SWE-CLARIN

RAPPORTER FRÅN SWE-CLARIN
SWE-CLARIN REPORT SERIES
sweclarin.se/scrs
[SCR-01-2019]

Characteristics of diachronic and historical corpora

Features to consider in a Swedish diachronic corpus

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

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

Contents

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 this report, as a first step towards a Swedish diachronic corpus, we investigate the structure and contents of diachronic (and historical) corpora available for other languages. The goal of this study is 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. In particular, for each corpus in the study, we focus on the following features:

Size

1. How large is the corpus?

Structure, balance and representativeness

1. What time period is covered?
2. Granularity (decades, 50-year periods, centuries etc.)
3. Genre distribution

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

[2] https://sweclarin.se/

Linguistic annotation

1. What levels of annotation are present? (part-of-speech, lemma, morphology, syntax, semantics, spelling harmonization)
2. Was the annotation performed manually or automatically?

Metadata

1. What metadata information is given to represent the texts in the corpus?
2. What metadata standard is used?

Accessibility

1. Are there search facilities, such as a web interface?
2. Is it possible to download data, and if so, in what format(s)?
3. What licence is needed for accessing the data?

For obvious reasons it is not possible to describe all existing diachronic and historical corpora, especially considering that the terms "diachronic", "historical" and "corpus" are far from well-defined and narrow concepts.[3] In the following, we will therefore focus on a selection of diachronic/historical textual datasets for a variety of European languages. Some of these are accessible from the CLARIN historical corpora webpage,[4] providing access to 71 historical corpora, covering most of the languages spoken in countries that are members (or observers) of the CLARIN infrastructure.

Section 2 gives an overview of the selected diachronic and historical corpora. Some previously suggested useful features in historical corpora are listed in Section 3. Finally, the findings are summarized, and conclusions drawn, in Section 4, including some directions for future work.

[3]For our purposes, we explicitly take a broad view on what constitutes a "corpus", which will not always conform to how this term is defined in the field of corpus linguistics. By and large, we have accepted individual authors' labeling of the datasets they present as "corpora", even when a dataset departs quite strongly from the prototypical notion of a "corpus", such as in the case of the Google Books Ngrams (Section 2.13)

[4]https://www.clarin.eu/resource-families/historical-corpora

2 Diachronic and historical corpora

2.1 Czech

2.1.1 Diachronic Corpus of Czech (DIAKORP)

The *Diachronic Corpus of Czech* (DIAKORP) (Kučera, Řehořková and Stluka 2015) is the diachronic section of the *Czech National Corpus*. The diachronic section contains approximately 4 million words, covering seven centuries of the Czech language development; from the 14th century to the 20th century. There are nine main genres included in the corpus: drama, informal, non-fiction, opinion, periodical, poetry, prose, reflection, and speech. These are further divided into subgenres, as described in more detail below. Regarding balance and representativeness, it is stated on the DIAKORP wikipage that:

> Due to the length of the time span aimed to be covered and due to the decision to include whole texts instead of samples, DIAKORP was not designed to be a representative nor balanced corpus (whether in terms of register variability or period size).
>
> https://wiki.korpus.cz/doku.php/en:cnk:diakorp, accessed 23-01-2019

Furthermore, two main principles have been observed for the inclusion of texts in the corpus:

1. Texts are transcribed, not transliterated.
2. Texts are provided with structure attributes, such as headlines, footnoes, verses etc.

The DIAKORP corpus is provided as a tokenised plaintext file, without any linguistic annotation. Metadata information is provided as plaintext headers, with information on:

- title
- author
- publication year
- genre (drama, informal, non-fiction, opinion, periodical, poetry, prose, reflection, or speech)
- sub-genre (humour, tragedy, diary, agriculture, astrology, chemistry, education, history, medicine, military, mystery, philosophy, veterinary medicine, magazine, paper, song, bible, letter, metaphorical, narrative, novel, story, topography, travel, dialogue or prayer)

The DIAKORP corpus is licensed under a Creative Commons non-commercial share-alike licence.[5]

More information on the DIAKORP corpus can be found here:

- DIAKORP wikipage: https://wiki.korpus.cz/doku.php/en:cnk:diakorp
- Kučera, K., A. Řehořková and M. Stluka. 2015. DIAKORP: diachronic corpus of Czech, version 6. Institute of the Czech National Corpus, Faculty of Arts, Charles University, Prague. http://www.korpus.cz/.

2.2 ENGLISH

There are several historical corpora available, both for British English and for American English. Some of the more well-known are the *ARCHER* corpus, containing both British and American text from the time period 1600–1999, the *Corpus of Historical American English* (COHA, 1810s–2000s), the *Hansard* corpus of British parliament speeches (1803–2005), the *Lampeter Corpus of Early Modern English Tracts* (1640–1740), the *Old Bailey* corpus of London criminal court records (1674–1913), the *TIME* corpus of Time magazine articles (1923–2006), and many more. In the following, we will take a closer look at the ARCHER corpus, the COHA corpus, and the Penn Parsed Corpora of Historical English. For another substantial diachronic English dataset, the Google Books corpus, see Section 2.13 below.

2.2.1 ARCHER

A Representative Corpus of Historical English Registers (ARCHER) (Biber, Finegan and Atkinson 1994) is a multi-genre historical corpus of British and American English covering the time period 1600–1999, for a range of written and speech-based registers. In total, the corpus contains 3.3 million words,

[5] http://creativecommons.org/licenses/by-nc-sa/4.0/

distributed over 1,710 files, where 1,075 files (approximately 2 million words) contain British English, and 635 files (approximately 1.3 million words) contain American English.

Regarding balance and representativeness, the corpus is divided into 50-year periods, with equally sized text samples per genre and language variety in each 50-year period. The corpus is intended to grow over time, adding more text while at the same time keeping the relative composition of the corpus intact.

The ARCHER corpus is not linguistically annotated, but some of the texts have been annotated with modern spelling variants based on the VARD tool (Baron and Rayson 2008). Metadata is given in TEI format, with information on:

- filename
- author, and gender of author
- publication date
- availability
- extent (number of words)
- genre
- language

The ARCHER corpus may be accessed on-site for consortium universities only, and online after signing a specific user agreement.

More information on the ARCHER corpus can be found here:

- ARCHER webpage: http://www.helsinki.fi/varieng/CoRD/corpora/ARCHER/updated%20version/introduction.html
- Biber, Douglas, Edward Finegan and Dwight Atkinson. 1994. ARCHER and its challenges: Compiling and exploring a representative corpus of historical English registers. *Creating and using English language corpora. papers from the 14th international conference on English language research on computerized corpora, 1993*, 1–13.

2.2.2 Corpus of Historical American English (COHA)

The *Corpus of Historical American English* (COHA) (Davies 2012) covers the time period from the 1810s to the 2000s. The total size of the corpus is 406,232,024 words, distributed over 107,000 texts, and divided into decades. The corpus is balanced, containing four genres: fiction, popular magazines,

newspapers, and non-fiction books. For each decade, all four genres are represented, except for the 1810s to the 1850s, during which newspapers are not present. Furthermore, for each decade, there is a division into roughly half fiction and half non-fiction texts. The COHA webpage states that:

> The corpus is balanced across decades for sub-genres and domains as well (e.g. by Library of Congress classification for non-fiction; and by sub-genre for fiction – prose, poetry, drama, etc). This balance across genres and sub-genres allows researchers to examine changes and be reasonably certain that the data reflects actual changes in the "real world", rather than just being artifacts of a changing genre balance.
>
> COHA webpage (https://corpus.byu.edu/coha/), accessed 21-01-2019

The size of the material increases for each decade. Hence, the 1810s collection contains a total of 1,181,022 words, whereas the 2000s collection contains a total of 29,479,451 words. It could also be noted that some of the texts included in the corpus are copyright-protected. To follow the US Fair Use Law for these texts, every 200 words, ten words have been removed and replaced with the @ sign.

The COHA corpus is lemmatized and tagged for part-of-speech using the CLAWS tagger (Leech, Garside and Bryant 1994). Manual inspection and correction was carried out by students for words that had a significantly higher frequency in the COHA corpus than in its contemporary counterpart, the *Corpus of Contemporary American English* (COCA) (Davies 2010a).

Metadata information is provided in MS Excel xls format, with information on:

- title
- author and life span of the author (when applicable)
- volume, issue and page (when applicable)
- publication year
- publication decade
- number of words
- publisher
- genre (fiction, popular magazines, newspapers or non-fiction books)
- sub-genre (fiction: drama, movie scripts, novels, poetry, or short stories)
- source

The COHA corpus is searchable via a web interface, where the user can search by words, phrases, or lemmas. It is also possible to use wildcards and more complex searches such as "a most + ADJ + NOUN". Furthermore, the user may search for collocates and synonyms, and compare results for different decades and genres. If purchased, the corpus can also be downloaded, either as plain text, or in a tab-separated format containing information on word form, lemma and part-of-speech, or in an SQL format. There are different fees for use in research by single researchers and universities, as well as for other kinds of use (including commercial use), see further https://www.corpusdata.org/purchase.asp.

More information on the COHA corpus can be found here:

- COHA webpage: https://corpus.byu.edu/coha/
- Davies, Mark. 2012. Expanding horizons in historical linguistics with the 400 million word Corpus of Historical American English. *Corpora* 7 (2): 121–157.

2.2.3 Penn Parsed Corpora of Historical English (PPCHE)

The *Penn Parsed Corpora of Historical English* (PPCHE) is a collection of three corpora:

- The Penn-Helsinki Parsed Corpus of Middle English, second edition (Kroch and Taylor 2000), containing approximately 1.2 million words, covering the time period 1150–1500, with division into centuries (1150–1250, 1250–1350 etc.).
- The Penn-Helsinki Parsed Corpus of Early Modern English (Kroch, Santorini and Delfs 2004), containing approximately 1.7 million words, distributed over 448 text samples, covering the time period 1500–1720. The texts are divided into three subperiods: 1500–1569, 1570–1639, and 1640–1720.
- The Penn Parsed Corpus of Modern British English, second edition (Kroch, Santorini and Diertani 2016) containing approximately 2.8 million words, distributed over 275 text samples, covering the time period 1707–1914.

The corpora contain both full texts and text samples of British English prose from the time period 1150–1914, and are manually annotated with part-of-speech and syntax. Metadata include philological and bibliographical information, as well as word counts, dialect and genre information.

The Penn parsed corpora are distributed on CD-ROM, with different licence

fees for individuals, departments and libraries. More information on the Penn parsed corpora can be found here:

- Penn Parsed Corpora of Historical English webpage: https://www.ling.upenn.edu/hist-corpora/

Apart from the Penn Parsed Corpora of Historical English, there are several other historical corpora, both for English and for other languages, that follow the Penn-Helsinki annotation scheme for linguistic annotation, see further Section 2.12.

2.3 FRENCH

2.3.1 Syntactic Reference Corpus of Medieval French (SRCMF)

The *Syntactic Reference Corpus of Medieval French* (SRCMF) (Stein 2019; Prévost and Stein 2013) is a dependency treebank containing 15 texts (approximately 251,000 words), covering the Old French period from 842 to 1278, with manual syntactic annotation. The SRCMF texts are extracted from two corpora of medieval French:

1. Base de Français Médiéval (Guillot, Marchello-Nizia and Lavrentiev 2007)
2. Nouveau Corpus d'Amsterdam (Kunstmann and Stein 2007)

The corpus is searchable via a web interface. The treebank can also be downloaded either as TigerSearch-ready files, in TigerXML format, or in the original annotation format (RDF). Some texts are freely downloadable, and some have a Creative Commons non-commercial share-alike licence.[6]

Metadata is given in TEI-XML format, with information on: author, title, date, edition, history (version, licence etc.), and format (TigerXML etc.). Furthermore, information on genre and word count is accessible via a table on the webpage.

More information on the SRCMF can be found here:

- SRCMF webpage: http://srcmf.org/
- Stein, Achim. 2019. Diachronic syntax based on constituency and dependency annotated corpora. *Linguistic Variation* 18 (1): 74–99.

[6] http://creativecommons.org/licenses/by-nc-sa/2.5

2.4 GEORGIAN

2.4.1 The Georgian National Corpus

The *Georgian National Corpus* (Gippert and Tandashvili 2015) (GNC) is an explicitly diachronic corpus, intended to cover the whole extent of the history of written Georgian (5th century CE – present day). It is stratified into three chronological layers, Old Georgian (5th–13th century), Middle Georgian (13th–18th century), and Modern Georgian (19th century onwards). It also contains a subcorpus of transcribed modern Georgian spoken varieties. The Old Georgian part comprises around 6 million tokens, Middle Georgian is represented with roughly 1.4 million tokens, and the Modern part weighs in at almost 200 million tokens. The corpus is morphologically annotated and lemmatized, and the historical components have had harmonized spellings added (while retaining original spellings).

The GNC is searchable via CLARINO's corpus search interface *Corpuscle*: http://clarino.uib.no/korpuskel/page?page-id=korpuskel-main-page

Texts are stored in the Corpuscle format, and their metadata provide information on (at least):

- title
- author
- creation date
- language variety
- text genre and subgenre(s)
- source
- edition
- licence

The GNC corpus pages list different licences for different components. The most common licence cited is a Creative Commons non-commercial licence.[7]

More information about the GNC can be found here:

- The GNC webpage: http://gnc.gov.ge/gnc/page
- Corpuscle: http://clarino.uib.no/korpuskel/corpus-list
- Gippert, Jost and Manana Tandashvili. 2015. Structuring a diachronic corpus: The Georgian National Corpus project. Jost Gippert and Ralf

[7] https://creativecommons.org/licenses/by-nc/4.0/

Gehrke (eds), *Historical corpora: Challenges and perspectives*, 305–322. Tübingen: Narr.

2.5 GERMAN

For German, we will have a closer look at four diachronic corpora: *Deutsches Textarchiv* (DTA), *GerManC*, *Reference Corpus of Middle High German* (ReM), and *Register in Diachronic German Science* (RIDGES).

2.5.1 Deutsches Textarchiv (DTA)

The *Deutsches TextArchiv* (DTA) core corpus (Geyken et al. 2010) contains 1,406 texts, within three main genres: 635 scientific texts, 523 fiction texts, and 248 non-fiction texts. The corpus is compiled with a distribution of texts that should be as balanced as possible with regard to the different disciplines and text types, with a text selection based on several sources, including bibliographies from history of literature, text selections from German Dictionaries, and recommendations from specialists in various disciplines. Some of the texts were manually transcribed, and some were OCRed and manually post-corrected.

The downloadable files are not linguistically annotated. However, in the search interface, automatic tokenisation, lemmatisation, and part-of-speech tagging have been performed, enabling the user to search for certain part-of-speech sequences, or all word forms connected to the same lemma etc.

Each text in the corpus is stored in an XML format, with metadata provided in accordance with the TEI standard, with information on:

- title
- author
- publication year
- editor
- edition
- number of images, tokens, types, and characters
- publisher
- licence
- language

From the DTA webpage,[8] it is possible to download the full DTA core corpus.

[8] http://www.deutschestextarchiv.de/download

The user could also choose to download specific subcorpora connected to the main genres (fiction, non-fiction or science), or subcorpora for different centuries (17th century, 18th century or 19th century). The texts are released under a Creative Commons non-commercial licence.[9]

More information on the DTA corpus can be found here:

- DTA webpage: http://www.deutschestextarchiv.de/
- Geyken, Alexander, Susanne Haaf, Bryan Jurish, Matthias Schulz, Jakob Steinmann, Christian Thomas and Frank Wiegand. 2010. Das Deutsche Textarchiv: Vom historischen Korpus zum aktiven Archiv. *Digitale Wissenschaft*, pp. 157–161.

2.5.2 GerManC

The *GerManC* corpus (Scheible et al. 2011; Durrell et al. 2012) was compiled with the aim of building a representative historical corpus of written German for the time period 1650–1800. A further aim was for the corpus to enable comparative studies of the historical development of grammar and vocabulary in English and German. Thus, the GerManC corpus follows the structure of the English ARCHER corpus (see further Section 2.2.1). For balance and representativeness, the corpus consists of text samples of about 2,000 words from eight different genres. Four of these genres are selected to represent orally oriented registers: drama, newspapers, sermons and personal letters. The other four genres are selected to represent more print-oriented registers: narrative prose (fiction or non-fiction), scholarly (i.e. humanities), scientific and legal texts. The complete corpus consists of approximately 800,000 words, distributed over 360 samples. The corpus is also divided into 50-year sections (1650–1700, 1700–1750 and 1750–1800), with an equal number of texts from each genre for each sub-period. Durrell et al. (2012) further state that:

> Given the areal diversity of German during this period, the corpus also aimed for representativeness in respect of region, and to this end broad regional divisions were adopted for the GerManC corpus, i.e. North German, West Central German, East Central German, South-West German (including Switzerland) and South-East German (including Austria), taking an equal number of texts for each genre and sub-period from these five regions.

[9] https://creativecommons.org/licenses/by-nc/3.0/de/

Each text sample has been transcribed in a double-keying process, to ensure transcription quality. Characters that are not present on an English keyboard are typed as Unicode strings. Furthermore, the texts have been automatically annotated with word tokens (GATE tokeniser), sentence boundaries (ANNIE sentence splitter), modernised spelling variants (Jurish 2012), lemmas, part-of-speech tags (STTS tagset), morphological tags, and grammatical functions (Bohnet 2010).

Metadata are given in TEI format, with the following information:

- title
- author
- publication date
- publication place
- name of file
- region
- genre
- details about selected sample

The GerManC corpus is distributed under a Creative Commons non-commercial share-alike licence.[10]

More information on the GerManC corpus can be found here:

- GerManC webpage: http://ota.ox.ac.uk/desc/2544
- 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.
- 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.

2.5.3 Reference Corpus of Middle High German (ReM)

The *Reference Corpus of Middle High German* (ReM) (Klein and Dipper 2016) is a corpus of diplomatically transcribed and manually annotated texts from the Middle High German period (1050–1350), containing approximately 2 million word forms. It was created by merging five existing corpora: *Kölner Korpus Hessisch-Thüringischer Texte*, *Bonner Korpus Mitteldeutscher Texte*,

[10] http://creativecommons.org/licenses/by-nc-sa/3.0/

Bochumer Mittelhochdeutschkorpus, *Korpus der Mittelhochdeutschen Grammatik*, and *Referenzkorpus Mittelhochdeutsch*.

The corpus is transcribed in two separate layers. In the diplomatic layer, the historical graphemes and word boundaries are preserved. Layout information, such as page or line breaks, refers to this layer. The second layer adapts word boundaries to the conventions of modern German and serves as the basis for further linguistic annotation. The texts have been annotated with part-of-speech tags (using the HiTS tagset), morphological information, and lemma.

The texts in the corpus are provided in an XML format, with metadata information on:

- title
- author
- publication/manuscript date
- publication place
- edition
- sample extent (e.g. pages)
- language (including area, region and language type)
- text source and language of the author (if translated)
- genre
- sub-genre
- names of the persons doing the pre-annotation, annotation, proofreading, collation and digitization of the text
- notes from the annotator and the transcriber
- library (where the source text can be found)
- medium (manuscript, print etc.)
- online link
- reference

The ReM corpus can be searched using the corpus query tool ANNIS (Krause and Zeldes 2016).[11] It is also possible to download the corpus in CorA-XML format. Furthermore, the user may download pdf versions of the transcribed texts (without annotation). The ReM corpus is licensed under a Creative Com-

[11] https://linguistics.rub.de/annis/annis3/REM/

mons share-alike licence.[12] More information on the ReM corpus can be found here:

- ReM webpage: https://www.linguistics.rub.de/rem/
- Klein, Thomas and Stefanie Dipper. 2016. Handbuch zum Referenzkorpus Mittelhochdeutsch. *Bochumer Linguistische Arbeitsberichte*, vol. 19.

2.5.4 Register in Diachronic German Science (RIDGES)

The *Register in Diachronic German Science* corpus (RIDGES) (Odebrecht et al. 2017) is a diachronic corpus of scientific texts printed from the 15th to the 20th century (1482–1914). Before the 15th century, scientific texts in Europe were mainly written in Latin, so the main motivation for creating the RIDGES corpus was to be able to study the development of a German scientific register, independent of Latin, when scientists began writing in German instead. New texts are continuously added to the corpus. Version 8 contains 60 texts (approximately 3 million words) within the genres of alchemy, astronomy, botany, gardening, kitchen, linguistics, medicine, and religion.

The RIDGES corpus is structured into four layers:

1. **ocr**
   An OCR layer, with the raw output from the OCR scanning.
2. **dipl**
   A diplomatic layer, where the word forms are transcribed exactly as found in the manuscript.
3. **clean**
   A cleaned layer, where graphical structures and special characters have been converted to a more easily processable format.
4. **norm**
   A harmonization layer, where orthography, phonology, morphology, and word formation (morphemes) are standardized to resemble present-day German word forms.

The texts have been automatically annotated for part-of-speech using the German STTS tagset,[13] for lemma using Treetagger (Schmid 1994), for morphology and for dependency relations based on the TIGER annotation scheme[14]

[12] http://creativecommons.org/licenses/by-sa/4.0/

[13] http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/TagSets/stts-table.html

[14] http://www.linguistics.ruhr-uni-bochum.de/~dipper/pub/tiger_annot.pdf

(prepared with Mate Tools[15]). Each text also contains metadata information describing:

- title
- author
- publication date
- genre
- sub-genre

The RIDGES corpus can be searched online using the ANNIS system. The texts are also downloadable in five formats: Paula XML, ANNIS, CoNLL, PTB and PML.

The corpus is released under a Creative Commons attribution licence.[16]

More information on the RIDGES corpus can be found here:

- RIDGES webpage: https://www.linguistik.hu-berlin.de/en/institut-en/professuren-en/korpuslinguistik/research/ridges-projekt
- Odebrecht, Carolin, Malte Belz, Amir Zeldes, Anke Lüdeling and Thomas Krause. 2017. RIDGES Herbology: designing a diachronic multi-layer corpus. *Language Resources and Evaluation* 51 (3): 695–725.

2.6 HUNGARIAN

2.6.1 Hungarian Generative Diachronic Syntax (HGDS)

Within the *Hungarian Generative Diachronic Syntax* (HGDS) project (Simon 2014), an annotated corpus of more than 3.2 million tokens was developed, including 47 Old Hungarian codices, 24 Old Hungarian minor texts, 244 letters, and 5 Bible translations from the Middle Hungarian period. Regarding balance and representativeness, it is stated on the HGDS webpage that:

> Our aim was to construct an annotated corpus comprising all extant texts from the Old Hungarian period (896–1526) and several texts from the Middle Hungarian period (1526–1772), which could provide answers to linguistically relevant problems.
>
> http://omagyarkorpusz.nytud.hu/en-descr.html, accessed 07-02-2019

[15] http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/matetools.html
[16] http://creativecommons.org/licenses/by/4.0/

Each word form in the corpus has been automatically OCR:ed and manually post-corrected. For some texts, additional annotation is given in CoNLL-U format,[17] with information on manually modernized spelling, automatically extracted lemma, and automatic morphological and syntactic analysis in accordance with the universal dependencies annotation scheme.[18]

The metadata used to describe the texts refer to the original codex and not to the printed edition. Metadata information includes title, author, year and edition, as well as information on annotation level and comments from the transcriber.

More information on the HGDS corpus can be found here:

- HGDS webpage: http://omagyarkorpusz.nytud.hu/en-descr.html
- Simon, Eszter. 2014. Corpus building from Old Hungarian codices. Katalin É Kiss (ed.), *The evolution of functional left peripheries in Hungarian syntax*, 224–236. Oxford: Oxford University Press.

2.7 NORDIC LANGUAGES

This section describes some of the historical corpora and text archives available for the Nordic languages: *Fornsvenska textbanken* (FSV), the *Hamburg Corpus of Old Swedish with Syntactic Annotations* (HaCOSSA), *Språkbanken's historical corpora*, the *Icelandic Parsed Historical Corpus* (IcePaHC), the *Faroese Parsed Historical Corpus* (FarPaHC), and the *Medieval Nordic Text Archive* (Menota).

2.7.1 Fornsvenska textbanken (FSV)

Fornsvenska textbanken (FSV) (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 form a balanced selection of texts, and the website also states that the texts are distributed without quality assurance.[19]

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 licens-

[17] https://universaldependencies.org/format.html
[18] https://universaldependencies.org/
[19] http://project2.sol.lu.se/fornsvenska/

ing information is provided, but note that the same texts are available as part of Språkbanken's historical corpora under a Creative Commons attribution licence.[20] (see Section 2.7.3 below),

More information on Fornsvenska textbanken can be found here:

- FSV 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.

2.7.2 HaCOSSA

The *Hamburg Corpus of Old Swedish with Syntactic Annotations* (HaCOSSA) (Höder 2011) 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 (geographical, theological, historic, natural science), and diplomas. Regarding the representativeness of the corpus, the HaCOSSA webpage states that:

> The bulk of the material represents the dominant text types of the Late Old Swedish period, namely different genres of religious/monastic prose (biblical, liturgical, and literary texts) but also profane texts (administrative and literary genres). As the main focus of the project was the study of contact-induced language change, the corpus consists mostly of texts that were translated from or otherwise influenced by Latin sources.
>
> https://corpora.uni-hamburg.de/hzsk/en/islandora/object/text-corpus%3Ahacossa, accessed 18-02-2019

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. However, the annotation is partial; only preverbal constituents and subjects are provided with full annotation. Further, clauses are annotated for their type. More rarely, other constituents besides the already mentioned may be annotated.

The corpus is available for non-commercial research and teaching. Access to

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

the corpus is granted by contacting *Fachinformationsdienst Nordeuropa* at the Kiel University Library.[21]

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. 2011. 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.

2.7.3 Språkbanken's historical corpora

Språkbanken at the University of Gothenburg makes a large number of historical Swedish corpora available for online search through the Korp corpus infrastructure (Borin, Forsberg and Roxendal 2012),[22] and also as downloadable datasets in XML format, comprising the texts plus any linguistic annotations and metadata added by the Korp corpus import pipeline.[23] The downloadable corpora are released under a Creative Commons attribution licence.[24]

The texts represent a heterogeneous mixture of genres and periods (Adesam et al. 2016), from the Old Swedish texts of Fornsvenska textbanken (see Section 2.7.1), over a collection of medieval letters digitized by the Swedish National Archives (*Svenskt diplomatarium*),[25] historical novels provided by *Litteraturbanken*, to a large body of older newspapers, the so-called *Kubhist* corpus (currently about one billion words, but version 2 which is now in the pipeline will have over 5 billion words covering the same timespan (1645–1926) as the present version, and improved OCR quality; Adesam, Dannélls and Tahmasebi 2019).

The format of the downloadable corpora is a bespoke XML format used by Språkbanken for all its downloadable corpora, and which basically comprises a light "XML-ification" of a CoNLL-like tabular format. Metadata information is given as attributes of the "text" element of this format. However, the individ-

[21] https://www.ub.uni-kiel.de/fach/sondersammlung/registrierung

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

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

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

[25] Svenskt diplomatarium also contains texts in other languages than Swedish, most notably Latin in which there in fact is more more text than in Swedish, but also at least German and Norwegian.

ual texts and corpora have varying amounts and kinds of metadata associated with them, and for this reason it is impossible to provide a list here.

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

- Korp search interface, historical mode:
  https://spraakbanken.gu.se/korp/?mode=all_hist#?lang=en
- Språkbanken'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.
- Adesam, Yvonne, Dana Dannélls and Nina Tahmasebi. 2019. Exploring the quality of the digital historical newspaper archive Kubhist. *Proceedings of DHN 2019*. Aachen: CEUR-WS.org. to appear.

2.7.4 The Swedish Culturomics Gigaword corpus

While the corpora described in the preceding section could be said to be diachronical in some sense, they are so only incidentally, by virtue of representing several historical periods of Swedish, and not by explicit design. However, Språkbanken also makes available for download a one-billion word diachronic corpus covering contemporary Swedish (1950–2015), the *Swedish Culturomics Gigaword* corpus (Eide, Tahmasebi and Borin 2016).[26] The corpus timestamp granularity is one year, but only starting in 1990 is there data for each consecutive year, while there are 9 years represented for the preceding 40-year period. For IPR reasons, the texts making up the Gigaword corpus are scrambled on sentence level. The resulting dataset is available under a Creative Commons attribution licence.[27]

The corpus is downloadable in decade chunks in a bespoke XML format used by Språkbanken for all its downloadable corpora, and which basically comprises a light "XML-ification" of a CoNLL-like tabular format. Metadata information is given as attributes of the "text" element of this format, with information on:

- year
- genre

[26] https://spraakbanken.gu.se/eng/resource/gigaword
[27] http://creativecommons.org/licenses/by/4.0/

More information on the Swedish Culturomics Gigaword corpus can be found here:

- the Swedish Culturomics Gigaword corpus resource page: https://spraakbanken.gu.se/eng/resource/gigaword
- Eide, Stian Rødven, Nina Tahmasebi and Lars Borin. 2016. The Swedish Culturomics Gigaword corpus: A one billion word Swedish reference dataset. *From digitization to knowledge 2016: Resources and methods for semantic processing of digital works/texts*, 8–12. Linköping: LiUEP.

2.7.5 Icelandic Parsed Historical Corpus (IcePaHC)

The *Icelandic Parsed Historical Corpus* (IcePaHC, version 0.9) (Rögnvaldsson et al. 2012) contains 60 texts in total, covering the time period from the 12th century to the 21st century. More specifically, the corpus contains 1,002,390 words, distributed over texts from 1150 to 2008. The corpus is balanced in the sense that it contains approximately 100,000 words from each century, where texts have been collected from five different genres:

- narrative texts (for all centuries except the 12th century)
- religious texts (for all centuries except the 15th and 21st century)
- biographies (17th and 18th century only)
- scientific texts (12th and 19th century only)
- law texts (13th century only)

One third of the texts were collected from text repositories on the internet, whereas another third were transcribed by students, and a few texts were received directly from scholars or publishing companies. The remaining texts were collected from the Project Gutenberg website,[28] the Internet Archive,[29] the Medieval Nordic Text Archive (see further Section 2.7.7), and from the Árni Magnússon Institute text archive.[30] For texts that are still under copyright, the corpus creators contacted the authors for permission to include them in the corpus. Only printed sources were included, and spelling was semi-automatically modernised, in order for existing natural language processing tools developed for present-day Icelandic to be applicable to the texts.

The IcePaHC corpus has been annotated with lemma, part-of-speech, and syn-

[28] http://www.gutenberg.org
[29] http://www.archive.org/
[30] http://www.lexis.hi.is/corpus/

tactic phrase structure. As a first step in the annotation process, students manually segmented the texts into sentences. After sentence segmentation, each sentence was run through the IceNLP tool (Loftsson and Rögnvaldsson) for tokenisation, part-of-speech tagging, lemmatisation and shallow parsing. Based on this, the final annotation was done manually, for the whole corpus.

The annotation scheme used is a slightly modified version of the Penn phrase structure annotation scheme. Rögnvaldsson et al. (2012) state that one advantage with this choice is that there is software developed for corpus annotation (CorpusDraw) and corpus search (CorpusSearch) (Randall 2005), based on the Penn scheme. Furthermore, they argue that since there are several diachronic and historical corpora available following the same scheme (see further Section 2.12), this makes it easier to compare for example old Icelandic to old stages of English.

For each text in the corpus, metadata information is provided in a tab-separated plaintext format, with the following entries:

- author
- birthdate
- textId
- textname
- edition (e.g. "Ectors saga. Late Medieval Icelandic Romances I, ed. Agnethe Loth, Kaupmannahöfn 1962. Bls. 81–169.")
- date
- genre
- wordcount
- sample (e.g. "Entire sample" or "Starting from Chapter 1")

The IcePaHC corpus is released under the free and open source LGPL licence.[31]

More information on the IcePaHC corpus can be found here:

- IcePaHC wiki page:
  http://www.linguist.is/icelandic_treebank/Icelandic_Parsed_Historical_Corpus_(IcePaHC)
- 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.

[31] https://opensource.org/licenses/lgpl-license

2.7.6 Faroese Parsed Historical Corpus (FarPaHC)

The *Faroese Parsed Historical Corpus* (FarPaHC) is a treebank of 53,000 words, manually annotated with part-of-speech and full phrase structures, using the annotation scheme of the Icelandic Parsed Historical Corpus (see further Section 2.7.5), which in turn is an adaptation of the annotation scheme used by the Penn Treebank and the Penn Parsed Corpora of Historical English (see further Section 2.2.3). The corpus contains three texts from the New Testament, published between 1823 and 1936, and the annotation for each sentence has been manually corrected. The corpus is released under an LGPL licence. Each text contains almost the same metadata information as the IcePaHC corpus, developed by the same researchers: author, life span of author, text id, text name, edition, spelling (modernized or not), publication date, genre, word count and sample.

More information on the FarPaHC corpus can be found here:

- FarPaHC webpage:
  https://einarfreyr.wordpress.com/2012/08/03/
  farpahc-0-1-53000-words-of-syntactically-parsed-hand-corrected-faroese/
- 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.

2.7.7 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.[32]

Metadata information is given in TEI format and includes:

- author
- title
- original manuscript date
- place of origin (country)
- language (Old Icelandic, Old Norwegian or Old Swedish)

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

- wordcount
- lemmatisation (none, partial or complete)
- morphological analysis (none, partial or complete)
- text representation level(s) (facsimile, diplomatic transcription and/or harmonized)
- transcription quality (low, medium or high)
- licence

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

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

2.8 Old Indo-European languages

2.8.1 The PROIEL treebank

The PROIEL family of treebanks (Eckhoff et al. 2018) constitute five corpora, following the same annotation scheme:

The PROIEL treebank (Haug and Jøhndal 2008) covering Ancient Greek and Latin, as well as the translations of the New Testament into Gothic, Classical Armenian and Old Church Slavonic

The TOROT treebank (Eckhoff and Berdicevskis 2015) covering Old Church Slavonic, Old East Slavic and Middle Russian

The ISWOC treebank (Bech and Eide 2014) covering Old English, Old French, Old Portuguese, and Old Spanish

The MAþiR treebank for Old Swedish

The Greinir skáldskapar treebank for Old Icelandic

Here, we will focus on the original PROIEL treebank (Haug and Jøhndal 2008), developed within the *Pragmatic Resources in Old Indo-European Languages* (PROIEL) project,[33] with the aim of facilitating comparative studies of the language in the Greek version of the New Testament and its translations into other Old Indo-European languages. Within the project, researchers are particularly interested in how the texts are structured linguistically in the different languages, and how this relates to pragmatic discourse phenomena, with a special focus on:

[33] https://www.hf.uio.no/ifikk/english/research/projects/proiel/

- word order
- discourse particles
- pronominal reference and the use of null pronouns
- expressions of definiteness
- the use of participles to refer to background events

The PROIEL treebank contains 614,512 tokens, distributed over 12 New Testament texts in Ancient Greek and Latin, as well as the translations of the New Testament into Gothic, Classical Armenian and Old Church Slavonic. The texts have been manually annotated with lemma and part-of-speech, as well as morphological and syntactic features. In addition, the texts are aligned, making it a parallel corpus.

The PROIEL treebank is released under a Creative Commons non-commercial share-alike licence,[34] and may be downloaded from GitHub,[35] in XML format or CoNLL format. The treebank is also searchable on the web, using either INESS Search[36] or Syntacticus.[37]

Metadata information is given in TEI format, with information on:

- text id and alignment id
- title
- language variant
- contact person, funder, distributor
- licence
- editor, annotator, and reviewer
- details on the electronic edition

More information on the PROIEL project and the PROIEL treebanks can be found here:

- PROIEL project webpage:
  https://www.hf.uio.no/ifikk/english/research/projects/proiel/
- PROIEL treebank webpage:
  https://proiel.github.io/

[34] https://creativecommons.org/licenses/by-nc-sa/4.0/
[35] https://github.com/proiel/proiel-treebank/
[36] http://clarino.uib.no/iness/treebanks
[37] http://syntacticus.org/

- Haug, Dag T. T. and Marius L. Jøhndal. 2008. Creating a parallel treebank of the Old Indo-European bible translations. *Proceedings of LaTeCH 2008*, 27–34. Marrakech: ELRA.
- Eckhoff, Hanne, Kristin Bech, Gerlof Bouma, Kristine Eide, Dag Haug, Odd Einar Haugen and Marius Jøhndal. 2018. The PROIEL treebank family: a standard for early attestations of Indo-European languages. *Language Resources and Evaluation* 52 (1): 29–65.

2.9 PORTUGUESE

For Portuguese, we will have a closer look at the *Tycho Brahe Corpus of Historical Portuguese* (TBCHP), and *Corpus do Português* (CdP).

2.9.1 Tycho Brahe Corpus of Historical Portuguese (TBCHP)

The *Tycho Brahe Corpus of Historical Portuguese* (TBCHP) (Galves and Faria 2017) is a corpus of texts written in Portuguese by authors born between 1380 and 1881. The corpus contains 3,302,666 words, distributed over 76 texts. In the search interface, texts are divided into 50 year-periods, within the following seven genres: letters, minutes, narrative text, dissertations, grammars, news, and theatre.

The Tycho Brahe corpus has been partly annotated for part-of-specch (44 texts, a total of 1,956,460 words) and phrase structure (20 texts, a total of 877,247 words). The annotation scheme used is a modified version of the Penn phrase structure annotation scheme. As a first step in the tagging task, an automatic tagger was used, implemented by Fabio Kepler at the University of São Paulo. The output was then manually revised and corrected using the graphical user interface Edictor.[38] Parsing was performed in a similar manner, using Dan Bickel's probabilistic parser for automatic analysis, and Annotald[39] for manual post-correction.

For each text in the corpus, metadata information is provided within square brackets. There is also a table with an overview of the different texts, with the following entries:

- code (i.e. text id)
- author
- birth (year)
- title

[38] https://edictor.net/

[39] https://annotald.github.io/

- words
- source text (either "edited" or "non-edited")
- modernized on transcription ("yes", "–" or "no")
- complete edition ("yes", "–" or "no")
- POS ("yes" or "no", since not all texts have a part-of-speech annotation)
- synt. ("yes" or "no", since not all texts have a syntactic annotation)

The Tycho Brahe corpus is available for free for educational and research purposes. The terms of use are described in more detail here: http://www.tycho.iel.unicamp.br/corpus/en/termos.html.

More information on the Tycho Brahe corpus can be found here:

- Tyco Brahe webpage:
  http://www.tycho.iel.unicamp.br/corpus/en/

2.9.2 Corpus do Português – Historical part (CdP)

The historical part of *Corpus do Português* (CdP)[40] is a 45 million word diachronic Portuguese corpus, distributed over nearly 57,000 texts, covering the time period from the 1300s to the 1900s. The texts from the 1900s (20 million words) are balanced over four genres: spoken, fiction, newspaper, and academic. Corpus do Português is searchable via a web interface, where the user can search by words, phrases, part-of-speech or lemma. It is also possible to use wildcards and to search for synonyms and collocates, and to compare European Portuguese to Brazilian Portuguese.

More information on the historical part of the CdP corpus can be found here:

- CdP webpage (historical part):
  https://www.corpusdoportugues.org/hist-gen/

2.10 Slovene

2.10.1 Reference Corpus of Historical Slovene (IMP-sl)

The *Reference Corpus of Historical Slovene* (IMP-sl) (Erjavec 2012) contains about 300,000 tokens from the time period 1584–1899, sampled from the resources developed within the Slovene part of the IMPACT project.[41] The texts have been manually annotated with modernised word forms as well as lemmas

[40] https://www.corpusdoportugues.org/hist-gen/

[41] Improving Access to Text: http://www.impact-project.eu/

and morphosyntactic annotation. As a first step in the annotation process, the corpus was automatically annotated using the ToTrTaLe tool (Erjavec 2011) for tokenisation, sentence segmentation, spelling modernisation, morphosyntactic tagging, and lemmatisation. In the second step, the automatically assigned annotations were manually checked and corrected, using the CoBaLT annotation editor (Kenter et al. 2012).

The corpus can be explored using a concordance search tool, or downloaded in XML format, under a Creative Commons attribution licence.[42] Metadata is given in TEI format, and include information on title, author, publication date, publisher, genre, and extent of the sample.

More information on the IMP-sl corpus can be found here:

- IMP language resources for historical Slovene:
  http://nl.ijs.si/imp/index-en.html
- Erjavec, Tomaž. 2012. The goo300k corpus of historical Slovene. *Proceedings of LREC 2012*, 2257–2260. Istanbul: ELRA.

2.11 Spanish

In the following, we will describe three historical corpora for Spanish in more detail: *Corpus del Español* (CdE), *Corpus Diacrónico del Español* (CORDE), and the *IMPACT-es Diachronic Corpus*.

2.11.1 Corpus del Español – Historical Part (CdE)

The historical part of *Corpus del Español* (CdE)[43] is a 100 million word diachronic Spanish corpus, distributed over more than 20,000 texts, covering the time period from the 1200s to the 1900s. The texts from the 1900s (20 million words) are balanced over four genres: spoken, fiction, newspaper, and academic. Corpus del Español is searchable via a web interface, where the user can search by words, phrases, part-of-speech or lemma. It is also possible to use wildcards and to search for synonyms and collocates. Furthermore, the texts in the corpus are divided into centuries, enabling the user to compare search results for different time periods. The linguistic annotation was performed semi-automatically, where the top 40,000 lemmas in the corpus were manually revised.

More information on the historical part of the CdE corpus can be found here:

[42] https://creativecommons.org/licenses/by/4.0/
[43] https://www.corpusdelespanol.org/hist-gen/

- CdE webpage (historical part):
  https://www.corpusdelespanol.org/hist-gen/

2.11.2 Corpus Diacrónico del Español (CORDE)

Corpus Diacrónico del Español (CORDE)[44] is a diachronic corpus of historical Spanish text, from "the beginning of the Spanish language" to 1974. The corpus is intended to be a balanced and representative corpus, including texts from different geographical areas, time periods and genres. In total, the corpus contains 250 million records, distributed over texts from 22 countries, extracted from five media (books, newspapers, journals, oral and miscellaneous). Furthermore, the texts are divided into nine main genres (further divided into several subgenres): science and technology, religion, politics and economy, arts, hobby, health, fiction, oral and miscellaneous. The following metadata fields are available for the texts:

- author
- title
- date
- medium
- region (typically country)
- genre
- subgenre

The corpus is freely available for search,[45] with fields for searching for specific authors, titles, media, geographical areas, genres, and subgenres. The user may also define a specific time period for the search, by specifying a start and end year.

More information on the CORDE corpus can be found here:

- CORDE webpage:
  http://www.rae.es/recursos/banco-de-datos/corde

2.11.3 IMPACT-es Diachronic Corpus of Historical Spanish

The *Impact-es Diachronic Corpus of Historical Spanish* (Sánchez-Martínez et al. 2013) comprises approximately 8 million words, extracted from over one hundred books. There is also a dictionary available, linking more than ten thou-

[44] http://www.rae.es/recursos/banco-de-datos/corde
[45] http://corpus.rae.es/creanet.html

sand lemmas with attestations of the different variants found in the documents. Both resources are released under a Creative Commons non-commercial share-alike licence.[46] The 7% most frequent words in the corpus have been annotated with lemma, part-of-speech, and their modern equivalent. Metadata is given following the TEI P5 guidelines,[47] with information on title, author, and publication date for the edition included in the corpus as well as for the source edition.

More information on the IMPACT-es corpus can be found here:

- IMPACT-es webpage:
  https://www.digitisation.eu/tools-resources/language-resources/impact-es/
- Sánchez-Martínez, F., I. Martínez-Sempere, X. Ivars-Ribes and R.C. Carrasco. 2013. An open diachronic corpus of historical Spanish. *Language Resources and Evaluation* 47 (4): 1327–1342. https://link.springer.com/article/10.1007%2Fs10579-013-9239-y.

2.12 Corpora with Penn-Helsinki annotation

For parsed historical corpora, there are several research groups that have decided to follow the Penn-Helsinki annotation scheme – i.e., a phrase-structure scheme – for linguistic markup. The use of the same annotation standard facilitates comparative studies between different languages based on these corpora.

On the Penn-Helsinki website[48] the following corpora are listed as annotated using the Penn-Helsinki annotation scheme (or slight modifications of it):

Parsed corpora of historical English

- York-Helsinki Parsed Corpus of Old English Poetry (Pintzuk and Plug 2002)
- York-Toronto-Helsinki Parsed Corpus of Old English Prose (Taylor et al. 2003)
- Brooklyn-Geneva-Amsterdam-Helsinki Parsed Corpus of Old English[49]
- Penn-Helsinki Parsed Corpus of Middle English, 2nd edition (PPCME2) (Kroch and Taylor 2000), see further Section 2.2.3

[46] https://creativecommons.org/licenses/by-nc-sa/3.0/
[47] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/
[48] https://www.ling.upenn.edu/hist-corpora/other-corpora.html
[49] http://www-users.york.ac.uk/~sp20/corpus.html

- Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME) (Kroch, Santorini and Delfs 2004), see further Section 2.2.3
- York-Helsinki Parsed Corpus of Early English Correspondence (PCEEC) (Taylor et al. 2006)
- Penn Parsed Corpus of Modern British English, 2nd edition (PPCMBE2) (Kroch, Santorini and Diertani 2016), see further Section 2.2.3
- Parsed Linguistic Atlas of Early Middle English (PLAEME)[50]
- Parsed Corpus of Middle English Poetry[51]

Parsed corpora of other languages

- **Historical French**: Modéliser le changement: les voies du français (Modelling change: the paths of French)[52]
- **Historical Icelandic**: Icelandic Parsed Historical Corpus (IcePaHC) (Rögnvaldsson et al. 2012), see further Section 2.7.5
- **Historical Portuguese**: Tycho Brahe Parsed Corpus of Historical Portuguese (Galves and Faria 2017), see further Section 2.9.1
- **Old Portuguese**: Word order and word order change in Western European languages (WOChWEL) Corpus[53]
- **Early Modern Portuguese and Spanish**: P.S. Post Scriptum - A Digital Archive of Ordinary Writing (CLUL 2014)
- **Old Japanese**: Oxford-NINJAL Corpus of Old Japanese (ONCOJ) (Oxford-NINJAL 2018)

In this report, we have also mentioned the Faroese Parsed Historical Corpus (FarPaHC), which uses the same annotation scheme.

2.13 The Google Books Ngram corpus and the Ngram Viewer

The *Google Books Ngram corpus* (Michel et al. 2011; Lin et al. 2012) – or simply *Google Ngrams* – consists of timestamped 1- to 5-grams with a frequency of at least 40 occurrences in the whole dataset[54] extracted from the books scanned in the Google Books project, comprising several million vol-

[50] https://github.com/rtruswell/PLAEME_current
[51] http://pcmep.net/
[52] http://www.voies.uottawa.ca/corpus_pg_en.html
[53] http://alfclul.clul.ul.pt/wochwel/oldtexts.html
[54] http://storage.googleapis.com/books/ngrams/books/datasetsv2.html

umes published between 1500 and 2008 in Chinese, (American and British) English, French, German, Hebrew, Italian, Russian, and Spanish. This dataset has been the basis of a number of large-scale data-driven computational studies on language change (in particular semantic change), concept history, cultural history and related areas. The dataset formed the basis for the so-called *culturomics* project (Michel et al. 2011; Aiden and Michel 2013) (see also Tahmasebi et al. 2015) and was subsequently released as a result of this project. In version 2, the tokens making up the n-grams have been automatically annotated for part of speech and UD dependency relations (Lin et al. 2012).

Although not a corpus in the normal sense of the term, since it does not consist of (carefully selected) texts (or text pieces), its inclusion in this report is motivated for at least two reasons:

(1) As just mentioned, Google Ngrams (together with COHA; see Section 2.2.2) has become the dataset of choice in computational studies of language change in (American) English over the last two centuries.

(2) It represents a concrete suggestion for dealing with IPR issues for in-copyright texts.[55]

An offshoot of the culturomics project was the *Google Ngram Viewer*,[56] an online interface where the user may search for words and phrases, and chart how these words and phrases have been used over time. There are several search possibilites available in the Google Ngram Viewer:

- **wildcard search**
- **inflectional search**
  All inflectional forms of a word may be searched by providing a lemma succeeded by the suffix _INF. For example, using the search query "book_INF a hotel" will display results for *book a hotel*, *booked a hotel*, *books a hotel*, and *booking a hotel*.
- **case-insensitive search**
- **part-of-speech**
  Doing a part-of-speech search, the user may search for nouns, verbs, adjectives, adverbs, pronouns, determiners, adpositions, numerals, conjunctions, and particles. In addition, there are special tags for the root of a parse tree, and for the start and end of a sentence. The part-of-speech tags can be used to search for example for verbs in general ("_VERB_"),

[55] The compiler of COHA takes a different approach; see Section 2.2.2.

[56] https://books.google.com/ngrams/info

or to search for a particular verb, such as "tackle_VERB_", which would exclude all cases where the word *tackle* has been analysed as a noun.

- **n-gram compositions**
  Using n-gram composition search, the user may for example compare ngrams of very different frequencies, or provide several search words to get a graph combining the multiple ngram series into one (useful for synonym search), or compare ngrams across different corpora (useful for example for studying the use of different concepts in American English versus British English, or for comparing different genres to each other).

More information on Google Ngrams and the Google Ngram Viewer can be found here:

- Google Ngram Viewer webpage:
  https://books.google.com/ngrams/info
- Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak and Erez Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science* 331 (6014): 176–182.
- Lin, Yuri, Jean-Baptiste Michel, Erez Lieberman Aiden, Jon Orwant, William Brockman and Slav Petrov. 2012. Syntactic annotations for the Google Books Ngram corpus. *Proceedings of ACL 2012*, 169–174. Jeju Island: ACL.

3 Useful features in historical corpora

Davies (2010b) lists a number of requirements on what search possibilites and levels of annotation historical corpora should offer, in order for the corpora to be truly useful for studying a wide range of linguistic phenomena. The features he lists are the following:

- **lexical search**
  The user should be able to search for a word or a phrase, and see the word or the phrase in its contexts, for example as concordances. In addition, for diachronic corpora it is important to offer possibilities for the user to study how the frequency of a word or a phrase has changed over time. Ideally, the user should also be able to search for words or phrases in a more advanced way, e.g. to search for all nouns that entered the language at a certain point in time, or all words that occur at least five times as often in the 13th century as in the 14th century.
- **morphological analysis**
  A morphologically annotated corpus enables the user to search by prefixes, suffixes and roots, and see the corresponding frequencies for these forms in different time periods. Another very useful natural language processing application in this context is *lemmatisation*, enabling the user to search for the base form of a word, to match all inflectional forms of that specific lemma.
- **syntactic analysis**
  With the term "syntactic", Davies (2010b) refers to *part-of-speech annotation*, enabling the user to search for certain part-of-speech sequences that could indicate a shift in syntactic structure at some point in time.
- **semantic search**
  The most basic form of semantic search in the context of historical corpora would be to enable the user to search for collocates of a given word or phrase. This could be very useful, since words that co-occur more of-

ten than would be expected by chance could give a good indication of the meaning of a word or a phrase. To facilitate studies on how meaning has changed over time, it would also be desirable to be able to compare collocates of a word or a phrase between different genres and time periods. In the most advanced setting, Davies (2010b) suggests that users should also be able to search by semantic field, rather than by words or phrases, e.g. searching for synonyms and antonyms, or for concepts such as "member of a family".

By and large, we would like to be able to offer a resource and tools which would allow historical linguists to move considerably beyond the state of the art in diachronic corpus linguistics as outlined by Davies (2010b) or the "quantitative historical linguistics" described by Jenset and McGillivray (2017), by being able to draw on the full language-technology expertise available in the Swe-Clarin consortium.

4 Summary and conclusions

As a preparatory step in the process of building a Swedish diachronic corpus, we have studied the structure and characteristics of a number of existing diachronic and historical corpora, and also listed specific features that in previous research have been identified as important for historical corpora to be useful for studying a wide range of linguistic phenomena. In Section 4.1, we start by summarising our findings on the main characteristics of diachronic and historical corpora, and how these could be taken into consideration in a Swedish diachronic corpus. In Section 4.2, we discuss metadata information and representation. Finally, some directions for future work are given in Section 4.3.

4.1 Characteristics of historical corpora

Table 1 gives an overview of some of the main characteristics of the diachronic and historical corpora decribed in this report.

As seen from the second column in the table, there are considerable differences in the size of the corpora, from 53,000 words in the Faroese Parsed Historical Corpus (FarPaHC), over 400 million words in the Corpus of Historical American English (COHA), to many billions of words in the Google Books Ngram Corpus. Large diachronic datasets are of course desirable for studies on linguistic change over time. On the other hand, it would be very time-consuming and resource-demanding to manually perform linguistic annotation for millions of words. Thus, the larger corpora are typically either not annotated at all, or automatically annotated, possibly with manual correction for a subpart of the corpus. The smaller corpora, on the other hand, are usually carefully annotated by humans with features such as part-of-speech, lemma, morphological and syntactic analysis, enabling the user to formulate more advanced search queries with high-quality results. For the manually annotated corpora in our study, the size varies between 53,000 words for the previously men-

tioned FarPaHC corpus and 8 million words for the Spanish IMPACT corpus (IMP-es). One idea for the forthcoming Swedish diachronic corpus would be to provide a core corpus of manually annotated text, offering advanced, high-quality search possibilities, as part of a larger, semi-automatically annotated corpus, enabling large-scale studies on linguistic change over time. In such an approach, the core corpus should ideally include the main time periods and genres covered in the corpus as a whole.

Name	**Size**	**Period** **Granularity**	**Balance**	**Annotation**	**Access**
ARCHER	3.3M	1600–1999 50 years	per genre per period	spelling (partly)	restricted
CdE	100M	1200s–1900s centuries	1900s only 4 genres	semi-automatic POS, lemma, semantics	search
CdP	45M	1300s–1900s N/A	1900s only 4 genres	semi-automatic? POS, lemma, semantics	search
COHA	400M	1810–2009 decades	4 genres 50% fiction 50% non-fict.	semi-automatic POS, lemma, semantics	search: free pay for download
CORDE	250M	?–1974 user-def.	geography, genre, time	no	search
DIAKORP	4M	1350–1939 N/A	no	no	CC BY-NC-SA
DTA	86M	1600–1899 centuries	3 genres	automatic POS (search)	CC BY-NC
FarPaHC	53K	1823–1936 N/A	no	manual POS, syntax, spelling	LGPL
FSV	1,2M	1162–1758 N/A	no	no	free
GerManC	800K	1650–1800 50 years	8 genres 2K/sample oral & print 5 regions	automatic POS, lemma, morph, syntax, spelling	CC BY-NC-SA
Google Books Ngrams	billions (millions of books)	1500–2008 1 year	no (8 languages)	automatic POS, syntax	CC BY (n-grams)
HaCOSSA	128K	1375–1550 N/A	no	manual morph, syntax	restricted non-comm.
HGDS	3.2M	1195–1626 N/A	no	automatic POS, lemma, morph, syntax, spelling	free
IcePaHC	1M	1150–2008 centuries	5 genres 100K/century	manual POS, lemma, morph, syntax, spelling	LGPL
IMP-es	8M	1482–1990 N/A	no	manual POS, lemma, spelling (auto)	CC BY-SA
IMP-sl	300K	1584—1899 N/A	no	manual lemma, msd, spelling	CC BY

Name	Size	Period Granularity	Balance	Annotation	Access
Menota	1.6M	1200s–1500s N/A	no	manual lemma, morph, spelling (partly)	CC BY-SA
PROIEL	1.4M	Old Indo-European N/A	no	manual POS, lemma, morph, syntax	CC BY-NC-SA
PPCHE	5.7M	1150–1914 centuries	?	manual POS, syntax	pay
ReM	2M	1050–1350	no	manual POS, lemma, morph	CC BY-SA
RIDGES	3M	1482–1914 N/A	science only	automatic POS, lemma, morph, syntax, spelling	CC BY
Språkbanken's historical corpora	1.3G (soon 6G)	1225–1926 N/A	no	automatic/ none varying	search (download: CC BY)
SRCMF	251K	842–1278 N/A	no	manual POS, lemma, syntax	CC BY-NC-SA
Swedish Culturomics Gigaword corpus	1G	1950–2015 1 year	5 genres	automatic POS, lemma morph, syntax, word senses	CC BY (sentence scrambled)
TBCHP	3M	1380–1881 50 years	7 genres	manual POS, syntax, spelling	research

Table 1: Characteristics of existing diachronic and historical corpora.

Apart from corpus size, there is also a noticeable difference in the time periods covered by the corpora in our study. This feature naturally relates to the purpose of the corpus in question. The Syntactic Reference Corpus of Medieval French (SRCMF) for example, clearly aims to cover the medieval period, whereas a corpus such as the Icelandic Parsed Historical Corpus (IcePaHC) intends to represent several stages of the Icelandic language, including present-day Icelandic. In the Swedish case, we aim for the latter corpus type, presumably starting from the Old Swedish time period (≈1200s–1500s), or possibly even including the Old Norse period (≈800s–1200s) (Bergman 1995). Before establishing the exact time boundaries of the Swedish diachronic corpus we will carefully investigate existing text resources for different stages of the Swedish language development, as well as the needs of the intended corpus users.

A related question to take into consideration when building a diachronic corpus is what subcorpora to include, concerning time periods (granularity) and

genres (balance and representativeness). Optimally, the corpus should be as fine-grained as possible, enabling the user to compare results for different centuries, 50-year periods, or even decades. In order for these comparisons to be truly comparable though, the composition of texts from different genres should then be equally distributed for the periods compared. Otherwise, results from studies purportedly investigating language change on the basis of such datasets might indicate differences between genres, rather than differences between certain time periods.

The COHA corpus contains four genres (fiction, popular magazines, newspapers, and non-fiction books), and is divided into decades, with all four genres represented for (nearly) every decade, making the corpus a good example of a balanced corpus. This corpus does however only cover a relatively short period of time (1810–2009). A problem when trying to cover several stages of a language throughout history, is that it might be hard to find the same kinds of text for all time periods. Newspapers, for example, did not exist in the older time periods. One way to overcome this could be to divide the texts into broader domains, such as fiction compared to non-fiction texts (as is also done in the COHA corpus), or to try to find at least one genre that exists for many time periods, e.g. legal texts or church documents. For the corpus to be representative of the language at a given point in time, several other genres should however be included as well, to reflect the actual text production during that period.

In addition to comparisons between decades etc., one could also imagine comparisons between linguistically motivated time periods, such as Old Swedish, Early Modern Swedish etc. Input from the intended users of the corpus, as well as an investigation into available text resources, will be needed in order to decide on the most useful and well-motivated divisions into time periods and genres in the Swedish diachronic corpus.

Concerning linguistic annotation, it has already been mentioned that some of the corpora in this study were manually annotated, whereas other were automatically or semi-automatically annotated, and that one idea for the Swedish corpus could be to have a manually annotated core corpus, and an unannotated, or automatically annotated, full corpus. One thing that is yet to be discussed is however what levels of annotation that would be most useful and doable to include in the corpus.

Most corpora in Table 1 include at least part-of-speech annotation and lemmatisation. These features are also in line with the levels of annotation listed as points 1–3 by Davies (2010b) as criteria in order for historical corpora to be useful for studying a wide range of linguistic phenomena (see further Section 3). One could thus conclude that this would be the minimum level of annota-

tion required for the forthcoming Swedish diachronic corpus. To be able to do more advanced studies on morphological and syntactic change, it would also be desirable to include more sophisticated morphological analysis than part-of-speech tagging, and also to add some level of syntactic annotation. Then the question arises what formalism to use for syntactic annotation; phrase structure grammar or dependency relations? In the context of diachronic corpora, it could be argued that there is a considerable amount of historical and diachronic corpora following the Penn-Helsinki phrase structure annotation scheme, see further Section 2.12. Choosing this formula would thus make results from the Swedish diachronic corpus easily comparable to results from any of these corpora. On the other hand, many corpora being developed today, both historical and present-day language corpora, follow the dependency grammar formalism. One advantage with the dependency grammar formalism is the *Universal Dependencies* framework,[57] aiming to provide a cross-linguistically consistent tagset applicable to all languages, facilitating comparisons between different languages.

Davies (2010b) also points out that semantic search is an important feature for studies on how the meaning of words and phrases has changed over time. To include information on synonyms and semantic fields in the corpus would be a quite advanced and costly annotation procedure. However, including a search for collocates of a given word or phrase would be more easily doable, and probably also very useful in this context, since words that co-occur more often than would be expected by chance could give a good indication of the meaning of a word or a phrase. Recently, so-called word embeddings have been applied to corpus-based studies of semantic change (Tahmasebi, Borin and Jatowt 2019).

A form of annotation that is specifically relevant to the historical domain is *spelling harmonization*, i.e. the process of mapping the historical spelling to some standard spelling variant (typically the modern spelling of a word). Spelling harmonization is a useful feature for a diachronic corpus in (at least) two ways. First of all, even if the corpus is to be manually annotated, this process typically includes a pre-processing step in the form of automatic linguistic annotation; the results of which are then manually revised and corrected. Since natural language processing (NLP) tools adapted to analysing historical text are rarely available, tools developed for present-day language may be used for processing historical text as well. However, NLP tools trained on present-day language are often confused by the varying and non-standardized spelling in historical texts, causing low-quality output. Studies have shown that transforming the spelling

[57] https://universaldependencies.org/

to a more modern spelling before applying the NLP tools has a significant positive effect on the results (Pettersson 2016). Secondly, spelling harmonization is useful for making the text more easily searchable by the user. If no harmonization is performed, the user needs to come up with all the different spellings of a word to search for it. If spelling harmonization is applied, the user may instead search for the standardised word form only, yielding results for the word in all its spelling variants.

The last column in Table 1 shows the licence for accessing the corpus. Some corpora are only searchable, without download possibilities. Others are downloadable, either freely available or only available for research, typically under some form of Creative Commons licence. In the case of the Swedish diachronic corpus, we aim for a freely accessible corpus, offering both a search interface and download possibilites, with the only possible exception for present-day copyright-protected material.

4.2 Metadata information and representation

An important issue to consider when building a corpus is what metadata information to include, and what standard to use for storing metadata. Table 2 gives and overview of some metadata information included for the corpora studied in this report.

Name	Title	Author	Year	Edition	Genre	Sub	#Words	Edited
ARCHER	x	x	x	–	x	–	x	-
COHA	x	x	x	x	x	x	x	–
CORDE	x	x	x	x	x	x	–	–
DIAKORP	x	x	x	–	x	x	–	–
DTA	x	x	x	x	–	–	x	–
FarPaHC	x	x	x	x	x	–	x	x
FSV	x	partly	x	x	x	–	-	x
GerManC	x	x	x	–	x	–	–	–
HGDS	x	x	x	x	–	–	–	x
IcePaHC	x	x	x	x	x	–	x	–
IMP-es	x	x	x	x	–	–	–	–
IMP-sl	x	x	x	–	x	–	–	x
Menota	x	x	x	x	–	–	x	x
PPCHE	x	x	x	x	x	?	x	?
ReM	x	x	x	x	x	x	–	–
PROIEL	x	–	–	x	–	–	–	–
RIDGES	x	x	x	–	x	x	–	–
SB-Korp	partly	partly	partly	–	partly	–	x	–
SRCMF	x	x	x	x	x	–	x	–
SwCGW	partly	partly	x	–	x	–	x	–
TBCHP	x	x	x	–	–	–	x	x

Table 2: Metadata included in existing diachronic and historical corpora.

As seen from the table, information on title, author and publication year is (al-

most) always provided (with the few exceptions being very old texts, where these features may be unknown). Since historical texts may occur in many versions, it is also important to state which edition the text represents. The genre, and possible sub-genre, of the text is another important feature to state in the metadata field, especially when aiming for a multi-genre corpus. Another common piece of information to provide is the number of words (and possibly also characters and bytes) in the text. The last column, "Edited", refers to information on edits that have been done in the text during transcription. This includes more formal edits like the representation of characters not included in the Unicode scheme, "correction" of line breaks inside words etc., as well as more advanced edits, such as spelling harmonization. Apart from the features listed in Table 2, some of the corpora studied also contain metadata information on features such as publisher, editor, transcriber, annotator, volume, issue, language variant (Early New Modern etc.), region in which the text was produced, availability (licence etc.), extent of the sample (if not the full text), levels of linguistic annotation and general notes. Ideally, it would of course be desirable to include as detailed metadata information as possible, to facilitate for the users of the corpus to identify texts of interest to them. However, especially in the case of old texts, even basic information such as author or publication year may be missing or unreliable, meaning that the level of metadata information available may vary greatly between different texts. Thus, for some texts it might only be possible to include limited and less reliable metadata information.

In addition to the metadata contents, it is also important to consider how the metadata should be represented. The most common formats in the corpora studied here are:

1. plaintext format (typically as headers preceding each text, with a standardised terminology), or
2. tab-separated CoNLL format, or
3. TEI standard (XML), or
4. a table listing all the texts and their metadata contents, typically in xls format, or as an HTML table

For the Swedish diachronic corpus, it could be an option to provide all four formats. If metadata is originally given as headers in plaintext files, in a standardised format following a specific terminology, this format could easily be mapped to a CoNLL or TEI format, and also exported as a csv file to be readable by Excel and similar programs.

4.3 FUTURE WORK

The next step in preparing the Swedish diachronic corpus is to investigate the needs and requirements of the intended users of the diachronic corpus, typically researchers in the humanities, and to make an inventory of available Swedish text resources that could be included in the corpus. Based on the findings from those studies, in combination with the results presented in this report, we will then decide on the contents and structure of the forthcoming Swedish diachronic corpus, by carefully considering what time period to cover, how to divide the corpus into genres and sub-periods, and what levels of linguistic annotation and metadata information to include, as well as what annotation and metadata standards to use.

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*. Aachen: CEUR-WS.org. to appear.

Aiden, Erez and Jean-Baptiste Michel. 2013. *Uncharted: Big data as a lens on human culture*. New York: Riverhead Books.

Baron, Alistair and Paul Rayson. 2008. VARD2: A tool for dealing with spelling variation in historical corpora. *Proceedings of the postgraduate conference in corpus linguistics*.

Bech, Kristin and Kristine Eide. 2014. The ISWOC corpus. http://iswoc.github.io. Department of Literature, Area Studies and European Languages, University of Oslo.

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

Biber, Douglas, Edward Finegan and Dwight Atkinson. 1994. ARCHER and its challenges: Compiling and exploring a representative corpus of historical English registers. *Creating and using English language corpora. papers from the 14th international conference on English language research on computerized corpora, 1993*, 1–13.

Bohnet, Bernd. 2010. Top accuracy and fast dependency parsing is not a contradiction. *Proceedings of COLING 2010*, 89–97. Beijing: ACL.

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

CLUL. 2014. P.S. Post Scriptum. Arquivo Digital de Escrita Quotidiana em Portugal e Espanha na Época Moderna. Online publication, http://ps.clul.ul.pt.

Davies, Mark. 2010a. The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing* 25 (4): 447–464.

Davies, Mark. 2010b. Creating useful historical corpora: a comparison of CORDE, the Corpus del Español, and the Corpus do Português. *Diacronía de las lenguas iberorromances: nuevas perspectivas desde la lingüística de corpus*, pp. 137–166.

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.

Eckhoff, Hanne, Kristin Bech, Gerlof Bouma, Kristine Eide, Dag Haug, Odd Einar Haugen and Marius Jøhndal. 2018. The PROIEL treebank family: a standard for early attestations of Indo-European languages. *Language Resources and Evaluation* 52 (1): 29–65.

Eckhoff, Hanne Martine and Aleksandrs Berdicevskis. 2015. Linguistics vs. digital editions: The Tromsø Old Russian and OCS treebank. *Scripta & e-Scripta* 14–15: 9–25.

Eide, Stian Rødven, Nina Tahmasebi and Lars Borin. 2016. The Swedish Culturomics Gigaword corpus: A one billion word Swedish reference dataset. *From digitization to knowledge 2016: Resources and methods for semantic processing of digital works/texts*, 8–12. Linköping: LiUEP.

Erjavec, Tomaž. 2011. Automatic linguistic annotation of historical language: ToTrTaLe and XIX century Slovene. *Proceedings of LaTeCH 2011*, 33–38. Portland, Oregon: ACL.

Erjavec, Tomaž. 2012. The goo300k corpus of historical Slovene. *Proceedings of LREC 2012*, 2257–2260. Istanbul: ELRA.

Galves, Charlotte and Pablo Faria. 2017. Tycho Brahe Parsed Corpus of historical Portuguese. http://www.tycho.iel.unicamp.br/corpus/en/.

Geyken, Alexander, Susanne Haaf, Bryan Jurish, Matthias Schulz, Jakob Steinmann, Christian Thomas and Frank Wiegand. 2010. Das Deutsche Textarchiv: Vom historischen Korpus zum aktiven Archiv. *Digitale Wissenschaft*, pp. 157–161.

Gippert, Jost and Manana Tandashvili. 2015. Structuring a diachronic corpus: The Georgian National Corpus project. Jost Gippert and Ralf Gehrke (eds), *Historical corpora: Challenges and perspectives*, 305–322. Tübingen: Narr.

Guillot, Céline, Christiane Marchello-Nizia and Alexeij Lavrentiev. 2007. La Base de Français Médiéval (BFM): états et perspectives. Pierre Kunstmann and Achim Stein (eds), *Le Nouveau Corpus d'Amsterdam. Actes de l'atelier de Lauterbad, 23–26 février 2006*. Stuttgart: Steiner.

Haug, Dag T. T. and Marius L. Jøhndal. 2008. Creating a parallel treebank of the Old Indo-European bible translations. *Proceedings of LaTeCH 2008*, 27–34. Marrakech: ELRA.

Höder, Steffen. 2011. 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.

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

Jurish, Bryan. 2012. Finite-state canonicalization techniques for historical German. Ph.D. diss., Humanwissenschaftliche Fakultätder Universität Potsdam.

Kenter, Tom, Tomaž Erjavec, Maja Žorga Dulmin and Darja Fišer. 2012. Lexicon construction and corpus annotation of historical language with the CoBaLT editor. *Proceedings of LaTeCH 2012*, 1–6. Avignon: ACL.

Klein, Thomas and Stefanie Dipper. 2016. Handbuch zum Referenzkorpus Mittelhochdeutsch. *Bochumer Linguistische Arbeitsberichte*, vol. 19.

Krause, Thomas and Amir Zeldes. 2016. ANNIS3: A new architecture for generic corpus query and visualization. *Digital Scholarship in the Humanities* 31, no. 1.

Kroch, Anthony, Beatrice Santorini and Lauren Delfs. 2004. Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME). CD-ROM, first edition, release 3.

Kroch, Anthony, Beatrice Santorini and Ariel Diertani. 2016. The Penn Parsed Corpus of Modern British English (PPCMBE2). CD-ROM, second edition, release 1.

Kroch, Anthony and Ann Taylor. 2000. The Penn-Helsinki Parsed Corpus of Middle English (PPCME2). CD-ROM, second edition, release 4.

Kunstmann, Pierre and Achim Stein. 2007. Le Nouveau Corpus d'Amsterdam. Pierre Kunstmann and Achim Stein (eds), *Le Nouveau*

Corpus d'Amsterdam. Actes de l'atelier de Lauterbad, 23–26 février 2006, 9–27. Stuttgart: Steiner.

Kučera, K., A. Řehořková and M. Stluka. 2015. DIAKORP: diachronic corpus of Czech, version 6. Institute of the Czech National Corpus, Faculty of Arts, Charles University, Prague. http://www.korpus.cz/.

Leech, Geoffrey, Roger Garside and Michael Bryant. 1994. CLAWS4: The tagging of the British National Corpus. *Proceedings of COLING 1994*, 622–628. Kyoto: ACL.

Lin, Yuri, Jean-Baptiste Michel, Erez Lieberman Aiden, Jon Orwant, William Brockman and Slav Petrov. 2012. Syntactic annotations for the Google Books Ngram corpus. *Proceedings of ACL 2012*, 169–174. Jeju Island: ACL.

Loftsson, Hrafn and Eiríkur Rögnvaldsson. . IceNLP: A natural language processing toolkit for Icelandic.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak and Erez Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science* 331 (6014): 176–182.

Odebrecht, Carolin, Malte Belz, Amir Zeldes, Anke Lüdeling and Thomas Krause. 2017. RIDGES Herbology: designing a diachronic multi-layer corpus. *Language Resources and Evaluation* 51 (3): 695–725.

Oxford-NINJAL 2018. Oxford-NINJAL Corpus of Old Japanese (Version 2018.9). Online publication, http://oncoj.ninjal.ac.jp/, accessed 5 February 2019.

Pettersson, Eva. 2016. Spelling normalisation and linguistic analysis of historical text for information extraction. Ph.D. diss., Department of Linguistics and Philology, Uppsala University, Uppsala.

Pintzuk, Susan and Leendert Plug. 2002. The York-Helsinki Parsed Corpus of Old English Poetry. Oxford Text Archive, first edition, http://www-users.york.ac.uk/~lang18/pcorpus.html.

Prévost, Sophie and Achim Stein. (eds) 2013. *Syntactic reference corpus of medieval French (SRCMF)*. Version 0.92. Lyon/Stuttgart: ENS de Lyon; Lattice, Paris; ILR University of Stuttgart.

Randall, Beth. 2005. *CorpusSearch 2 users guide*. http://corpussearch.sourceforge.net/CS-manual/Contents.html: University of Pennsylvania, Philadelphia.

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.

Schmid, Helmut. 1994. Probabilistic part-of-speech tagging using decision trees. *International Conference on New Methods in Language Processing*, 44–49.

Simon, Eszter. 2014. Corpus building from Old Hungarian codices. Katalin É Kiss (ed.), *The evolution of functional left peripheries in Hungarian syntax*, 224–236. Oxford: Oxford University Press.

Stein, Achim. 2019. Diachronic syntax based on constituency and dependency annotated corpora. *Linguistic Variation* 18 (1): 74–99.

Sánchez-Martínez, F., I. Martínez-Sempere, X. Ivars-Ribes and R.C. Carrasco. 2013. An open diachronic corpus of historical Spanish. *Language Resources and Evaluation* 47 (4): 1327–1342. https://link.springer.com/article/10.1007%2Fs10579-013-9239-y.

Tahmasebi, Nina, Lars Borin, Gabriele Capannini, Devdatt Dubhashi, Peter Exner, Markus Forsberg, Gerhard Gossen, Fredrik Johansson, Richard Johansson, Mikael Kågebäck, Olof Mogren, Pierre Nugues and Thomas Risse. 2015. Visions and open challenges for a knowledge-based culturomics. *International Journal on Digital Libraries* 15 (2–4): 169–187.

Tahmasebi, Nina, Lars Borin and Adam Jatowt. 2019. Survey of computational approaches to lexical semantic change. *arXiv.org*, no. 1811.06278. https://arxiv.org/abs/1811.06278.

Taylor, Ann, Arja Nurmi, Anthony Warner, Susan Pintzuk and Terttu Nevalainen. 2006. The York-Helsinki Parsed Corpus of Early English Correspondence (PCEEC). Oxford Text Archive, first edition, http://www-users.york.ac.uk/~lang22/PCEEC-manual/index.htm.

Taylor, Ann, Anthony Warner, Susan Pintzuk and Frank Beths. 2003. The York-Toronto-Helsinki Parsed Corpus of Old English Prose (YCOE). Oxford Text Archive, first edition, http://www-users.york.ac.uk/~lang22/YcoeHome1.htm.