods or even centuries, whereas contemporary Swedish could be divided into for example 25-year periods or decades.

In the longer term, corpus compilation could include OCR and transcription of resources currently not available in a digitised format, to reach the goals of representativeness. However, for the first version of the corpus the aim is to include resources that are already available in a digitised format. These texts will then be converted to a clearly defined technical format (or several equivalent formats). Furthermore, as detailed metadata information as possible should be added to each text, preferably in a TEI-compatible format. The genre classification should be well-defined and as intuitive as possible. To mirror general text properties while at the same time providing a level of specificity, the genre classification could be divided into main genre and sub-genre. Another piece of metadata information that is worth discussing concerns dating. It is sometimes hard to know when a certain text was written, especially for older texts. Even if a document has been assigned a date, this date sometimes refers to the date when the original text was written, and sometimes to the dating of the manuscript in which the actual version of the text is found, which could be a copy of the text written centuries later, and edited to an unknown degree as compared to the original text. Any uncertainty about the dating should be clearly stated in the metadata information.

Furthermore, we aim for download possibilites for all texts that are not protected by copyright. Long-term goals include a graphical user interface as well, providing user-friendly search possibilites, with concordance search and linguistic annotation. Due to the large amounts of text planned to be included in the corpus, a fully manual annotation is unfortunately not an option. One idea could however be to manually post-correct the linguistic annotation for parts of the corpus, especially older parts for which automatic methods are known to perform worse. One could also think of manual annotation for the predefined subcorpora mentioned above, compiled with the goal to be balanced and representative, for use in comparative studies. Here, crowdsourcing or "expert sourcing" could be an option.

For enhanced search, we also aim for spelling harmonisation (also known as 'spelling normalisation') for the older texts (and possibly also spelling standardisation for contemporary non-standard text, such as social media text), enabling the user to search for a single word form (typically corresponding to the modern standard spelling), such as *skriva* 'to write', and receive results for other spellings of the same word form as well (such as *skrifva* or *skriffa*).

The answers to the user questionnaire clearly point in the direction of adding all sorts of texts written in Swedish to the corpus, including language varieties such as Finland Swedish, American Swedish and different Swedish dialects, but not texts mainly written in another language, such as Latin or German, even if the text was written within the Swedish borders and with Swedish people as the intended readers (but see the discussion in Section 4.2). When possible, we intend to add information on language variety to the metadata information stated for each text.

Last but not least, the users responding to the questionnaire point out that it could be a good idea to include a reference group of language historians to give input on the structure and contents of the diachronic corpus, during the development of the corpus.

Acknowledgements

This project was funded by Swe-Clarin.[67] We are also very grateful to all persons who have contributed with data and knowledge to this report, in particular to those who generously made the effort to answer the user questionnaire.

[67] https://sweclarin.se/

References

Adesam, Yvonne, Malin Ahlberg, Peter Andersson, Lars Borin, Gerlof Bouma and Markus Forsberg. 2016. Språkteknologi för svenska språket genom tiderna. *Studier i svensk språkhistoria 13*, 65–87. Umeå: Umeå University.

Adesam, Yvonne, Dana Dannélls and Nina Tahmasebi. 2019. Exploring the quality of the digital historical newspaper archive Kubhist. *Proceedings of DHN 2019*, 9–17. Aachen: CEUR-WS.org.

Ågren, Maria, Rosemarie Fiebranz, Erik Lindberg and Jonas Lindström. 2011. Making verbs count. The research project 'Gender and Work' and its methodology. *Scandinavian Economic History Review* 59 (3): 271–291.

Ansaldo, Umberto and Lisa Lim. 2017. Editorial. *Language Ecology* 1 (1): 1–3.

Bergman, Gösta. 1995. *Kortfattad svensk språkhistoria*. 5th. Stockholm: Prisma Magnum.

Borin, Lars, Markus Forsberg and Johan Roxendal. 2012. Korp – the corpus infrastructure of Språkbanken. *Proceedings of LREC 2012*, 474–478. Istanbul: ELRA.

Davies, Mark. 2012. Expanding horizons in historical linguistics with the 400 million word Corpus of Historical American English. *Corpora* 7 (2): 121–157.

Delsing, Lars-Olof. 2002. Fornsvenska textbanken. Svante Lagman, Stig Örjan Olsson and Viivika Voodla (eds), *Nordistica tartuensia 7*, 149–156. Tallinn: Pangloss.

Durrell, Martin, Paul Bennett, Silke Scheible and Richard J. Whitt. 2012. The GerManC corpus. Technical Report, School of Languages, Linguistics and Cultures, The University of Manchester, Manchester.

Evert, Stefan. 2019. *The IMS Open Corpus Workbench (CWB) – CQP query language tutorial, CWB version 3.4.16*. The CWB Development Team.

Haugen, Einar. 1972/1971. The ecology of language. Anwar S. Dil (ed.), *The ecology of language: Essays by Einar Haugen*, 325–339. Stanford: Stanford University Press. (Reprinted from *The Linguistic Reporter* 13[1/S25]: 19–26. Winter 1971. Washington, DC: Center for Applied Linguistics).

Hickey, Raymond. 2010. *The handbook of language contacct*. Oxford: Wiley-Blackwell.

Höder, Steffen. 2011a. The Hamburg Corpus of Old Swedish with Syntactic Annotation (HaCOSSA). Archived in Hamburger Zentrum für Sprachkorpora. Version 1.0. Publication date 2011-06-30. http://hdl.handle.net/11022/0000-0000-9D16-7.

Höder, Steffen. 2011b. 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.

Jenset, Gard B. and Barbara McGillivray. 2017. *Quantitative historical linguistics*. Oxford: Oxford University Press.

Johannessen, Janne Bondi, Øystein Alexander Vangsnes, Joel Priestley and Kristin Hagen. 2014. 2: A multilingual speech corpus of North-Germanic languages. Tommaso Raso and Heliana Mello (eds), *Spoken corpora and linguistic studies*, 69–83. Amsterdam: John Benjamins Publishing Company.

Kesäniemi, Joonas, Turo Vartiainen, Tanja Säily and Terttu Nevalainen. 2018. Open science for English historical corpus linguistics: Introducing the Language Change Database. *Proceedings of DHN 2018*, 51–62. Aachen: CEUR-ws.org.

de Marneffe, Marie-Catherine and Christopher Potts. 2017. Developing linguistic theories using annotated corpora. Nancy Ide and James Pustejovsky (eds), *Handbook of linguistic annotation: Volume 1*, 411–438. Dordrecht: Springer.

Pettersson, Eva and Lars Borin. 2019. Characteristics of diachronic and historical corpora: Features to consider in a Swedish diachronic corpus. Swe-Clarin Report Series (SCRS), no. SCR-01-2019. https://sweclarin.se/sites/sweclarin.se/files/diachronic-corpora-sweclarin-v3.pdf.

Rögnvaldsson, Eiríkur, Anton Karl Ingason, Einar Freyr Sigurðsson and Joel Wallenberg. 2012. The Icelandic Parsed Historical Corpus (IcePaHC). *Proceedings of LREC 2012*, 1977–1984. Istanbul: ELRA.

Scheible, Silke, Richard J. Whitt, Martin Durrell and Paul Bennett. 2011. A gold standard corpus of Early Modern German. *Proceedings of the 5th Linguistic Annotation Workshop*, 124–128. Portland, Oregon: ACL.