Seminar 06491: Digital Historical Corpora

Lou Burnard
Milena Dobreva
Norbert Fuhr
Anke Lüdeling

The seminar brought together scholars from (historical) linguistics, (historical) philology, computational linguistics and computer science who work with collections of historical texts. These texts or digital libraries or corpora[1] are collected for a number of different purposes such as lexicography, history, linguistics, philology etc. This, naturally, leads to different decisions in their design and architecture.

However, there are many issues that are common to many projects working with historical texts. These include:

Standards and methods of digitization: historical texts have to be digitized from different sources. Sometimes it is necessary to digitize directly from a manuscript or early print. In these cases it is not possible to use current OCR technology, and the texts have to be double keyed (for example according to the standards developed in the Kompetenzzentrum Retrodigitalisierung in Trier). Newer texts can sometimes be scanned and OCRed, although even the relatively 'clean' 19th century newspaper texts are often problematic. Fraktur and some other scripts (e.g. old Cyrillic scripts) also pose problems for OCR.

For some research questions it is possible to work with editions. In these cases the digitization itself is not an issue (if the editions are new). It has to be decided, however, how to deal with a critical apparatus.

Design (composition) of corpora: While literary scholars often work on one text (or a small number of related texts), many research questions in linguistics and lexicography require a collection of several texts. Corpus design is, of course, always an issue in corpus construction. Ideally a matrix of the necessary parameters (text type, author, time etc.) is constructed and all 'cells' are filled with the appropriate texts. For older time periods this is often not possible since the texts might not have survived. A 'skewed' corpus, of course, only permits certain research questions.

Standards and methods of annotation: For many research questions it is not sufficient to have the 'naked' text. The texts need to be annotated with further information. The texts need (a) header annotation (information about the whole text), (b) positional annotation (annotation for each token), and (c) structural annotation. The Text Encoding Initiative and other groups have developed suggestions for historical texts (the most detailed suggestions pertain to the header annotation). Annotation often cannot be done automatically since older texts are less standardized than newer texts – it is difficult to develop statistical or rule-based methods. It is necessary to discuss possible automation. It is also necessary to develop good annotation tools for manual or semi-automatic annotation.

Corpus architecture: Most large modern corpora are stored in some table or tree format. Such architectures might not be the best option for historical corpora since they cannot accommodate conflicting annotation. Therefore one has to think about alternatives like multi-layer models or database models.

[1] Henceforth we will speak of corpora even though some of the text collections would not be considered corpora by some scholars.

Search and exploitation of corpora: It is necessary to develop search machines that can deal with non-standard data, either by using some kind of fuzzy mechanism or by doing some internal normalization. It is also necessary to think about quantitative methods for relatively small datasets.

We will briefly summarize the main points of each contribution before giving an overall summary.

Construction and Design of Historical Corpora

Kurt Gärtner (Trier): *The Middle High German Text Archive*. Kurt Gärtner presented a number of interconnected lexicon projects for Middle High German. The lexicons and many of the source texts for the lexical entries have been digitized and are available online at http://germazope.uni-trier.de/Projects/WBB/woerterbuecher/.

Eva Dyllong (Duisburg-Essen): *A Multifunctional Historical Document Research System.* Eva Dyllong presented a system for the collection, digitization, storage, fuzzy retrieval, communication, annotation and visualisation of historic documents.

Manfred Markus (Innsbruck): *Wright's English Dialect Dictionary – architecture and retrieval.* Wright's English Dialect Dictionary (published around 1900) is the largest historical dialect dictionary for English. Manfred Markus described the dictionary, its composition, entry structure and collection methods. He then explained how it was digitized and how it can be researched online.

Tomaz Erjavec (Ljubljana): *Capture of historical corpora.* Tomaz Erjavec spoke about corpus storage and annotation methods. Since many historical linguists are not at the same time computational linguists and might therefore not be able (or willing) to use very specific tools he suggested using regular Windows-based tools for annotation. Erjavec then presented a Slovenian case study.

Roland Meyer (Regensburg): *Design issues in a corpus for historic linguistics.* Roland Meyer presented the architecture of the Regensburg Diachronic Corpus of Russian. His focus was on the annotation methods and tools (especially the annotation tool ACT) and search procedures.

Amir Zeldes (Berlin): *A parallel corpus of historical Polish.* Amir Zeldes described the construction and architecture of a parallel diachronic Bible corpus of Polish.

Nikola Ikonomov (Sofia): *A dialect corpus*. Nikola Ikonomov presented a project for the preservation and digital conversion of historical Bulgarian dialect recordings.

Milena Dobreva (Sofia): *Issues*. Milena Dobreva summarized the presentations about the design and construction of historical corpora. She also informed us about the specific situation in Bulgaria.

Corpus Storage and Encoding

Viktor Baranov (Izhevsk): *Storage and encoding in the Manuscript system and its application to research.* The Manuscript system offers a multi-layer architecture for historical Russian corpora and annotation and retrieval facilities.

James Cummings (Oxford): *Migration of Legacy Texts.* The Oxford Text Archive stores all textual resources that are produced in publicly funded projects in the UK. Older texts have to be converted to a sustainable format. James Cummings presented the OTA solutions.

Astrid Ensslin (Manchester): *GerManC.* The GerManC corpus is a corpus of Early New High

German newspaper texts. Astrid Ensslin presented the corpus design and discussed first results of studies on standardization in ENHG.

Anke Lüdeling (Berlin): *DeutschDiachronDigital.* Anke Lüdeling presented a Germany-wide initiative to build a diachronic corpus for German (DeutschDiachronDigital). The proposed corpus architecture should be flexible enough to encode conflicting annotation layers. Each annotation layer should be standardized in order to enable a uniform search.

Lou Burnard (Oxford): *The Text Encoding Initiative.* Lou Burnard described the aims and structure of the Text Encoding Initiative and presented the P5 recommendations for manuscript encoding.

Andreas Witt (Tübingen): *TEI P5 and Multiple Hierarchies.* Andreas Witt then presented several TEI recommendations for encoding multiple hierarchies and discussed their strengths and weaknesses.

Corpus Annotation

Andreas Witt (Tübingen): *Corpus annotation and sustainability, interoperability, long-term storage.* In a second talk Andreas Witt provided strategies for long-term storage of digital data. He is involved in a sustainability project (funded by the German Science Foundation) that works on recommendations for encoding standards and ontologies for annotation tagsets.

Mila Vulchanova (Trondheim): *Part-of-speech tagging of participles in Old Bulgarian.* Participles in Old Bulgarian are difficult to categorize because they have verbal properties as well as nominal properties. The talk discussed the different functions of participles and the consequences for part-of-speech tagging of Old Bulgarian.

Paul Rayson (Lancaster): *Tagging Historical Corpora.* Most automatic taggers are trained on modern language data. Paul Rayson showed which tagging problems could occur in historical corpora where spelling and word order are less standardized. He also presented algorithms to normalize spelling.

Meike Klettke (Rostock): *The cadastral register of the town Wismar. Structure Mining in historical documents.* Meike Klettke described information extraction techniques for converting historical registers into a formatted database.

Greg Crane (Boston): *Automatic annotation, named entity identification.* Named entity recognition is especially important for historical texts because many places changed their names over time and person names are often spelled in several different ways. Greg Crane presented methods to uniquely identify place names using GPS coordinates.

Stefan Evert (Osnabrück): *The Web as corpus.* Working with historical corpora is difficult because the texts are less standardized than modern (newspaper) texts. Similar problems occur in other less standardized text types. Stefan Evert showed some parallels between historical texts and computer-based communication.

Corpus Retrieval and Indexing

Mark Davies (Provo): *Competing demands of size, speed, and annotation with historical corpora.* Mark Davies showed how a relational database management system can be used for various kinds of analyses of annotated historical corpora.

Andrea Ernst-Gerlach (Duisburg-Essen): *Retrieval in text collections with historic spelling* Andrea Ernst Gerlach presented a machine learning approach for generating historic spelling variants via transformation rules.

Thomas Pilz (Duisburg-Essen): *Search in non-standard databases.* Thomas Pilz discussed specific metrics for dealing with problems in optical character recognition, transcription and historical spelling.

Markus Heller (München): *Technologies for processing historical German texts*. Markus Heller described methods for OCR postcorrection, XML indexing and fast approximative search in a historical corpus.

Jaap Kamps (Amsterdam): *A cross-language approach to historical document retrieval*. Jaap Kamps investigated the application of cross-language information retrieval methods for searching in historical texts.

Alexander Karosseit & Ulf Leser (Berlin): *Large historical text corpora based on database management systems*. Alexander Karosseit talked about the representation of multiple annotations of texts as concurrent XML markup, and discussed the development of an appropriate query language.

Lou Burnard (Oxford): *Xaira*. Lou Burnard presented a system for providing linguistically-motivated search facilities for collections of richly encoded XML documents.

Corpus Exploitation

Jean Daniel Fekete (Paris): *Information Visualization for Corpora*. Jean Daniel Fekete showed how the concept of preattentive perception can be used for designing visualization methods easing understanding and navigation of corpora.

Wolfram Luther (Duisburg-Essen): *Rule-based search in historical text databases - Visualization techniques*. Wolfram Luther described several methods used to visualize rule sets for deriving non-standard spellings.

Mark Davies (Provo): *Semantic and syntactic change*. Mark Davies presented quantitative methods to model language change in large historical databases.

Thorsten Vitt & Fotis Jannidis (Darmstadt): *TextGrid - Grid-based philological infrastructure*. Thorsten Vitt presented an ongoing project for developing an infrastructure enabling the collaborative edition, publication, annotation and analysis of texts.

Amir Zeldes: *Machine translation*. Amir Zeldes showed how a parallel corpus can be exploited for annotation by annotation projection.

Greg Crane (Boston): *E-philology.* Greg Crane discussed how methods and standards in philology might change with the availability of large digital libraries.

Ulf Leser & Anke Lüdeling (Berlin): *Bioinformatics methods to detect language relationships*. Texts can be viewed as strings, similar to strings of proteins on DNA. Ulf Leser showed how bioinformatics methods can be used to calculate (genetic) relationships between languages.

Summary

The purpose of this seminar was twofold: First we wanted to inform each other about the decisions each of us had taken in building a historical corpus and discuss the options. Second, we wanted to build an international network of people working with historical corpora and explore the options for further partnerships or projects. We think that both goals were reached.

The seminar was very interesting and stimulating. In the final discussion of the workshop, a ‘grand picture’ of the research issues in the area of digital historic corpora was developed (see Figure 1). Here the arcs represent enabling/supporting methods. As can be seen from this picture,

the major goal is the research on large historical corpora, which requires work on the areas pointing to it directly or indirectly. A researcher's workbench should support personalization, collaboration as well as problem solving. It must be complemented by tools for the annotation and the analysis of corpora, as well as providing functions for visualization, browsing and retrieval (especially for spelling variants). These methods should first be applied to and tested on small corpora, before they can be used for large corpora. In this context, evaluation also plays a major role. For large corpora (stored in digital libraries), the choice of an appropriate architecture is a crucial issue.
Another issue that was of interest to all participants is quality control and standardization.

Publications:
The participants of the workshop plan a common publication. The first choice for such a publication is *Digital Humanities Quarterly*. The editors have been contacted.

Projects:
The participants of the workshop think about collaborations and projects. Milena Dobreva (Sofia) is in charge of searching for appropriate EU calls. There will also be cooperations ands possible projects between subsets of the partcipants (already planned: Transcoop application between Tufts University and Humboldt University, BMBF applications).

Further Workshops:
(Subsets of) the participants will meet for further workshops to discuss specific issues. Two such workshops have already taken place:
April 2007: Diachrone Corpora, historische Syntax und Texttechnologie, Universität Frankfurt
Mai 2007: The Million Books Workshop, Tufts University Boston