UPPSALA
UNIVERSITET

# The HistCorp Collection of Historical Corpora and Resources

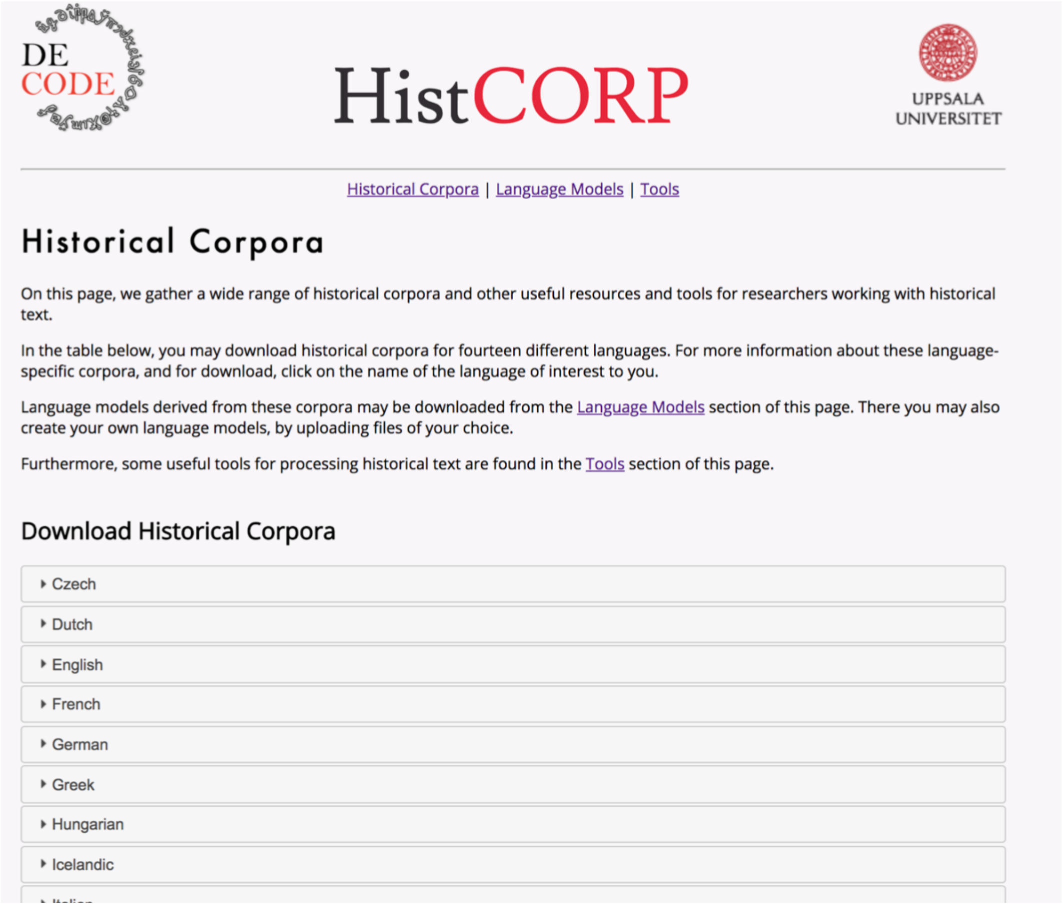
Eva Pettersson & Beáta Megyesi
Uppsala University
firstname.lastname@lingfil.uu.se

Digital Humanities in the Nordic Countries
3rd Conference
Helsinki, March 7-9, 2018

# THE HISTCORP PLATFORM

HTTP://STP.LINGFIL.UU.SE/HISTCORP/

- Freely available open platform with historical resources:
  1. Historical corpora
  2. Language models
  3. Tools



HistCORP

Historical Corpora | Language Models | Tools

## Historical Corpora

On this page, we gather a wide range of historical corpora and other useful resources and tools for researchers working with historical text.

In the table below, you may download historical corpora for fourteen different languages. For more information about these language-specific corpora, and for download, click on the name of the language of interest to you.

Language models derived from these corpora may be downloaded from the Language Models section of this page. There you may also create your own language models, by uploading files of your choice.

Furthermore, some useful tools for processing historical text are found in the Tools section of this page.

### Download Historical Corpora

- Czech
- Dutch
- English
- French
- German
- Greek
- Hungarian
- Icelandic
- Italian

# Aims and Motivation

- Corpora and tools for historical text not easy to find
- Existing historical corpora:
  - Different formats (no well-established format)
    - Text encoding issues
    - Often not possible to use the same tools for processing corpora of different formats
    - Not always a well-documented format
  - Unclear copyright and terms of use

# Aims and Motivation

- Corpora and tools for historical text not easy to find
- Existing historical corpora:
  - Different formats (no well-established format)
    - Text encoding issues
    - Often not possible to use the same tools for processing corpora of different formats
    - Not always a well-documented format

      ↳ *Time-consuming and hard to extract adequate information from the corpora!*
  - Unclear copyright and terms of use

# Aims and Motivation

- Corpora and tools for historical text not easy to find
- Existing historical corpora:
  - Different formats (no well-established format)
    - Text encoding issues
    - Often not possible to use the same tools for processing corpora of different formats
    - Not always a well-documented format

↳ *Time-consuming and hard to extract adequate information from the corpora!*

  - Unclear copyright and terms of use

**Our solution: to gather corpora and tools for historical text in one place, and in a uniform, standardised format, with clearly stated terms of use**

# 1. HISTCORP CORPORA

# CORPORA

- Monitoring corpus
  Intended to grow over time

- Currently 14 (European) languages

  CZECH
  DUTCH
  ENGLISH
  FRENCH
  GERMAN
  GREEK
  HUNGARIAN
  ICELANDIC
  ITALIAN
  LATIN
  PORTUGUESE
  SLOVENE
  SPANISH
  SWEDISH

- Approximately 181,000,000 words (tokens)

- Freely available for download

# CURRENT CORPORA

**Icelandic**
*Icelandic Parsed Historical Corpus*
1150-2008

**Spanish**
*IMPACT-es Diachronic Corpus*
1481-1962

**Slovene**
*Scherrer & Erjavec Historical Dictionaries*
1750-1900

**Portuguese**
*Tycho Brahe Parsed Corpus of Historical Portuguese*
1380-1881

**Greek**
*Ancient Greek Dependency Treebank*

**German**

| | |
|---|---|
| *Deutsches Textarchiv* | 1600-1899 |
| *GerManC* | 1650-1800 |
| *Reference Corpus MHG* | 1050-1350 |

**Dutch**

| | |
|---|---|
| *Compilation* | 1250-2000 |
| *Gutenberg* | 1400-1875 |

**English**
*Lampeter Corpus of Early Modern English Tracts*
1600-1800

**Czech**

| | |
|---|---|
| *Diakorp* | 1350-1939 |
| *Gutenberg* | 1890-1897 |

**French**
*Paris Speech in the Past*
1296-1790

**Latin**
*Ancient Latin Dependency Treebank*

**Italian**
*Gutenberg* 1300-1897

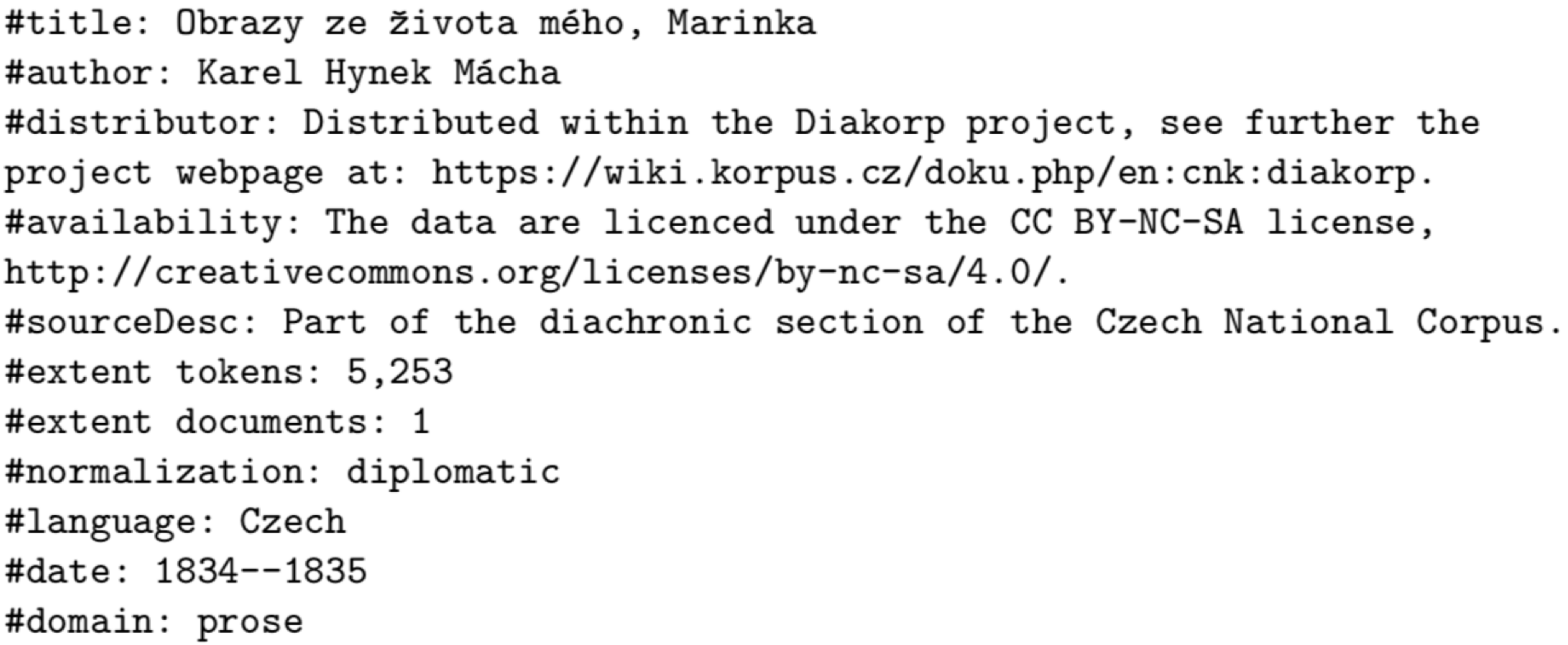
**Hungarian**
*Hungarian Generative Diachronic Syntax*
1440-1539

**Swedish**

| | |
|---|---|
| *Fornsvenska Textbanken* | 1350-1758 |
| *Gender and Work* | 1527-1812 |
| *Gutenberg* | 1789-1902 |
| *Academic Protocols* | 1624-1699 |

# Corpus format

- Each corpus is well-documented, and available in a uniform, standardised format
  - UTF-8
  - Metadata in TEI-compatible format



```
#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
```

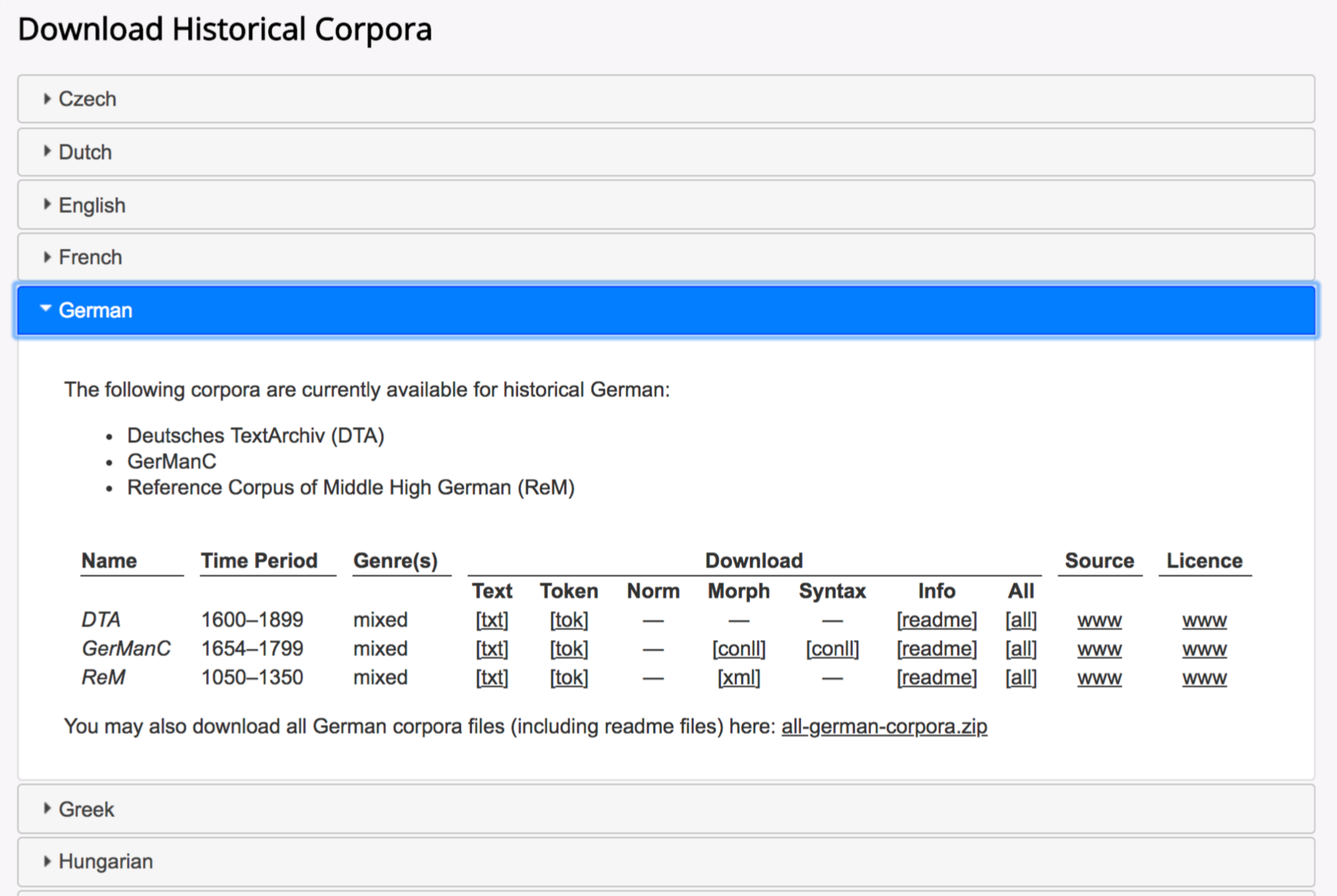
# Corpus format (cont.)

- Links to:
  - Corpus source
  - Licence information

- Downloadable formats for all corpora:
  - Plain text format
  - Tokenised format (one word on each line) – *UDPipe*
  - Readme file with information about the corpus

- Additional formats for applicable corpora
  - Normalisation (linking historical spelling to modern spelling)
  - Morphologically annotated (part of speech, inflection etc)
  - Syntactically annotated

# Corpus download interface



## Download Historical Corpora

- ▸ Czech
- ▸ Dutch
- ▸ English
- ▸ French
- ▾ German

The following corpora are currently available for historical German:

- Deutsches TextArchiv (DTA)
- GerManC
- Reference Corpus of Middle High German (ReM)

| Name | Time Period | Genre(s) | Download | | | | | | | Source | Licence |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Text | Token | Norm | Morph | Syntax | Info | All | | |
| *DTA* | 1600–1899 | mixed | [txt] | [tok] | — | — | — | [readme] | [all] | www | www |
| *GerManC* | 1654–1799 | mixed | [txt] | [tok] | — | [conll] | [conll] | [readme] | [all] | www | www |
| *ReM* | 1050–1350 | mixed | [txt] | [tok] | — | [xml] | — | [readme] | [all] | www | www |

You may also download all German corpora files (including readme files) here: all-german-corpora.zip

- ▸ Greek
- ▸ Hungarian

# II. HistCorp Language Models

# Language Models

- Statistics on sequences of words or characters in a language
- Common word sequences in the Lampeter corpus:

*as well as*

*out of the*

*can not be*

*as to the*

*that it is*

*ought to be*

- Common character sequences in the Lampeter corpus:

*the*

*ing*

*tha*

*ion*

*ere*

*thi*

*con*

*wit*

*oth*

- Useful for research on e.g. language change and language identification

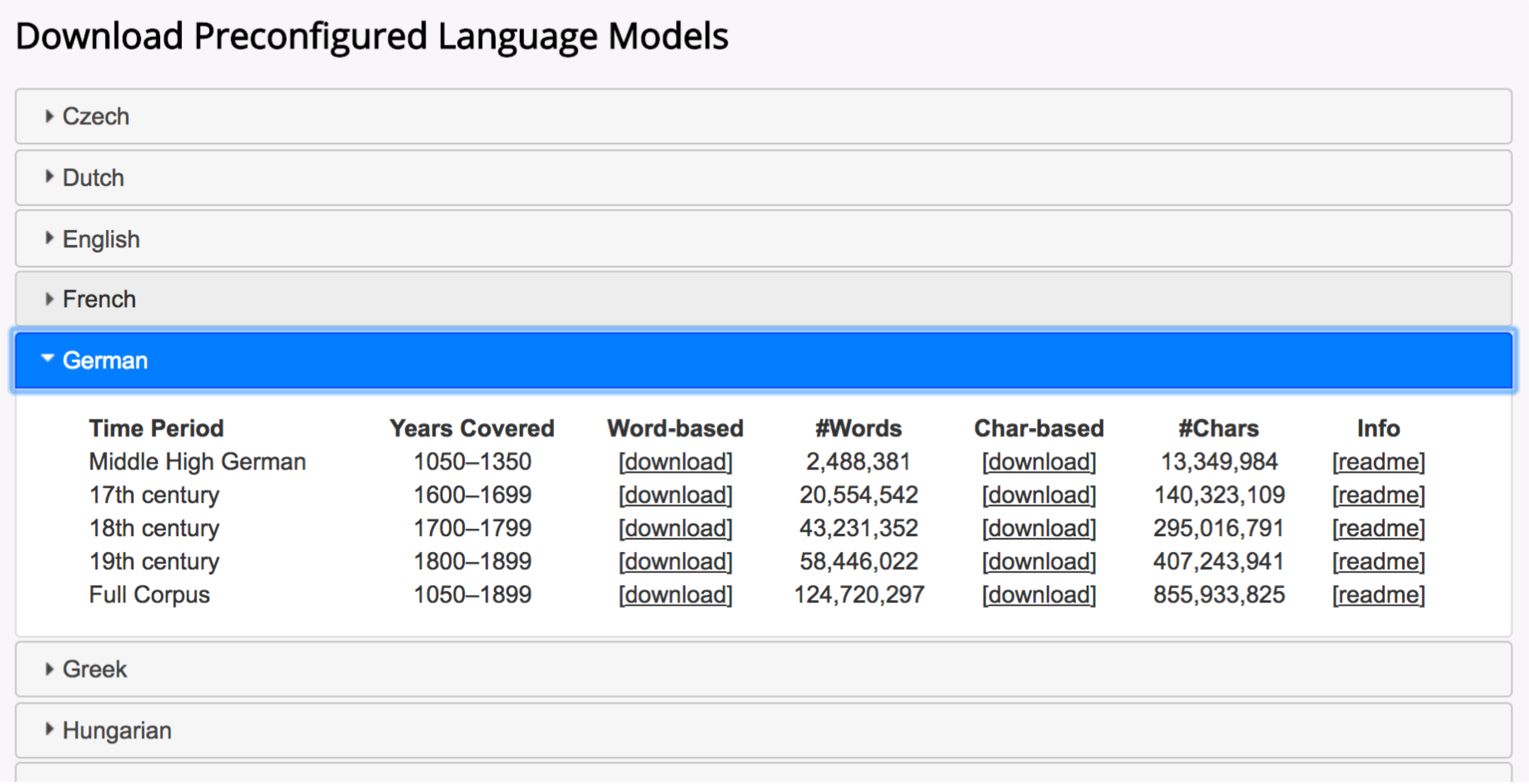
# HISTCORP LANGUAGE MODELS

- On the HISTCORP platform, the user may:
  - download preconfigured language models based on the corpora available on HISTCORP
    or
  - create his/her own language models based by uploading corpora of his/her choice

- The HISTCORP language models are created using the freely available IRSTLM Toolkit*

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

# LM download interface



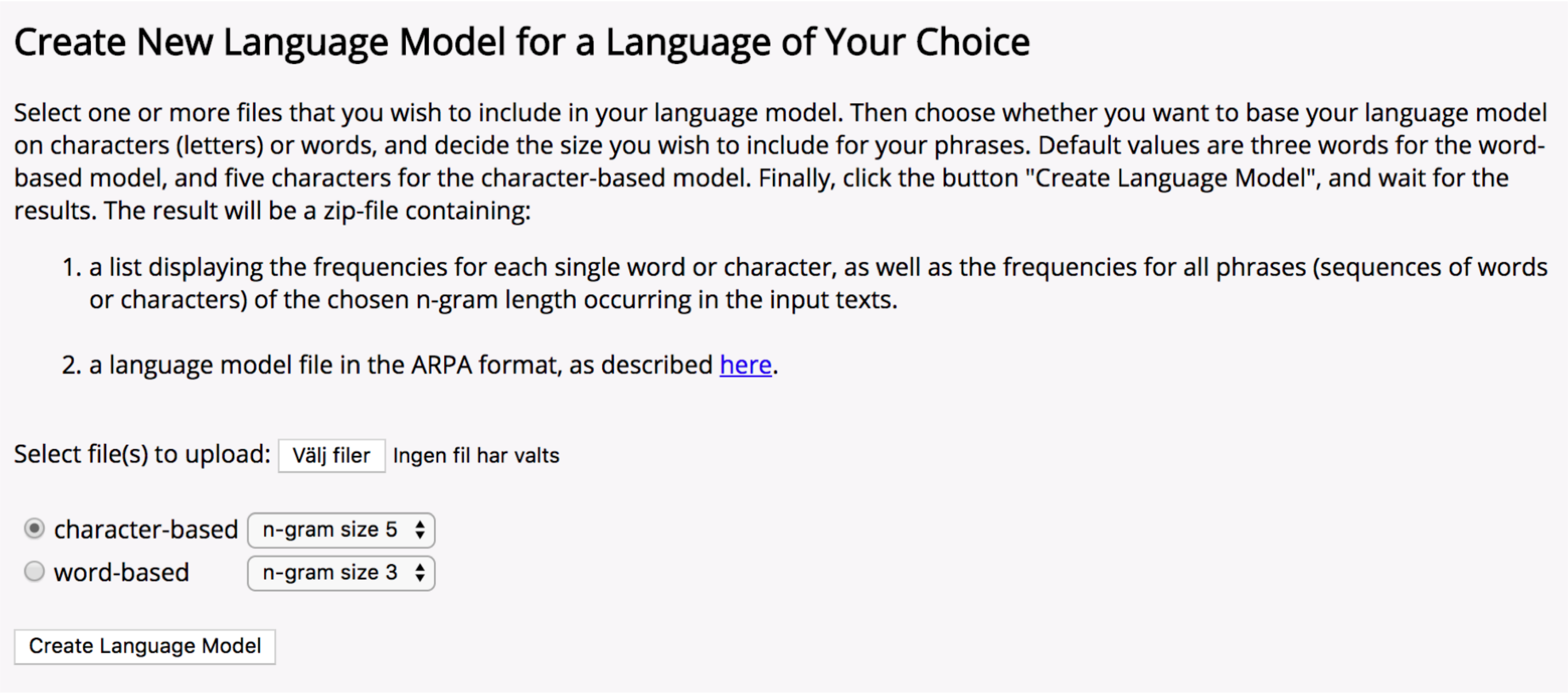
## Download Preconfigured Language Models

- Czech
- Dutch
- English
- French
- German

| Time Period | Years Covered | Word-based | #Words | Char-based | #Chars | Info |
|---|---|---|---|---|---|---|
| Middle High German | 1050–1350 | [download] | 2,488,381 | [download] | 13,349,984 | [readme] |
| 17th century | 1600–1699 | [download] | 20,554,542 | [download] | 140,323,109 | [readme] |
| 18th century | 1700–1799 | [download] | 43,231,352 | [download] | 295,016,791 | [readme] |
| 19th century | 1800–1899 | [download] | 58,446,022 | [download] | 407,243,941 | [readme] |
| Full Corpus | 1050–1899 | [download] | 124,720,297 | [download] | 855,933,825 | [readme] |

- Greek
- Hungarian

# LM DIY INTERFACE



## Create New Language Model for a Language of Your Choice

Select one or more files that you wish to include in your language model. Then choose whether you want to base your language model on characters (letters) or words, and decide the size you wish to include for your phrases. Default values are three words for the word-based model, and five characters for the character-based model. Finally, click the button "Create Language Model", and wait for the results. The result will be a zip-file containing:

1. a list displaying the frequencies for each single word or character, as well as the frequencies for all phrases (sequences of words or characters) of the chosen n-gram length occurring in the input texts.

2. a language model file in the ARPA format, as described here.

Select file(s) to upload: Välj filer Ingen fil har valts

● character-based n-gram size 5
○ word-based n-gram size 3

Create Language Model

# III. HISTCORP TOOLS

# TOOLS

- Tools for automatic processing of historical text

- Currently contains tools for spelling normalisation
  - Translating the historical spelling to a modern spelling
  - Useful for e.g. searching historical text or as a preprocessing step for further analysis of the text

*To the moost noble & Worthiest Lordes moost ryghtful & wysest conseille to owre lige Lorde the Kyng compleynen if it lyke to yow the folk of the Mercerye of London as a membre of the same citee*

*To the most noble and worthiest Lords most rightful and wisest council to our liege Lord the King complain if it likes to you the folk of the Mercery of London as a member of the same city*

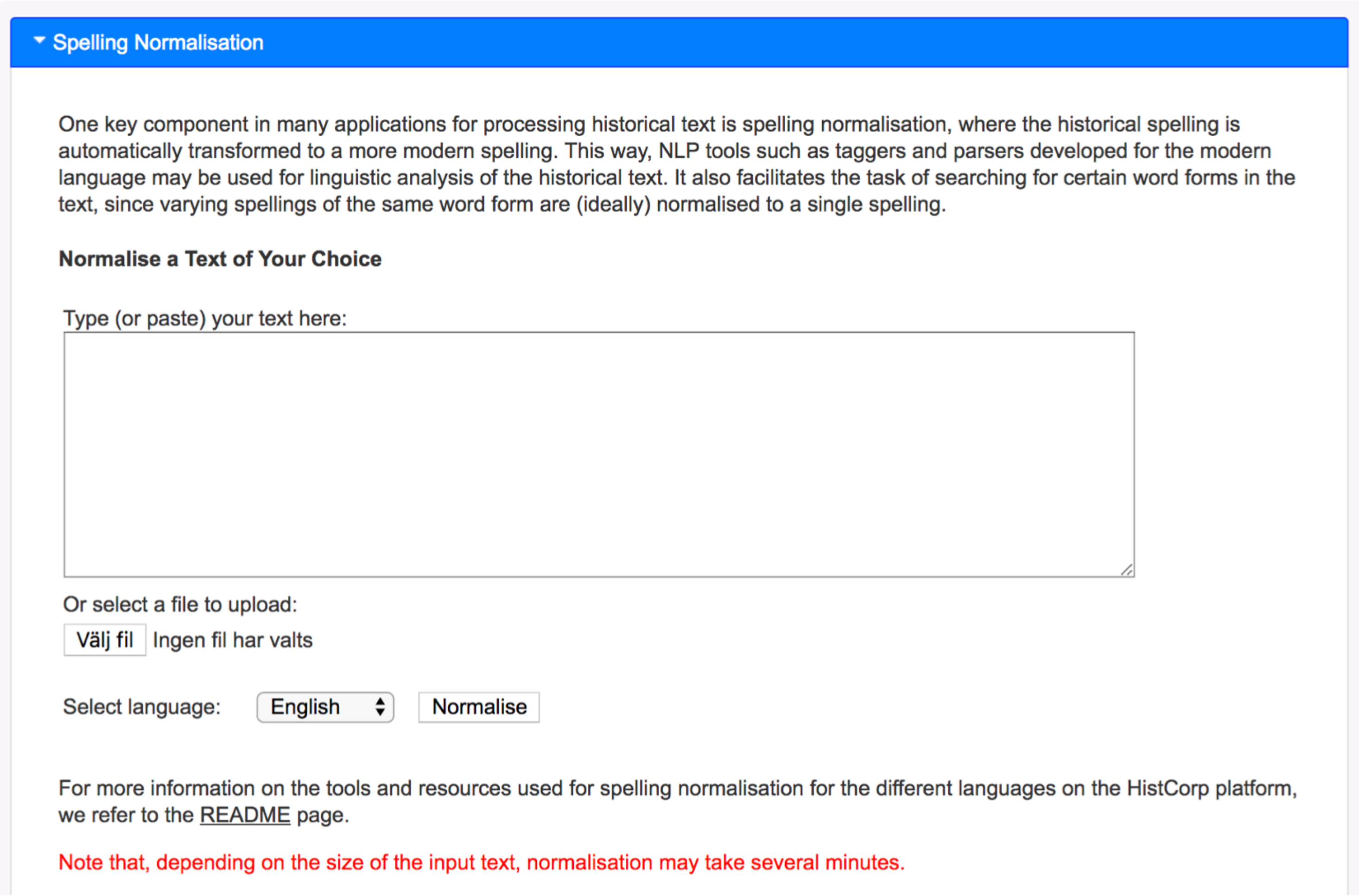
# Spelling Normalisation

- On HISTCORP, spelling normalisation is provided in 2 ways:

1. Type a text or upload a file with historical spelling. The output will be a file with the text normalised to a more modern spelling

2. Download the HISTNORM package, to install the normalisation program on your own computer

# Normalisation interface



Spelling Normalisation

One key component in many applications for processing historical text is spelling normalisation, where the historical spelling is automatically transformed to a more modern spelling. This way, NLP tools such as taggers and parsers developed for the modern language may be used for linguistic analysis of the historical text. It also facilitates the task of searching for certain word forms in the text, since varying spellings of the same word form are (ideally) normalised to a single spelling.

**Normalise a Text of Your Choice**

Type (or paste) your text here:

Or select a file to upload:

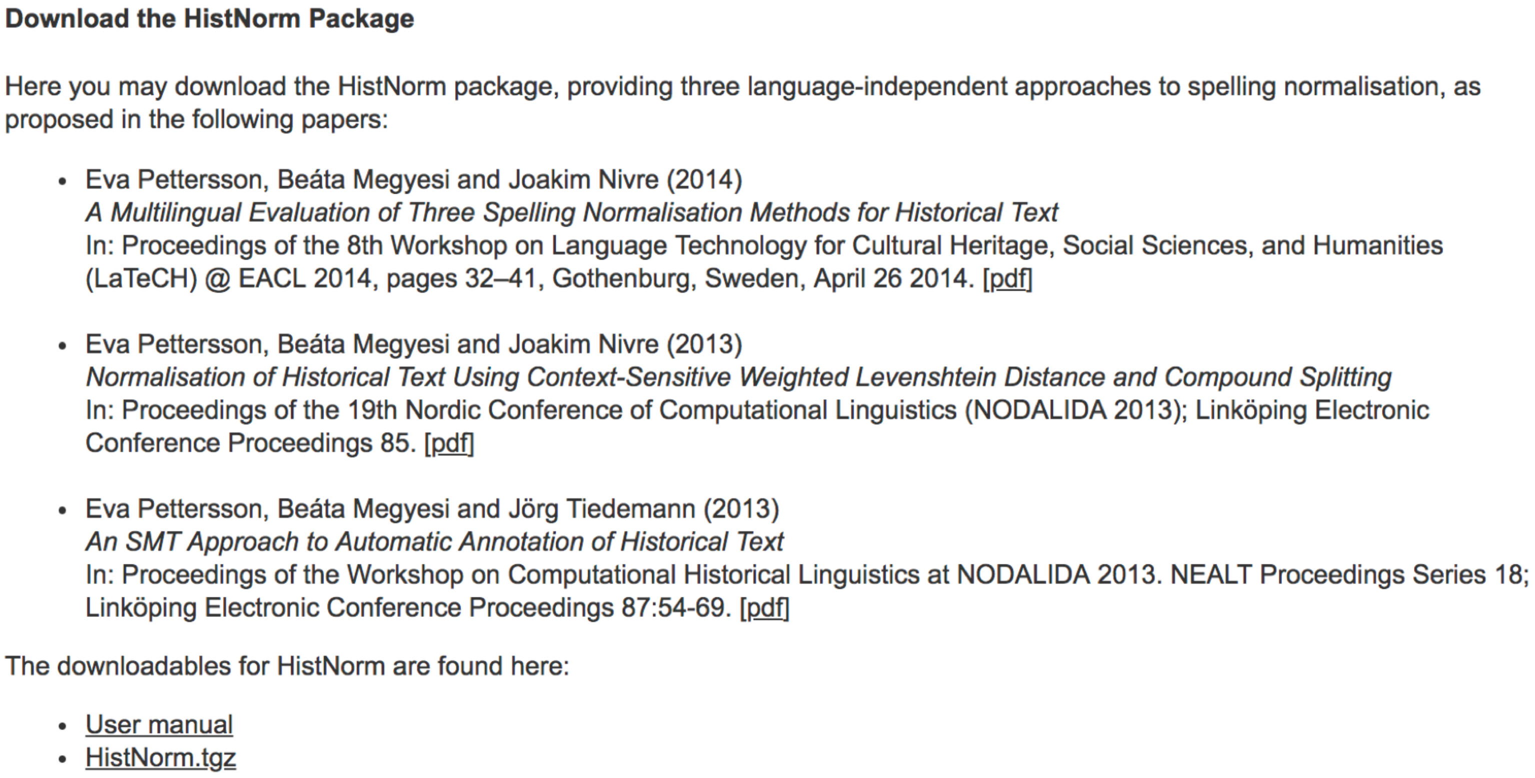
Ingen fil har valts

Select language: English

For more information on the tools and resources used for spelling normalisation for the different languages on the HistCorp platform, we refer to the README page.

Note that, depending on the size of the input text, normalisation may take several minutes.

# Download Normalisation Package



**Download the HistNorm Package**

Here you may download the HistNorm package, providing three language-independent approaches to spelling normalisation, as proposed in the following papers:

- Eva Pettersson, Beáta Megyesi and Joakim Nivre (2014)
  *A Multilingual Evaluation of Three Spelling Normalisation Methods for Historical Text*
  In: Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH) @ EACL 2014, pages 32–41, Gothenburg, Sweden, April 26 2014. [pdf]

- Eva Pettersson, Beáta Megyesi and Joakim Nivre (2013)
  *Normalisation of Historical Text Using Context-Sensitive Weighted Levenshtein Distance and Compound Splitting*
  In: Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013); Linköping Electronic Conference Proceedings 85. [pdf]

- Eva Pettersson, Beáta Megyesi and Jörg Tiedemann (2013)
  *An SMT Approach to Automatic Annotation of Historical Text*
  In: Proceedings of the Workshop on Computational Historical Linguistics at NODALIDA 2013. NEALT Proceedings Series 18; Linköping Electronic Conference Proceedings 87:54-69. [pdf]

The downloadables for HistNorm are found here:

- User manual
- HistNorm.tgz

# Conclusion and Future Work

- HistCorp – A freely available open-access platform containing historical corpora and resources for (currently) 14 languages:
  - Corpora
    - Standardised, uniform and well-documented format
    - Easy to find metadata and licence information
  - Language models
    - Download predefined language models or
    - Create your own language model
  - Tools
    - Run or download spelling normalisation tools
- We plan to continuously add more corpora and tools