The HistCorp Collection of Historical Corpora and Resources

Eva Pettersson and Beáta Megyesi

Uppsala University

Abstract. We present the HistCorp collection, a freely available open platform aiming at the distribution of a wide range of historical corpora and other useful resources and tools for researchers and scholars interested in the study of historical texts. The platform contains a monitoring corpus of historical texts from various time periods and genres for 14 European languages. The collection is taken from well-documented historical corpora, and distributed in a uniform, standardised format. The texts are downloadable as plaintext, and in a tokenised format. Furthermore, a subset of the corpus contains information on the modern spelling variant, and some of the texts are also annotated with part-of-speech and syntactic structure. In addition, preconfigured n-gram language models and spelling normalisation tools are provided to allow the study of historical languages.

Keywords: digital humanities, historical corpora, language models, spelling normalisation, HistCorp

1 Introduction

Historical text, and tools for processing historical text, are of great interest to historians, literary scholars and other researchers in humanities, as well as for researchers in computational linguistics and digital humanities. However, corpora containing historical text are not easily found. Furthermore, since there is no well-established standard for the format of historical corpora, the corpora that do exist are often available in different formats, that are not always well documented. Thus, it might be time-consuming and hard for the user to extract the adequate information from these corpora. In addition, copyright issues and terms of use are often unclear. There are also many cases where the corpus at hand is only available via a web-based search interface, giving no possibility to download the actual corpus in a machine-readable format, which might be desirable for the research topic in question.

Another useful resource in corpus studies is *language models*, representing the probability for certain combinations of words (or letters) to occur in a sequence in a specific language (or sublanguage). For example, the sequence *he is* would have a significantly higher probability to occur in the English language than

the sequence *he are*. These statistics are computed automatically from corpora of any kind. Within the domain of historical text, language models could be created for texts from different time periods, providing clues on how language has changed over time; information that could be of interest to for example historical linguists.

Furthermore, searching historical text for particular words and/or structures is trickier than searching modern text, mainly due to the peculiar and variable spelling in historical text. Therefore, natural language processing tools especially designed for handling historical text might be useful as an aid in the information extraction phase, for example *spelling normalisation tools* (as described in more detail in Section 5).

In this paper, we present the HistCorp collection of historical corpora and resources: `http://stp.lingfil.uu.se/histcorp/`. The aim of our work is to provide a platform for researchers working with historical text, where we gather a wide range of historical corpora and other useful resources and tools in one place, available for download in a well-defined and uniform format. The HistCorp platform consists of three entries:

1. Historical Corpora
   Historical corpora are available for download in a plain text format and in a tokenised format (separated into words and sentences). For some corpora there are also other formats available for download, containing information on for example modernised spelling, or morphological and syntactic annotation. See further Section 3.

2. Language Models
   The user may download preconfigured n-gram language models built on the corpora available on the HistCorp platform. There is also a possibility for the user to upload his/her own text files to create a language model based on these specific files. See further Section 4.

3. Tools
   The Tools part of the HistCorp platform provides tools for facilitating the process of analysing historical text. See further Section 5.

2 The Decode Project

The HistCorp platform was created within the Decode project[1], a project aiming at the development of computer-aided tools for (semi-)automatic decoding of enciphered historical manuscripts, so called ciphertexts. Ciphertexts are important historical records, encrypted to keep the content of the message hidden from others than the intended receiver(s). Examples of such material include diplomatic correspondence, intelligence reports, alchemistic and scientific writings,

[1] `http://stp.lingfil.uu.se/~bea/decode/`

private letters, diaries, and sources related to secret societies. There are thousands of such historical texts in archives and libraries, waiting to be discovered and decrypted.

Within the rather young and highly interdisciplinary field of historical cryptology, researchers from areas such as language technology, computer science, history, linguistics and philology make efforts to develop an infrastructure to be able to systematically decode ciphertexts. In order to develop algorithms for (semi-)automatic decryption of historical ciphers, there is a need for all three aspects of the HistCorp platform. Historical corpora reflecting the language of the time, as well as language models derived from these resources, are useful for language identification and more well-informed hypotheses and guesses in the decryption phase. In addition, natural language processing tools based on methods such as spelling normalisation techniques could be used for mapping several word forms with different spellings to the same normalised form, facilitating the decoding process, as well as the analysis of the decoded text.

3 Historical Corpora

The corpus part of the HistCorp platform is to be considered as a *monitoring corpus*, intended to grow over time, as more texts and more languages are added. The aim is to collect *diplomatic transcriptions* of historical text, that is texts with minimal or no editorial intervention in the digitization process. Another quality aspect is that the included texts are taken from well-documented corpus collections, rather than just sampling historical texts from the Internet. In addition, the texts included on the HistCorp platform should be free for download.[2] Additional terms of use (e.g., regarding redistribution, commercial use etc.) are clearly stated in the licence information for each corpus. As mentioned above, the corpus collection is not static, meaning that more corpora will be added over time. At the time of writing, there are corpora for 14 languages available on the HistCorp platform, as presented in Table 1.

[2] There are a few exceptions to this, but then the procedure for getting access to the corpus is clearly stated in the README file for that particular corpus.

Czech	
1) Diakorp: the diachronic section of the Czech National Corpus (Kučera and Stluka 2011)	1350–1939
2) Gutenberg texts (`http://www.gutenberg.org/`)	1890–1897
Dutch	
1) Compilation Corpus Historical Dutch (Coussé 2010)	1250–2000
2) Gutenberg texts (`http://www.gutenberg.org/`)	1400–1875
English	
Lampeter Corpus of Early Modern English Tracts (`http://ota.ox.ac.uk/desc/3193`)	1600–1800
French	
Paris Speech in the Past (`http://ota.ox.ac.uk/desc/2423`)	1296–1790
German	
1) Deutsches TextArchiv (Geyken et al. 2010)	1600–1899
2) GerManC (Durrell et al. 2007)	1650–1800
3) Reference Corpus of Middle High German (Klein and Dipper 2016)	1050–1350
Greek	
Ancient Greek Dependency Treebank	not stated
Hungarian	
Hungarian Generative Diachronic Syntax (Simon to appear)	1440–1539
Icelandic	
Icelandic Parsed Historical Corpus (Rögnvaldsson et al. 2012)	1150–2008
Italian	
Gutenberg texts (`http://www.gutenberg.org/`)	1300–1897
Latin	
Ancient Latin Dependency Treebank	not stated
Portuguese	
Tycho Brahe Parsed Corpus of Historical Portuguese (Galves and Faria 2010)	1380–1881
Slovene	
Historical-to-modern dictionaries as described in Scherrer and Erjavec (2016)	1750–1900
Spanish	
IMPACT-es diachronic corpus (Sánchez-Martínez et al. 2013)	1481–1962
Swedish	
1) Fornsvenska Textbanken (Delsing 2002)	1350–1758
2) Gender and Work (Ågren et al. 2011)	1527–1812
3) Gutenberg texts (`http://www.gutenberg.org/`)	1789–1902
4) Protocols from the Academic Consistory of Uppsala University 1968	1624–1699

Table 1. Languages and corpora currently available on the HistCorp platform. The first column lists the name of the corpus, whereas the second column states the time period in which the texts in each corpus were written.

3.1 Corpus Format

Since the corpora on the HistCorp platform are collected from many different sources, the source format and annotation level differ between the corpora. First, the text encoding for representing language specific characters is not the same for all source corpora. In addition, some corpora are only available in a plain text format, whereas others are annotated with for example morphological and/or syntactic information. For annotated corpora, the linguistic information may be represented in a tab-separated format, in an XML-based format, in a parenthesis-based format etc. Apart from the representation of the text content and the linguistic annotation, metadata (such as time period, author, genre etc.) may also be given in many shifting formats.

One aim with the HistCorp platform is to provide corpora in a uniform and well-documented format. Therefore, every corpus is converted to the Unicode UTF-8 encoding scheme[3], and standardised into a plain text format with metadata stated in a TEI-compatible format at the top of each file. Furthermore, each corpus is also provided in a uniform tokenised format, with one token on each line, and blank lines separating each sentence. Since the metadata information, as well as the plain text format and the tokenisation format, are the same for all corpora, it will be possible for the user to apply the same tools for metadata extraction and linguistic processing of all corpora. Regarding morphological and syntactic annotation, it is a much trickier task to standardise this information over all corpora. Thus, the source format for representing these features are currently kept unchanged.

All in all, each corpus on the HistCorp platform is possibly available in five different download formats:

1. A plain text format, with metadata stated at the top of each file, in a TEI-compatible format (see further Section 3.3). For corpora that were originally only available in an XML format, the text elements are automatically extracted from the XML file, to create a plain text file.

2. A tokenised format, with one token on each line, and a blank line separating sentences, as illustrated in the following example taken from the Gender and Work corpus (Ågren et al. 2011):

```
Nils
Olofsson
i
Ås
ähr
af
Daniel
Mårtensson
på
```

[3] http://www.unicode.org/

```
Åsen
stämbd
till
tingz
```

For corpora originally lacking tokenisation, UDPipe (Straka and Straková 2017) is used with the CoNLL 2017 Shared Task baseline models (Straka 2017), to create the tokenised version of the corpus.

3. A format containing information on historical and modern spelling of the text. This applies to corpora where all or some of the words are available both in their original spelling, and in a manually modernised spelling. This kind of annotation is presented in a format with one token pair on each line, with a tab separating the historical spelling from the modernised spelling, as illustrated in the following example taken from the the Gender and Work corpus (Ågren et al. 2011):

```
Nils        Nils
Olofsson    Olofsson
i           i
Ås          Ås
ähr         är
af          av
Daniel      Daniel
Mårtensson  Mårtensson
på          på
Åsen        Åsen
stämbd      stämd
till        till
tingz       tings
```

4. A morphologically annotated format. This applies only to corpora where morphological information is available in the original source corpus. The original annotation format is then kept in the HistCorp download as well, as illustrated in the following example taken from the Tycho Brahe Parsed Corpus of Historical Portuguese (Galves and Faria 2010):

```
O/D amor/N puro/ADJ ,/, belíssima/ADJ-S-F Genoveva/NPR ,/,
é/SR-P muito/Q raro/ADJ ./.
```

5. A syntactically annotated format. This applies only to corpora where syntactic information is available in the original source corpus. The original annotation format is then kept in the HistCorp download as well, as illustrated in the following example taken from the Tycho Brahe Parsed Corpus of Historical Portuguese (Galves and Faria 2010):

```
((IP-MAT (NP-SBJ (D O)
 (N amor)
 (ADJP (ADJ puro)))
 (, ,)
 (NP-VOC (ADJ-S-F belíssima) (NPR Genoveva))
 (, ,)
 (SR-P é)
 (ADJP (Q muito) (ADJ raro))
 (. .))
```

3.2 Presentation Format

On the HistCorp platform, each corpus is presented in 12 columns, as illustrated in Figure 1. The first column states the name of the corpus, whereas the second and third column gives information on the time period and genres contained in the corpus. Furthermore, there are five download columns, corresponding to the different annotation levels described in Section 3.1, that is: a) plain text format, b) tokenised format, c) spelling normalisation, d) morphological annotation, and e) syntactic annotation. The ninth column contains a link to a short README file for the corpus, with information on the contents and format of the corpus. The tenth column provides the possibility to download all the corpora files in one go, instead of downloading only the plain text format or the morphologically annotated file etc. The eleventh column contains a link to the source page, from which the corpus was originally downloaded, whereas the last column states the terms of use for each corpus, typically in the form of a link to the specific licence adhering to the corpus in question.

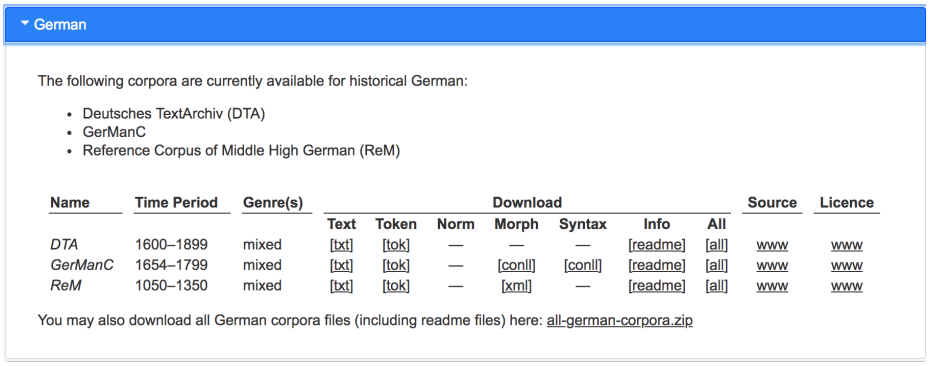

Fig. 1. Download format for historical corpora.

In the specific example of the German language, illustrated in Figure 1, it could be noted that the *Deutsches TextArchiv* corpus contains texts from the time

period 1600–1899, and that the texts are available in a plain text format and in a tokenised format. The *GerManC corpus* on the other hand contains texts from the time period 1654–1799, and is available on all annotation levels, except for the spelling normalisation level. The third corpus, *Reference Corpus of Middle High German*, contains older texts (1050–1350) and is available in a plain text format, in a tokenised format, and in a morphologically analysed format.

3.3 Metadata

Many corpora contain extra-textual information (in the following referred to as *metadata*), such as the title of the text, the name of the author, the year in which the text was written and so on. This is very useful information, but the way this information is structured often differs between different source corpora, making it hard for the user to extract the relevant information. Therefore, in the HistCorp files, the metadata information has been converted to a consistent format that is identical for all corpora, regardless of the original source corpus format. The metadata information is stored at the top of each plain text file, with a hash sign (#) preceding each new piece of metadata information, as illustrated in Figure 2, showing the metadata information available for one of the texts in the diachronic section of the *Czech National Corpus* (Kučera and Stluka 2011). Some metadata are extracted from the metadata information given in the source corpora, whereas some metadata have been added during the creation of the HistCorp files (for example the number of tokens).

```
#title: Obrazy ze života mého, Marinka
#author: Karel Hynek Mácha
#distributor: Distributed within the Diakorp project, see further the
project webpage at: https://wiki.korpus.cz/doku.php/en:cnk:diakorp.
#availability: The data are licenced under the CC BY-NC-SA license,
http://creativecommons.org/licenses/by-nc-sa/4.0/.
#sourceDesc: Part of the diachronic section of the Czech National Corpus.
#extent tokens: 5,253
#extent documents: 1
#normalization: diplomatic
#language: Czech
#date: 1834--1835
#domain: prose
```

Fig. 2. Metadata representation in the HistCorp files.

To make the metadata information compatible with metadata contained in other corpora, we use the Text Encoding Initiative (TEI) standard for naming the metadata elements[4], but in the simplified format illustrated in Figure 2 instead

[4] http://www.tei-c.org/

of the full XML format usually associated with the TEI standard. This way, the metadata information is more comprehensible for the human user, while at the same time offering a format that is straightforward to convert to the traditional TEI XML format if needed. The most common metadata elements occurring in the HistCorp files are:

- AUTHOR
  Name of the author of the text.
- AVAILABILITY
  Terms of use for the corpus, typically with a link to the actual licence.
- DATE
  The year or time period during which the text was written.
- DISTRIBUTOR
  Information on the person or organisation providing the source corpus.
- DOMAIN
  Information on the genre of the text.
- EXTENT
  Divided into EXTENT DOCUMENTS, for the number of subtexts within the text, and EXTENT TOKENS, for the number of tokens (words and punctuations) in the text. The number of tokens are calculated using the Unix command `wc -w`.
- LANGUAGE
  The language in which the text is written.
- SOURCEDESC
  A short description of the contents of the text.
- TITLE
  The title of the text, possibly divided into TITLE MAIN for the main title, and TITLE SUB for the subheading.

4 Language Models

On the Language Models part of the HistCorp platform, the user may download preconfigured n-gram language models based on the corpora in the HistCorp collection of texts. The language models contain statistical information on the occurrence of sequences of words and characters in a language or a sublanguage, and are useful for many purposes, such as language identification and research on language change.

The language models provided on the HistCorp platform were created using the IRSTLM toolkit[5] (Federico et al. 2008) with n-gram size 5 for character-based language models, and n-gram size 3 for word-based language models. The HistCorp language models are typically divided into centuries, as illustrated in Figure 3, showing the download format for the Swedish language models.

[5] `https://github.com/irstlm-team/irstlm`

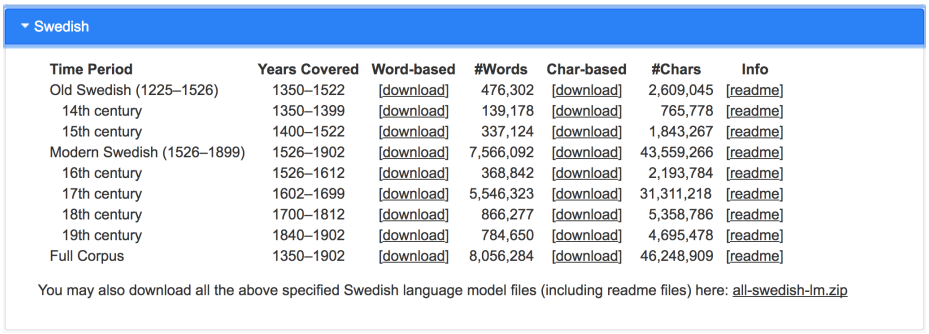

Fig. 3. Download format for historical language models.

As seen from the Figure, the user may choose to download a *word-based* or a *character-based* language model for the time period of interest. The table also provides information on the number of words or characters covered by each language model. The last column of the download table contains a link to a short README file, describing the contents of the language model, and how it was created. Below the table is a link for downloading a language model containing all the texts for the language in question, instead of downloading separate language models for each time period.

Apart from downloading preconfigured language models, the user may also upload one or more text files to the HistCorp platform, to generate his/her own language model, as illustrated in Figure 4.

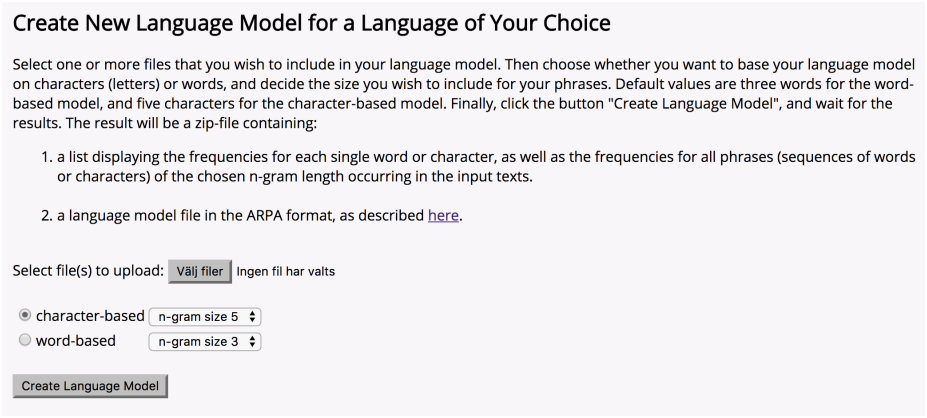

Fig. 4. The HistCorp interface for creating new language models.

This gives the user the possibility to upload files not contained in the HistCorp collection of texts, and to define his/her own criteria for time periods and genres included in the resulting language model. The default values for how many sequential units to include in the language models are three words for the word-based models, and five characters for the character-based model, but the user has the possibility to choose a value between one and ten for both types of language models.

5 Tools

The Tools section of the HistCorp platform provides the possibility to download tools especially designed for processing historical text. Here, the HistNorm package for automatic spelling normalisation of historical text is available for download (Pettersson et al. 2014). *Spelling normalisation* is the process of automatic modernisation of the spelling in historical text, a method that could be used for several purposes. One is to facilitate search in historical text, since the user then can search for the standardised spelling of a word, and find all occurrences of that word form in the text, regardless of spelling variation. Searching for the same word form in the text in its original spelling on the other hand, would mean that the user would have to guess what different spelling variants to enter for that particular word form. Another benefit of spelling normalisation, is that natural language processing (NLP) tools such as taggers for morphological analysis and parsers for syntactic analysis are generally trained on modern language texts, and do not perform well on texts with the variable historical spelling. However, studies have shown that spelling normalisation of historical text prior to the application of taggers and parsers trained on modern language data significantly improves the performance of the NLP tools (Pettersson 2016).

In the Tools section, the user also has the possibility to download predefined datasets suitable for training spelling normalisation systems for different languages, as illustrated in Figure 5. These datasets contain both a set of historical spellings mapped to their corresponding modern spellings, and references to modern language resources that could also be useful in training a spelling normalisation tool (depending on the normalisation method chosen).

We plan to add more tools for spelling normalisation, as well as other useful tools for processing historical text, in the near future.

6 Conclusion

In this paper, we have presented the HistCorp collection consisting of historical corpora and resources for 14 languages from various time periods, as a freely available open-access resource. The historical corpora are taken from well-documented corpus collections with minimal or no editorial intervention. The corpora for various languages are provided in a uniform and well-documented format, converted to the Unicode UTF-8 encoding scheme, and standardised into

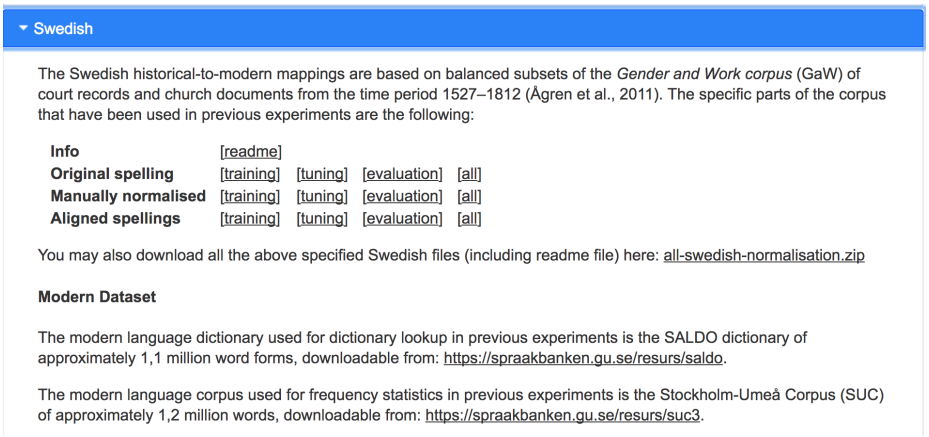

Fig. 5. Download format for historical-to-modern datasets useful for training spelling normalisation systems.

plaintext. TEI-compatible metadata is used for describing extra-textual information about the texts.

Each corpus is presented in several downloadable formats. Apart from plaintext, the texts are also available in a tokenised format. In addition, where applicable, tokens are annotated with their modern spelling as well as part-of-speech and morphological features, and the sentences are marked-up with syntactic dependencies.

The platform also provides several preconfigured n-gram language models generated from the various texts in the collection. The language models contain statistical information on co-occurrence sequences of characters and words in a particular dataset, which might be useful for historical linguistic studies of language change. The user can also create his/her own language model by uploading texts of their interest.

In addition, HistCorp contains downloadable tools for the automatic processing of historical texts including spelling normalisation to create standardised spelling across texts to facilitate search, and predefined datasets for training spelling normalisation systems for various languages. In the near future, we plan to add more historical datasets and tools to the platform.

Acknowledgement

This work has been supported by the Swedish Research Council DECODE project grant E0067801.

References

Buchholz, S. and Marsi, E.: CoNLL-X shared task on Multilingual Dependency Parsing. Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL-X) 149–164, Association for Computational Linguistic (2006).

Coussé, Evie: Een digitaal compilatiecorpus historisch Nederlands. *Lexikos* 20:123–142 (2010).

Delsing, Lars-Olof: Forsvenska textbanken. Lagman, S., Olsson, S. Ö. and Voodla, V. (ed.): *Nordistica Tartuensia 7* 149–156 (2002).

Durrell, M., Ensslin, A. and Bennett, P.: The GerManC project. *Sprache und Datenverarbeitung* 31:71–80 (2007).

Federico, M., Bertoldi, N. and Cettolo, M.: IRSTLM: an open source toolkit for handling large scale language models. Proceedings of Interspeech 1618–1621 (2008).

Galves, C. and Faria, P.: Tycho Brahe Parsed Corpus of Historical Portuguese (2010). url: http://www.tycho.iel.unicamp.br/ tycho/corpus/en/index.html.

Geyken, A., Haaf, S., Jurish, B., Schulz, M., Steinmann, J., Thomas, C. and Wiegand, F.: Das Deutsche Textarchiv: Vom historischen Korpus zum aktiven Archiv. *Digitale Wissenschaft* 157–161 (2010).

Klein T. and Dipper, S.: Handbuch zum Referenzkorpus Mittelhochdeutsch *Bochumer Linguistische Arbeitsberichte* 19 (2016).

Kučera, K. and Stluka, M.: DIAKORP: Diachronní korpus, version 5. Ústav Českého národního korpusu FF UK, Praha (2011). url: http://www.korpus.cz.

Pettersson, E., Megyesi, B. and Nivre, J.: A Multilingual Evaluation of Three Spelling Normalisation Methods for Historical Text. Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH) 32–41 (2014)

Pettersson, Eva.: Spelling Normalisation and Linguistic Analysis of Historical Text for Information Extraction. Doctoral thesis (2016).

Rögnvaldsson, E., Ingason, A. K., Sigurdsson, E. F. and Wallenberg, J.: The Icelandic Parsed Historical Corpus (IcePaHC). Proceedings of the Eighth International Conference on Language, Resources and Evaluation (LREC) 1977–1984 (2012)

Sallander, Hans: Akademiska konsistoriets protokoll. Acta Universitatis Upsaliensis, Uppsala (1968).

Sánchez-Martínez, F., Martínez-Sempere, I., Ivars-Ribes, X. and Carrasco, R.C.: An open diachronic corpus of historical Spanish. *Language Resources and Evaluation* 47(4):1327–1342 (2013).

Scherrer, Y. and Erjavec, T.: Modernising historical Slovene words. *Natural Language Engineering* 22(6):881—905 (2016).

Simon, Eszter.: Corpus building from Old Hungarian codices. *Katalin È Kiss (ed.): The Evolution of Functional Left Peripheries in Hungarian Syntax*, Oxford University Press (to appear).

Straka, M. and Straková, J.: Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies 88–99, Association for Computational Linguistics (2017).

Straka, Milan: CoNLL 2017 Shared Task - UDPipe Baseline Models and Supplementary Materials LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University (2017). url: http://hdl.handle.net/11234/1-1990.

Ågren, M., Fiebranz, R., Lindberg, E. and Lindström, J.: Making Verbs Count. The research project 'Gender and Work' and its methodology. *Scandinavian Economic History Review* 59(3):271—291 (2011).