Digital Philology

Mats Dahllöf, Beáta Megyesi, Joakim Nivre, Eva Pettersson

UPPSALA UNIVERSITET

SWE-CLARIN

Overview

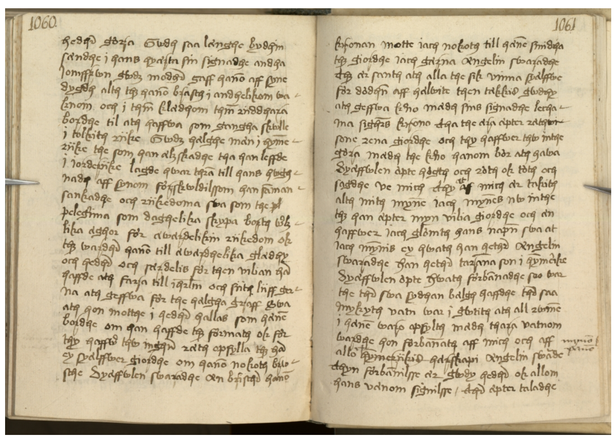
- How can computational linguistics support text-based research in the humanities and social sciences?
- Digital Philology is a subfield within Digital Humanitites dealing with the digitization and automatic processing of text-based sources thereby allowing scholars in the humanities and social sciences to study large amounts of data more systematically.

SWE-CLARIN

- SWE-CLARIN is the Swedish branch of the "Common Language Resources and Technology Infrastructure" ERIC and focuses on the development of resources and tools that can be used for research in digital humanities and social sciences.
- Participants: The Swedish Language Bank,The Language Council of Sweden, Swedish National Data Service, Digisam, KTH, Linköping, Lund, Stockholm and Uppsala University.

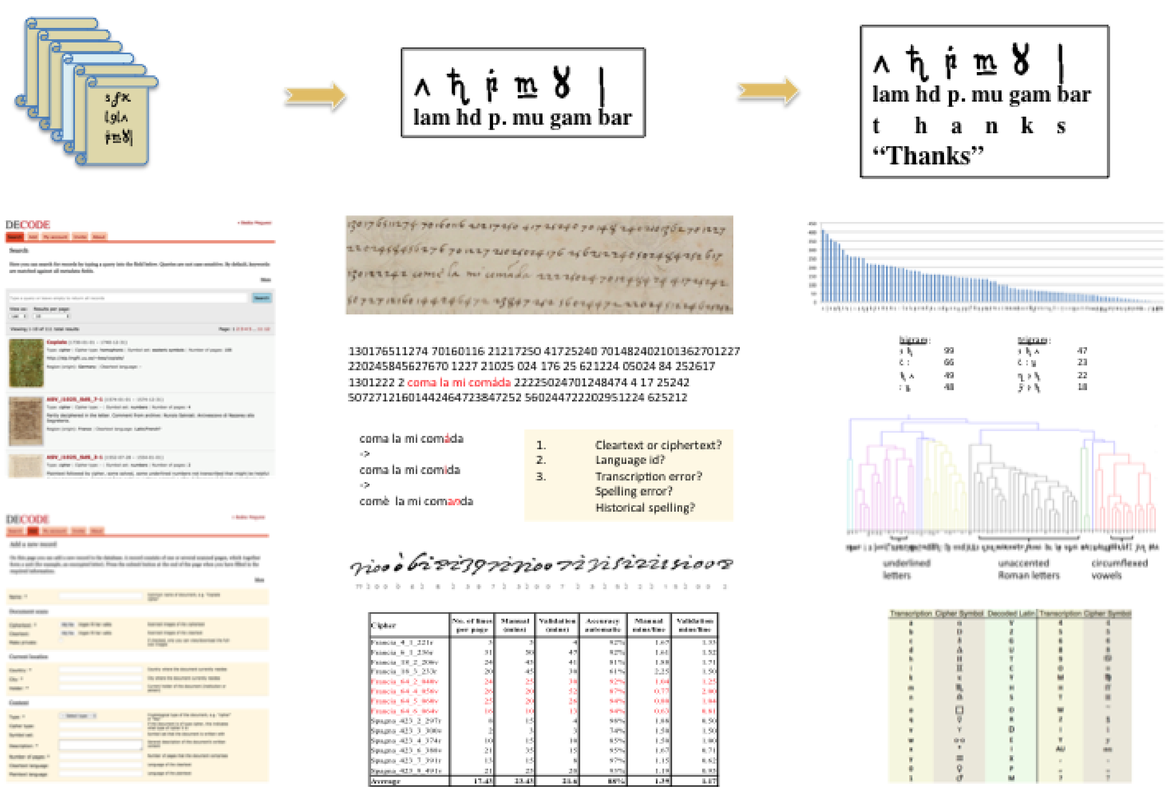
From Quill To Bytes

- Computational analysis of historical handwritten material, mainly medieval charters and manuscripts.
- Cooperation involving several subjects: image analysis, philology, computational linguistics, and history, in a series of projects.
- Word spotting and transcription.
- Scribe attribution and dating.



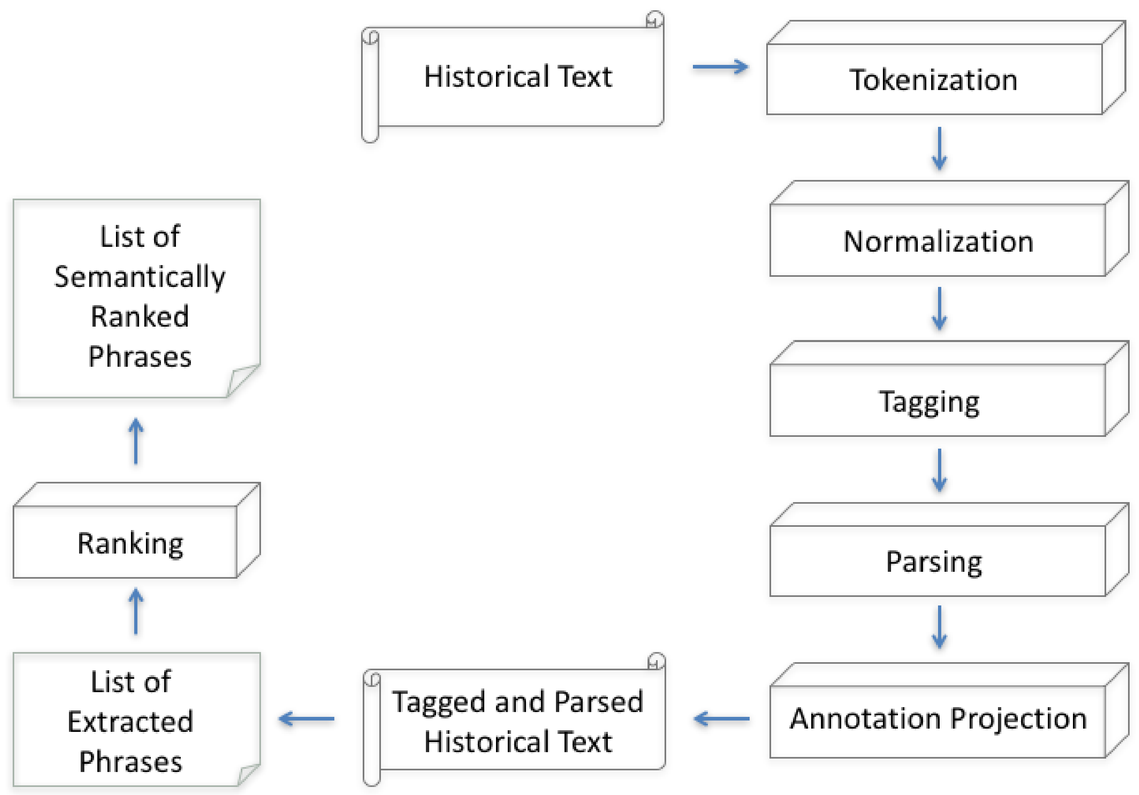
DECODE

- Automatic decoding of historical secret writings
  - Create a database of commonly occurring cipher types, codes and keys from Early modern times (1400-1850)
  - Automatically categorize different types of ciphers
  - Develop algorithms/tools to decrypt various types



HistSearch

- A tool for automatic information extraction from historical text
- Applied to the Gender and Work project:
  - Historians are interested in what men and women did for a living in the Early Modern Swedish society (appr. 1550–1800)
  - HistSearch automatically searches text for phrases describing working activities
  - Spelling normalization is a crucial step, translating the original spelling to a standardized spelling, prior to tagging and parsing



SWEGRAM

- SWEGRAM is a web-based tool for automatic annotation and quantitative linguistic analysis of Swedish texts.
- SWEGRAM allows users to create their own corpus or compare texts on various linguistic levels. We used SWEGRAM for the creation of the Uppsala Corpus of Student Writings consisting of 1.5 million tokens.