UPPSALA UNIVERSITET

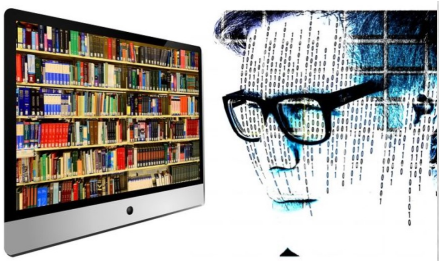
DIGGING THE PAST

Digital Philology and the Analysis of Historical Sources

Beáta Megyesi
Department of Linguistics and Philology, Uppsala University
beata.megyesi@lingfil.uu.se

Language Technology: Research and Development
August 30, 2022

UPPSALA UNIVERSITET



Digital Philology

- Digital Philology is a subfield of digital humanities.

  ”An academic field concerned with the application of computational tools and methods to traditional humanities disciplines such as literature, history, and philosophy.” (Oxford English Dictionary)

- Rather new research area, constantly changing.

- DP is a combination of subjects in:
  - classical philology: linguistics and literary studies
  - computer science: computational methods and tools.

- It involves (fairly) large amounts of written text data: historical or modern digitized text or images.

UPPSALA UNIVERSITET

Topics

- Contributions might cover a wide range of topics about the **accessibility** and **analysis** of various languages or sources from Ancient to modern languages.

- The studies can involve:
  - (semi-)automatic transcription,
  - annotation on various linguistic levels, and/or
  - document analysis and text mining;

- The methods used can include old-school rule-based formalisms, traditional corpus linguistics, classical machine learning methods, or deep learning.

UPPSALA UNIVERSITET

Aims and Motivation

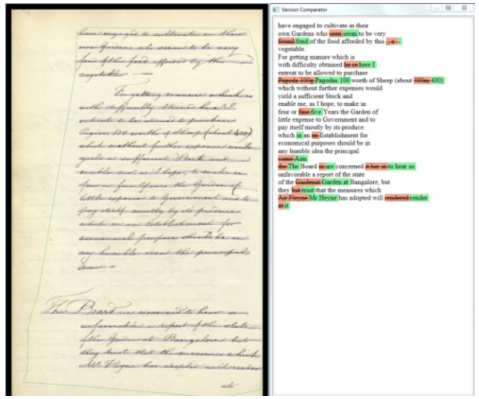
Discover the past to understand the present.

- Historical text constitutes a rich source of information
- Not easily accessed
- Many texts are not digitized
- Lack of language technology tools to handle even digitized historical text
- Leads to time-consuming manual work for historians, philologists and other researchers in humanities

UPPSALA UNIVERSITET

(Some) Challenges with Historical Text

- Often not digitized
  - hand-written or printed
  - transcription is needed
  - ➢ digitization: sources, formats, metadata
  - ➢ hand-written text recognition (HTR) or
  - ➢ optical character recognition (OCR)



- Different and inconsistent spelling:
  both diachronic and synchronic spelling variance
  due to lack of spelling conventions,
  Write as you speak – various dialects

mig
migh
mik
mic
mich
mech

pronoun mig (‘me/myself’) in the Swedish book of prayers *Svenska tideboken* (1525)

UPPSALA UNIVERSITET

Spelling Variation Extreme

- The word **tiuvel** (Teufel) ‘devil’ occurs 733 times in *Reference Corpus of Middle High German* with 90 different spellings:

dievel diuel diufal diuual diu=uil diuvil divel divuel divuil divvel dufel duoifel duovel duuel duuil duvel duvil dvofel dvuil dwowel lieuel loufel teufel tevfel thufel thuuil tiefal tiefel tiefil tieuel tie=uel tieuil tieuuel tieuuil tievel ti=evel tie=vel tievil tifel tiofel tiuel tiufal tiufel tiufil tiufle tiuil tiuofel tiuuel tiuuil tiuval tiuvel tiuvil tivel tivfel tivil tivuel tivuil tivvel tivvil tivwel tiwel tubel tubil tueuel tufel tufil tuifel tuofel tuouil tuovel tuovil tuuel tuuil tuujl tuvel tuvil tvfel tvivel tvivil tvouel tvouil tvovel tvuel tvuil tvvel tvvil tyefel tyeuel tyevel tyfel

UPPSALA UNIVERSITET

Vocabulary

- Different vocabulary (often with Latin influences)
  - New words enter the language (e.g., tech development)
  - Old words become less frequent or eventually disappear
  - Examples:

Early New High German Words (1350–1650) *:

Old Form	Modern	Gloss
liberei/librari	Bibliothek	'library'
triangel	Dreieck	'triangle'
akkord	Vertrag	'treaty'

* Salmons (2012): *A History of German – What the past reveals about today's language*

UPPSALA UNIVERSITET

Morphology and Syntax

- Different (and inconsistent) morphology
  - Shift in inflection from strong to weak paradigm

Historical English	Modern English*
old – elder – eldest	old – older – oldest

- Long(er) sentences
- Inconsistent use of or lack of punctuation
- Different syntax and inconsistent word order
- Code-switching

Substantial differences between texts from different time periods, genres, and authors

UPPSALA UNIVERSITET

Annotating Historical Texts

Two main approaches:

1. Train a tagger and/or a parser on historical data
   - Straightforward
   - Data sparseness issues

2. Spelling Normalisation
   - Automatically translate the original spelling to a modern spelling, before performing tagging and parsing
   - Enables the use of NLP tools available for the modern language
   - Does not take syntactic differences and changes in vocabulary into account

UPPSALA UNIVERSITET

Non-Standard Language Data

- The same methods that are used for NLP for historical text have also been used for modern text, such as Twitter data
- Spelling normalisation useful before tagging/parsing

seein that ad makes me wanna listen to dat song rite now

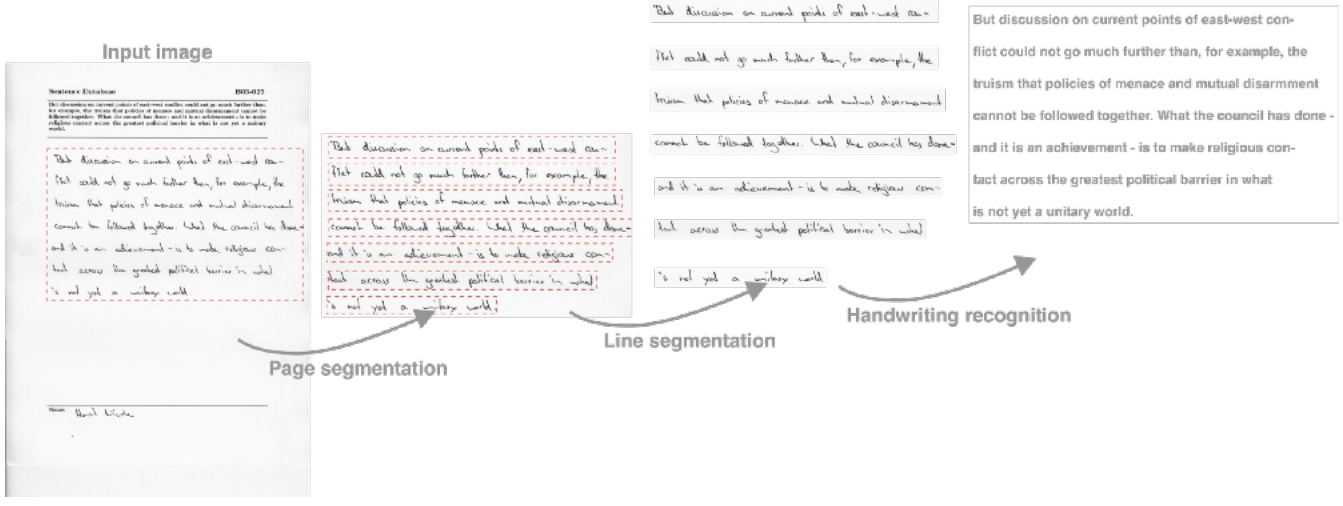
Example from Clark & Araki (2011)

UPPSALA UNIVERSITET

Possible topics (not limited to!)

Digital processing:

- (semi-)automatic transcription of historical sources to produce a digitized text from images;
- spelling normalization to automatically transform the original spelling to a modern standard spelling in order to be able to apply tools developed for modern language;
- detection of cleartext in enciphered sequences in historical ciphers

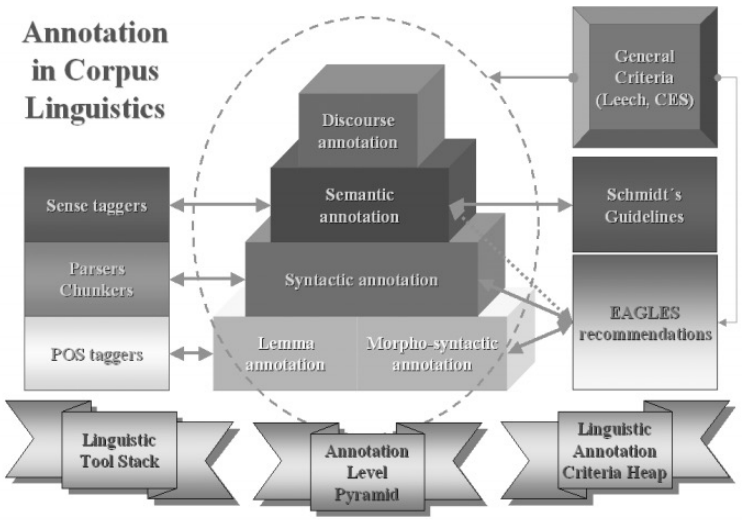


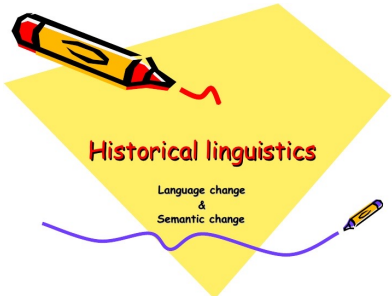
UPPSALA UNIVERSITET

Possible topics (not limited to!)

Annotation and analysis of sources by computational approaches:

- linguistic annotation: PoS tagging and syntactic parsing;
- language identification in historical sources incl. codeswitching
- linguistic analysis incl. comparative studies across genres, time periods
- detecting historical language change: morphologic, syntactic, or semantic



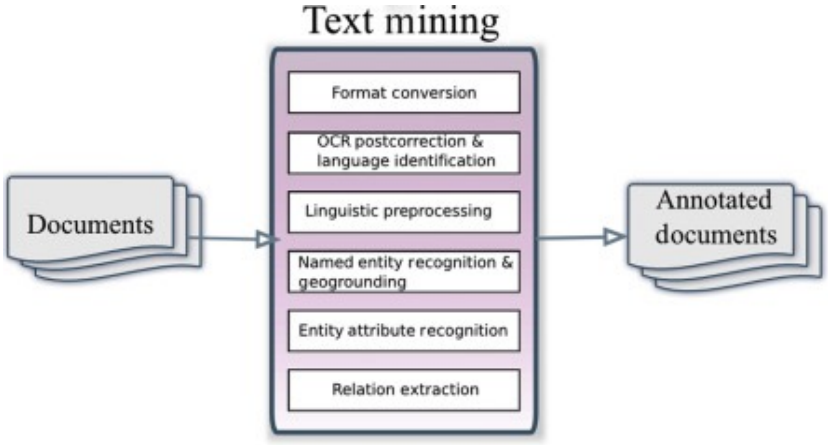


UPPSALA UNIVERSITET

Possible topics (not limited to!)

Applications:

- topic modeling/genre/register detection in historical and/or modern texts;
- comparative studies of some aspects in literature;
- summarization of historical sources;
- decryption of historical sources;



UPPSALA UNIVERSITET

Previous Year's Projects: some examples

Normalisation

- *An Evaluation of Different Approaches to Spelling Normalization of English Historical Text*
- *Character-based SMT and NMT for Historical Text Normalization*
- *An Evaluation of NMT Models on Historical Spelling Normalization* (COLING paper)

Tagging Historical Text

- *Evaluating Two Modern POS Taggers on Historical Novels by Daniel Defoe*

NLP for Twitter

- *A Pipeline for Twitter Lexical Normalization*

Machine Translation

- *Translating Middle Egyptian to Modern English with NMT*

[supervised by Eva Pettersson]

UPPSALA UNIVERSITET

Papers

1. Data

Michael Piotrowski (2012) Natural Language Processing for Historical Texts, chapter 3: *Spelling in Historical Texts*

Michael Piotrowski (2012) Natural Language Processing for Historical Texts, chapter 6: *Handling Spelling Variation*

Daniel van Strien, Kaspar Beelen, Mariona Coll Ardanuy, Kasra Hosseini, Barbara McGillivray, and Giovanni Colavizza. (2020) Assessing the impact of OCR quality on downstream NLP tasks. In ICAART (1), pages 484–496.

---

2. Tools

Michael Piotrowski (2012) Natural Language Processing for Historical Texts, chapter 7: *NLP Tools for Historical Languages*

Marcel Bollmann, *A Large-Scale Comparison of Historical Text Normalization Systems. In Proceedings of NAACL-HLT 2019, pages 3885–3898 Minneapolis, Minnesota, June 2 - June 7, 2019*

Michael A. Hedderich, Lukas Lange Heike Adel, Jannik Strötgen & Dietrich Klakow A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568 June 6–11, 2021.

---

3. Applications

Eduard Hovy and Julia Lavid (2010). Towards a ‘Science’ of Corpus Annotation: A New Methodological Challenge for Corpus Linguistics. International Journal of Translation Vol. 22, No. 1, Jan-Jun 2010

Xutan Peng, Yi Zheng, Chenghua Lin and Advaith Siddharthan. (2021) Summarising Historical Text in Modern Languages**.** In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics.

Adam Hammond, Julian Brooke, and Graeme Hirst (2013). A Tale of Two Cultures: Bringing Literary Analysis and Computational Linguistics Together. Proceedings of the Second Workshop on Computational Linguistics for Literature, pages 1–8.

UPPSALA
UNIVERSITET

Questions?