06491 Abstracts Collection
Digital Historical Corpora - Architecture, Annotation, and Retrieval
— Dagstuhl Seminar —

Lou Burnard[1], Milena Dobreva[2], Norbert Fuhr[3] and Anke Lüdeling[4]

[1] Oxford Univ. Computing Services, UK
lou.burnard@oucs.ox.ac.uk
[2] Bulgarian Academy of Sciences, BG
dobreva@math.bas.bg
[3] Univ. Duisburg-Essen, DE
norbert.fuhr@uni-due.de
[4] HU Berlin, DE

Abstract. From 03.12.06 to 08.12.06, the Dagstuhl Seminar 06491 "Digital Historical Corpora - Architecture, Annotation, and Retrieval" was held in the International Conference and Research Center (IBFI), Schloss Dagstuhl. During the seminar, several participants presented their current research, and ongoing work and open problems were discussed. Abstracts of the presentations given during the seminar as well as abstracts of seminar results and ideas are put together in this paper. The first section describes the seminar topics and goals in general. Links to extended abstracts or full papers are provided, if available.

Keywords. Historical corpora, digitization, corpus design, corpus architecture, search, retrieval, standardization

06491 Summary – Digital Historical Corpora

The seminar "Digital Historical Corpora" brought together scholars from (historical) linguistics, (historical) philology, computational linguistics and computer science who work with collections of historical texts. The issues that were discussed include digitization, corpus design, corpus architecture, annotation, search, and retrieval.

Keywords: Historical corpora, digitization, corpus design, corpus architecture, search, retrieval, standardization

Joint work of: Burnard, Lou; Dobreva, Milena; Fuhr, Norbert; Lüdeling, Anke

Extended Abstract: http://drops.dagstuhl.de/opus/volltexte/2007/1039

Dagstuhl Seminar Proceedings 06491
Digital Historical Corpora - Architecture, Annotation, and Retrieval
http://drops.dagstuhl.de/opus/volltexte/2007/1056

Information-Analytical System "Manuscript": technologies and tools of creation of electronic collections of ancient and medieval documents

Victor Baranov (Izhevsk State Technical University, RUS)

The paper is devoted to the possibilities of the Manuscript system (http://manuscripts.ru/) designed for preparation of electronic scientific publications of ancient manuscripts on the Internet.

The primary consideration is given to the specialized modules of the system ensuring 1) input, storage, editing and processing of materials in the database, 2) textologic, linguistic and paleographic analyses of manuscripts/texts and 3) preparation of dummy copies and publication of manuscripts and research apparatus.

All modules interact with a common database allowing processing text/manuscript units organized into hierarchies and nets, their relationships and values that adequately reflect modeled objects and their relationships. The report also shows the possibilities of the system modules for a comprehensive study of texts and their units.

Keywords: Full-text database, electronic editions on the Internet, ancient and medieval manuscripts

Extended Abstract: http://drops.dagstuhl.de/opus/volltexte/2007/1041

TEI P5: state of the art

Lou Burnard (Oxford University Computing Services, GB)

The paper presents an update on the current state of development of the Text Encoding Initiative's Recommendations for the encoding of machine readable text. Since the last major edition in 2000, which saw the conversion of the Guidelines into XML, there has been substantial activity on adding new content in areas of particular interest to historical corpus builders. The TEI has also reinvented itself as a membership initiative and set up mechanisms for its continued development and maintenance. This presentation contrasts "old" and "new" TEI, and gives a brief overview of some specific technical enhancements to the system, in particular the use of a class system to facilitate expansion and customization, and also new features supporting internationalization.

Keywords: TEI, XML, encoding standards

Joint work of: Burnard, Lou; Rahtz, Sebastian

Introducing Xaira

Lou Burnard (Oxford University Computing Services, GB)

This presentation introduces Xaira, an XML Aware Retrieval and Indexing Architecture. Xaira is a system for providing linguistically-motivated search facilities for collections of richly encoded XML documents, originally developed at Oxford for use with the British National Corpus, now re-engineered for use with any collection of XML data. It combines text searching facilities with traditional XML query features. The software is available under an open source licence. The presentation briefly describes the motivation and architecture of the system, and includes examples of how it may be used via a Windows GUI.

Keywords: XML, query languages, linguistic data, search engines

Joint work of: Burnard, Lou; Dodd, Tony

New tricks from an old dog: An overview of TEI P5

Lou Burnard (Oxford University Computing Services, GB)

This paper presents an update on the current state of development of the Text Encoding Initiative's Guidelines for Electronic Text Encoding and Interchange. Since the last major edition in 2002, which saw the conversion of the Guidelines into XML, there has been substantial activity on adding new content in areas of particular interest to historical corpus builders. The TEI has also reinvented itself as a membership initiative and set up mechanisms for the continued development and maintenance of the Guidelines. We contrast "old" and "new" TEI, and give a brief overview of some recent technical enhancements to the system intended to facilitate expansion and customization of the scheme.

Keywords: Text Encoding XML TEI Standards Interchangeability

Full Paper: http://drops.dagstuhl.de/opus/volltexte/2007/1042

XAIRA : software for language analysis

Lou Burnard (Oxford University Computing Services, GB)

This paper describes a software architectiure developed at Oxford University Computing Services (OUCS) over the last decade for the analysis of large or small text corpora, in any language, using rich or only minimal XML markup.

Keywords: XML TEI XAIRA concordance corpus linguistic analysis

Extended Abstract: http://drops.dagstuhl.de/opus/volltexte/2007/1043

Migrating Legacy Electronic Texts at the Oxford Text Archive

James Cummings (Oxford University Computing Services, GB)

The Oxford Text Archive is one of the oldest archives of academic electronic texts – it celebrated its thirtieth birthday this year. Unfortunately, in the early days before TEI XML became such a useful standard, many deposits were received in a large variety of unstandardised or entirely unique individual markup schemes. I will talk on the problems of legacy data migration, individualistic forms of markup, and present a few of the results of a pilot study the OTA undertook to investigate some of its resources, concentrating on those in unknown and early formats. The pilot study found a number of vaguely-termed 'Unknown Markup' or 'Plain Text' formats which it was able to reclassify into known markup formats, and suggested a number of possible routes for conversion of a subset of this legacy material. One example which was used as a case study, for the conversion of COCOA-encoded verse drama will be shown.

Keywords: Oxford Text Archive, OTA, Legacy Data Migration, COCOA, TEI XML

Competing demands of size, speed, and annotation with historical corpora

Mark Davies (Brigham Young University, USA)

Three issues facing any corpus are those of size, speed, and extensive annotation (e.g. lemma or part of speech). While it is quite easy to create a corpus that achieves two of the three goals, achieving all three with the same corpus (architecture) is quite difficult.

This presentation will focus on three historical corpora that we have created, where we feel that all three goals have been met. These include the Corpus del Espanol (100 million words, 1200s-1900s, www.corpusdelespanol.org), the Corpus do Portugues (45 million words, 1300s-1900s, www.corpusdoportugues) and the Corpus of Historical English (37 million words, 1000s-1900s, view.byu.edu/che). The architecture for these corpora is also similar to that created for the VIEW interface to the British National Corpus (view.byu.edu). Finally, we have proposed to create a 200 million word corpus of English (1500s-1900s), which will be based on the same architecture.

The architecture for these corpora relies on relational databases, and the power of the architecture is due to the indexing on the tables and to the powerful SQL joins between the tables. In terms of size, the architecture allows corpora of 200+ million words or more. In terms of speech, the architecture allows for very fast queries of the corpora - typically less than one or two seconds for even complex queries on the largest of corpora. Finally, the architecture allows for a very wide range of annotation - whether it is part of speech, lemma, semantic information, etc. Due to the modularized relational database architecture, there is absolutely no "performance hit" as an unlimited number of levels of annotation are added to the corpus.

All of this results in a corpus (interface) that provides the end user with a wide range of query types. Users can search by word, phrase, wildcards, part of speech, lemma, and semantically-related words. They can easily find the frequency across different time periods and in different registers, and see the results either as tables or charts. They can limit the queries by the frequency in different time periods and registers. Users can also easily find the collocates of a given word (raw frequency or z-score), and in one step they can compare the collocates of competing words, or the collocates of a given word in different time periods or registers. Finally, due to the modularized nature of the relational database architecture, they can easily use the semantic information in linked thesauruses (or WordNets) to search the corpus by semantic fields, and they can also create "customized lists" of these words for re-use in subsequent sessions.

Keywords: Corpus, corpora, historical, diachronic, annotation, relational databases

Participles in Old Bulgarian: Issues of annotation

Mila Dimitrova-Vulchanova (Norwegian Univ. of Science & Technology, N)

In this presentation we address participle data from Old Bulgarian based in a specialized corpus of nominal expressions from Codex Suprasliensis (http://www.hf.ntnu.no/hf/adm/forskning/prosjekter/balkansim/databases.html) and problems related to strategies for POS-tagging and annotation. We review existing solutions in comparable historical corpora (the Penn-Helsinki corpus of Middle English; the ACT corpus) and modern corpora (The BNC) and discuss the advantages and problems arising from adopting a specific strategy in the latter corpora. Our conclusion is that while some very general strategy (e.g., underspecification, ambiguity tags, word-specific tags) can carry over from a particular set of data to another and across languages, most decisions remain language-specific and require in-depth study of the properties of a specific paradigm and its syntactic manifestations.

Joint work of: Dimitrova-Vulchanova, Mila; Vulchanov, Valentin

Participles in Old Bulgarian: issues of annotation

Mila Dimitrova-Vulchanova (Norwegian Univ. of Science & Technology, N)

In this presentation we address participle data from Old Bulgarian based in a specialized corpus of nominal expressions from Codex Suprasliensis (http://www.hf.ntnu.no/hf/adm/forskning/prosjekter/balkansim/databases.html) and problems related to strategies for POS-tagging and annotation.

We review existing solutions in comparable historical corpora (the Penn-Helsinki corpus of Middle English; the ACT corpus) and modern corpora (The BNC) and discuss the advantages and problems arising from adopting a specific strategy in the latter corpora. Our conclusion is that while some very general strategy (e.g., underspecification, ambiguity tags, word-specific tags) can carry over from a particular set of data to another and across languages, most decisions remain language-specific and require in-depth study of the properties of a specific paradigm and its syntactic manifestations. For praticiples in particular we believe that a two-level approach with annotation for POS and morphological properties separately from syntactic annotation will be most reasonable.

Keywords: Annotation strategies, POS tagging, underspecification, Old Bulgarian, participle constructions

Joint work of: Dimitrova-Vulchanova, Mila; Vulchanov, Valentin

Are there any Easy Corpora Solutions for a Digitisation Department?

Milena Dobreva (Bulgarian Academy of Sciences, BG)

The talk discusses some of the current tasks of the Digitisation Heritage department and raises issues related to corpora creation and exploitation.

Keywords: Digitisation, newspapers, archival documents, 19 century, Bulgarian, Russian, language variety

A Multifunctional Historical Document Research System

Eva Dyllong (Universität Duisburg-Essen, D)

In this talk, the key components of a multifunctional historical document research system are discussed. An ongoing project which aims at creating a representative corpus of documents that reflect the impact of the German philosopher Friedrich Nietzsche in the period 1865-1945 forms the case study for the system.

The realisation of the system includes several working fields: the collection of relevant historical documents, the digitization and choice of a suitable library-oriented data standards for archival storage, the design and implementation of a database, the development of fuzzy techniques for searching on documents with a non-standard orthography, the preparation of communication, annotation and visualisation tools, and the design of a user interface adapted for heterogeneous user group ranging from interested amateurs to experts.

Keywords: Literature database, digitization and archival storage

Extended Abstract: http://drops.dagstuhl.de/opus/volltexte/2007/1045

GerManC - Towards a Methodology for Constructing and Annotating Historical Corpora

Astrid Ensslin (The University of Manchester, GB)

Our paper focuses on the one hand on the challenges posed by the structural variability, flexibility and ambiguity found in historical corpora and evaluates methods of dealing with them on the other.

We are currently engaged in a project which aims to compile a representative corpus of German for the period 1650-1800. Looking at exemplary data from the first stage of this project (1650-1700), which consists of newspaper texts from this period, we first aim from the perspective of corpus linguistics to identify the problems associated with the morphological, syntactical and graphemic peculiarities that are characteristic of that particular stage. Specific phenomena which significantly complicate automatic tagging, lemmatisation and parsing include, for instance, 'abperlende' (Admoni 1980; Demske-Neumann 1990), i.e. complex and often asyndetic syntax; non-syntactic, prosodic, virgulated punctuation (Demske et al. 2004; cf. Stolt 1990), inflectional variability (e.g. Admoni 1990; Besch & Wegera 1987), as well as partly unsystematic and almost experimental allomorphic and allographic (Kettmann, 1992) diversity.

Secondly, we outline a methodology which is intended to facilitate the construction and annotation of such corpora which antedate linguistic standardisation. This is informed by 'conventional' and innovative tagging techniques and tools, which are evaluated in terms of utility and accuracy. Finally, we attempt to evaluate the degree to which annotation tools for specialist corpora of this kind can be developed which will substitute for manual or semi-automated annotation.

Keywords: Early Modern German, newspaper corpus, GerManC, variation, annotation, tagging

Joint work of: Ensslin, Astrid; Durrell, Martin; Bennett, Paul

Extended Abstract: http://drops.dagstuhl.de/opus/volltexte/2007/1046

TEI and Microsoft: a marriage made in...

Tomaž Erjavec (Jozef Stefan Institute - Ljubljana, SLO)

In several on-going projects we were faced with the dilemma of how to reconcile our goal of delivering standardly encoded historical documents, yet have the actual editing and annotation performed by researchers and students who had no knowledge of XML and TEI, and, for the most part, no interest in learning them. The solution we developed consists of allowing the annotators use familiar and flexible editors, such as Microsoft Word (for structural annotation of documents) and Excel (for word-level linguistic annotation) and automatically

converting these into TEI. Given the unconstrained nature of such editors this sounds like a recipe for disaster. But the solution crucially depends on a dedicated Web service, to which the annotators can up-load their files; these are then immediately converted to XML/TEI and from it back to a visual format, either HTML or Excel XML, and presented to the annotators. These then get immediate feedback about the quality of their encoding in the source, and can thus correct errors before they accumulate; and the responsibility for the correct encoding rests with the annotators, rather than with the developers of the conversion procedure. The paper describes the web service and details its use in three projects. The main conclusions are that the proposed solution is appropriate for shallow encodings, and nevertheless does require producing detailed annotation guidelines.

Keywords: Text encoding, manual annotation, open standards, XML, Microsoft

Full Paper: http://drops.dagstuhl.de/opus/volltexte/2007/1047

Retrieval in text collections with historic spelling

Andrea Ernst-Gerlach (Universität Duisburg-Essen, D)

In the presentation a new approach for retrieval in texts with non-standard spelling will described, which is important for historic texts e. g. in English or German. The non-standard spelling produces problems when searching in the historic parts of digital libraries. Most users will enter basic form search terms in their contemporary language which differs from the historic language used in the documents. In order to solve this problem, our project deals with the research and development of a search engine where the user can formulate queries in contemporary language for searching in documents with an old spelling that is possibly unknown to the user.

For this purpose, an automatic approach for the generation of spelling variants based on evidences of contemporary full word forms and their corresponding spelling variants will be presented. The algorithm produces a set of probabilistic rules. These probabilities can be considered for ranking in the retrieval stage. The overall architecture of the search system will be described. Given a search term in its basic word form, we use a dictionary of contemporary German for finding all full word forms. Then we apply generated transformation rules (derived from training data) for generating historic spelling variants.

Keywords: Retrieval, digital libraries, non-standard spelling

Digital historical corpora and the Web as corpus

Stefan Evert (Universität Osnabrück, D)

At first sight, digital historical corpora and the Web as corpus would seem to represent opposite ends of the scale of approaches to corpus building.

While the former are relatively small collections of texts with rich manual annotations, the latter is big and messy and its proponents will happily compromise on annotation quality if this allows them to accumulate larger amounts of data.

Surprisingly, a closer look shows that the two fields face many similar problems, including the (semi-)automatic annotation of nonstandard language, the identification and representation of document variants, and retrieval with approximate ("fuzzy") corpus queries. In my talk, I will discuss a number of common problems and suggest ways in which the two approaches can benefit from each other.

Keywords: Web as corpus, annotation, corpus query, document variants

Information Visualization for Corpora

Jean-Daniel Fekete (INRIA Futurs - Orsay, F)

Interfaces for accessing digital corpora are currently based on web forms or search queries. We present here a new type of interaction to better understand and navigate these corpora: Information Visualization. Relying on the human visual perception, Information Visualization enables users to perceive important features very quickly and to perform queries using interaction instead of textual syntax.

We present three applications: one on a corpus of 1000 historic documents encoded using TEI/XML and analyzed using several visualization tools, a second designed to encode and navigate in modern draft manuscripts and a third for accessing a large digital library using a zoomable user interface.

Keywords: Information Visualization, navigation, interaction, dynamic queries

The Middle High German Text Archive

Kurt Gärtner (Universität Trier, D)

The medieval period of the German language in Middle and Southern German speaking areas from ca. 1050 to 1350/1400 is called Middle High German (MHG). The language of this period has been transmitted in a great number of literary texts, among them the famous classical literary works from around 1200 like the Parzival, the Tristan, and the Nibelungenlied, and also famous works of Germanic law like the Sachsenspiegel.

Many texts of this period are available on the internet, however, most of them do not meet the requirements of academic users. The need of reliable texts arose when we began to plan and then to compile the new MHG Dictionary (MHGD). Moreover, reliability of all electronic resources is essential for being used in literary and linguistic research because the new medium should offer the texts as reliable as the traditional book medium, otherwise we would have to wait to long before serious humanities' scholars will use them.

The compilation of a dictionary on historical principles requires in general a threefold historical corpus. This applies also to the new MHGD for which the following electronic resources have been created:

1. The already existent dictionaries to MHG, i.e. the "old" dictionaries, now electronically available, interlinked, and in an optimal state of accuracy (www.MWV.uni-trier.de).
2. The texts published in scholarly editions after the completion of the "old" dictionaries. Out of these texts a sufficient number (ca. 100) together with their edition-inherent material (introduction, critical apparatus, glossaries) have been digitized in a joint NSF/DFG-funded project (www.MHGTA.uni-trier.de).
3. A number of lemmatized texts, out of which an archive of quotations has been drawn (Belegarchiv, not yet available for general use).

The third of these three components consists of a small number of texts which mostly have been chosen from the texts of the second component; for choosing them the following criteria were essential: time when a text has appeared, place where it was composed, and genre to which it belongs to. The source texts of the "old" dictionaries are also classified according to these features.

The paper deals with various aspects of interlinking these components and preparing them for complex and sophisticated research questions and retrieval which go far beyond that what the book medium with its alphabetical order as the main way of structuring and searching the vocabulary of a language period would offer.

Keywords: Digital Text Archives, Lexicography, Middle High German, Dictionaries

Technologies for Processing Historical German Texts

Markus Heller (CIS - Universität München, D)

Historians have long been dreaming of electronic source editions, which obviate material access and preserve originals as well as facilitate simultaneous access by multiple researchers. First approaches like the ECHO (European Cultural Heritage Online) project demonstrate that it is possible to offer historical sources online and that this way of presentation and publication is generally well usable in historical research.

Our beginning cooperation with the Bayerische Staatsbibliothek (BSB) may lead to the production of a massive corpus of automatically transcribed (OCR'ed) scans of historical print documents. Since the BSB is not the only library with scanning activities, but given that the Deutsche Forschungsgemeinschaft has funded a voluminous retro digitization project, the overall next step is to convert the scanned images into semistructured corpora and to make them accessible and usable by historians, paleographers and linguists with all of their access and query methods.

We understand that the properties of the produced documents, but also the scientific interest of the mentioned user disciplines necessitate access technologies which are not provided by traditional approaches. The technologies which we have been developing in the fields of OCR postcorrection, XML indexing and fast approximative search in combination seem to fulfill all the requirements for an efficient and versatile management framework of an XML-encoded corpus of historical documents.

We will present current and ongoing work in the mentioned fields, yet with a special focus on historical corpora. We will also display the insights laid out in a recent paper: Together with the German linguistics department of Munich university our working group has developed a linguistic workbench for early modern high German texts to heuristically classify word forms, using a Brill-algorithm based distance measure to cover the spelling variance. The paper has been accepted for the IJCAI-2007 Workshop on Analytics for Noisy Unstructured Text Data in Hyderabad, India.

Joint work of: Heller, Markus; Klaus Schulz

Information Access to Historical Documents from the Early New High German Period

Markus Heller (CIS - Universität München, D)

With the new interest in historical documents insight grew that electronic access to these texts causes many specific problems. In the first part of the paper we survey the present role of digital historical documents. After collecting central facts and observations on historical language change we comment on the difficulties that result for retrieval and data mining on historical texts. In the second part of the paper we report on our own work in the area with a focus on special matching strategies that help to relate modern language keywords with old variants. The basis of our studies is a collection of documents from the Early New High German period. These texts come with a very rich spectrum on word variants and spelling variations.

Keywords: Historical documents, information access, Early New High German, historical language, information retrieval, word similarity, approximate matching

Joint work of: Hauser, Andreas; Heller, Markus; Leiss, Elisabeth; Schulz, Klaus U.; Wanzeck, Christiane

Toward creation of a Digital Corpus of Bulgarian Dialects

Nikola Ikonomov (Bulgarian Academy of Sciences, BG)

In this presentation, we describe our considerations related to the creation of a digital corpus of Bulgarian dialects.

The dialectological archive of Bulgarian language consists of more than 250 audio tapes. All tapes were recorded between 1955 and 1965 in the course of regular dialectological expeditions throughout the country. The records typically contain interviews with inhabitants of small villages in Bulgaria. The topics covered are usually related to such issues as birth, everyday life, marriage, family relationship, death, etc. Only a few tapes contain folk songs from different region of the country.

Taking into account the progressive deterioration of the magnetic media and the realistic prospects of data loss, the Institute for Bulgarian Language at the Academy of Sciences launched in 1997 a project aiming at restoration and digital preservation of the dialectological archive. Within the framework of this project more than the half of the records was digitized, de-noised and stored on digital recording media. Since then restoration and digitization activities are done in the Institute on a regular basis. As a result a large collection of sound files has been gathered. Our further efforts are aimed at the creation of a digital corpus of Bulgarian dialects, which will be made available for phonological and linguistic research. Such corpora typically include besides the sound files two basic elements: a transcription, aligned with the sound file, and a set of standardized metadata that defines the corpus.

In our work we will present considerations on how these tasks could be realized in the case of the corpus of Bulgarian dialects. Our suggestions will be based on a comparative analysis of existing methods and techniques to build such corpora, and by selecting the ones that fit closer to the particular needs. Our experience can be used in similar institutions storing folklore archives, history related spoken records etc.

Keywords: Digital corpus, dialects, dialectology, linguistics, corpus linguistics, transcription, phonetics, phonology

Joint work of: Ikonomov, Nikola; Dobreva, Milena

CREATION OF A DIGITAL CORPUS OF BULGARIAN DIALECTS

Nikola Ikonomov (Bulgarian Academy of Sciences, BG)

The paper presents our considerations related to the creation of a digital corpus of Bulgarian dialects. The dialectological archive of Bulgarian language consists of more than 250 audio tapes. All tapes were recorded between 1955 and 1965 in the course of regular dialectological expeditions throughout the country. The

records typically contain interviews with inhabitants of small villages in Bulgaria. The topics covered are usually related to such issues as birth, everyday life, marriage, family relationship, death, etc. Only a few tapes contain folk songs from different regions of the country.

Taking into account the progressive deterioration of the magnetic media and the realistic prospects of data loss, the Institute for Bulgarian Language at the Academy of Sciences launched in 1997 a project aiming at restoration and digital preservation of the dialectological archive. Within the framework of this project more than the half of the records was digitized, de-noised and stored on digital recording media. Since then restoration and digitization activities are done in the Institute on a regular basis. As a result a large collection of sound files has been gathered.

Our further efforts are aimed at the creation of a digital corpus of Bulgarian dialects, which will be made available for phonological and linguistic research. Such corpora typically include besides the sound files two basic elements: a transcription, aligned with the sound file, and a set of standardized metadata that defines the corpus. In our work we will present considerations on how these tasks could be realized in the case of the corpus of Bulgarian dialects. Our suggestions will be based on a comparative analysis of existing methods and techniques to build such corpora, and by selecting the ones that fit closer to the particular needs. Our experience can be used in similar institutions storing folklore archives, history related spoken records etc.

Keywords: Phonology, corpus, corpus linguistics, audio archive, digitization, restoration, metadata, alignment

Joint work of: Ikonomov, Nikola; Dobreva, Milena

Extended Abstract: http://drops.dagstuhl.de/opus/volltexte/2007/1048

A Cross-Language Approach to Historic Document Retrieval

Jaap Kamps (University of Amsterdam, NL)

Our cultural heritage, as preserved in libraries, archives and museums, is made up of documents written many centuries ago.

Large-scale digitization initiatives, like DigiCULT, make these documents available to non-expert users through digital libraries and vertical search engines.

For a user, querying a historic document collection may be a disappointing experience. Natural languages evolve over time, changing in pronunciation and spelling, and new words are introduced continuously, while older words may disappear out of everyday use. For these reasons, queries involving modern words may not be very effective for retrieving documents that contain many historic terms.

Although reading a 300-year-old document might not be problematic because the words are still recognizable, the changes in vocabulary and spelling can make

it difficult to use a search engine to find relevant documents. To illustrate this, consider the following example from our collection of 17th century Dutch law texts. Looking for information on the tasks of a lawyer (modern Dutch: *advocaat*) in these texts, the modern spelling will not lead you to documents containing the 17th century Dutch spelling variant *advocaet*.

Since spelling rules were not introduced until the 19th century, 17th century Dutch spelling is inconsistent. Being based mainly on pronunciation, words were often spelled in several different variants, which poses a problem for standard retrieval engines.

We therefore define Historic Document Retrieval (HDR) as the retrieval of relevant historic documents for a modern query. Our approach to this problem is to treat the historic and modern languages as different languages, and use cross-language information retrieval (CLIR) techniques to translate one language into the other.

Keywords: Historic Documents, Information Retrieval, Spelling variation, Modernizing Spelling, 17th Century Dutch

Extended Abstract: http://drops.dagstuhl.de/opus/volltexte/2007/1049

Full Paper:
http://staff.science.uva.nl/∼kamps/publications/2006/kool:cros06.pdf

See also: Marijn Koolen, Frans Adriaans, Jaap Kamps, and Maarten de Rijke. A cross-language approach to historic document retrieval. In Mounia Lalmas et al., editor, Advances in Information Retrieval: 28th European Conference on Information Retrieval (ECIR 2006), volume 3936 of Lecture Notes in Computer Science, pages 407-419. Springer Verlag, Heidelberg, 2006.

The cadastral register of the town Wismar (1677-1680) – structure mining in historical documents

Meike Klettke (Universität Rostock, D)

The cadastral registers of the town Wismar (1677 &#150; 1838) contain a lot of historical information that could be useful for other systems, for instance for historical geographical systems.

The documents are electronically available but they are unstructured text documents. For using it, the information has to be analysed and stored in databases. In the talk, a method is suggested that associates markup information to the text items in the cadastral registers. For that, in a first step layout structures and implicit structural characteristics are analysed with specialized parsers. Second, the historical texts itself are analysed with dictionary-based methods using phonetic encoding and Levenshtein distance for determining the similarity between tokens in the text and entries in the dictionaries. In the third step, semantic rules are applied for checking and improving the results. The above enumerated methods enrich the texts with markup &#150; so that XML documents are generated, these XML documents are finally stored in relational databases (DB2).

Full Paper:
http://www.informatik.uni-rostock.de/~meike/Wismarer_Grundbuecher.pdf

Detecting the relationship between languages using bioinformatics methods

Ulf Leser (HU Berlin, D)

Languages, as well as biological species, change over time. If enough changes have accumulated in a given population, new languages (or dialects) emerge, in a similar manner as new species emerge during evolution. Based on this similarity, methods from bioinformatics can be used to estimate the historic relationships between languages based on a comparison of text testimonies. In the first half of the talk, we present methods for deriving language trees from similarity of words and texts. We present the result of a small study on various versions of the "Vater Unser" which highlight properties and pitfalls of these methods. In particular, language contact, which breaks the hierarchical organization of languages in trees, is a common phenomenon that must be taken into consideration. In the second half, we present the results of a simulation study where we estimated the effect of language contact on the quality of derived phylogenetic language trees and networks using four different algorithms. Our findings suggest that, although tree-based methods can cope well with a small degree of "noise" (i.e. contact), network-based methods are superior in the general case.

Keywords: Language history; phylogeny; language evolution

Joint work of: Leser, Ulf; Lüdeling, Anke; Hochmuth, Mirko

Rule-based search in historical text databases - Visualization techniques

Wolfram Luther (Universität Duisburg-Essen, D)

The talk describes several techniques used to visualize among other aspects the productivity of rule sets in deriving non-standard spellings. The treemap or similar visualizations help find typical replacement sequences depending on the localization of the spellings and their epoch. The study conducted proves that treemaps ease the understanding of rule hierarchies, the detection of productive and non productive rules and the evaluation of a rule's importance. They also provide better search performance.

An interactive visualization over a map is showing isoglosses running between different regions of Germany and clusters text samples of different epochs and their writings.

Furthermore, allograph variants are displayed using adequate data types.

Keywords: Non-standard spellings, visualization techniques, treemaps, rule base optimization

Extended Abstract: http://drops.dagstuhl.de/opus/volltexte/2007/1051

DeutschDiachronDigital - a diachronic corpus of German

Anke Lüdeling (HU Berlin, D)

In my talk I present the concept for the design and the architecture of a diachronic corpus of German, as developed in the project initiative DeutschDiachronDigital.

Keywords: Corpus, database

DeutschDiachronDigital - A Diachronic Corpus of German

Anke Lüdeling (HU Berlin, D)

The talk describes the design and the architecture of a diachronic corpus of German.

Keywords: Historical corpora

Extended Abstract: http://drops.dagstuhl.de/opus/volltexte/2007/1050

Joseph Wright's EDD Computerisd: architecture and retrieval

Manfred Markus (Universität Innsbruck, A)

After a brief introduction of the project SPEED (Spoken English in Early Dialects), started officially for three years on 1 July 2006, the paper will describe the eight parameters that the entries of the dictionary consist of, from lemma to comment, and then demonstrate with a few examples where the problems of database parsing are. As a result of philological interpretation, the eight parameters and their often deviant and heterogeneous content will be featured as nineteen database items, which are mapped by the search mask. In my presentation I will focus on some of the items which have turned out to be particularly complex, namely dialect and meaning.

Keywords: Corpus lingusitics, historical English, dialects, lexicography

Joint work of: Markus, Manfred; Heuberger, Reinhard; Onysko, Alexander

Joseph Wright's English Dialect Dictionary (1898-1905) Computerised: architecture and retrieval routine

Manfred Markus (Universität Innsbruck, A)

The Innsbruck government-funded project SPEED (Spoken English in Early Dialects), scheduled for 2006 to 2009, has the aim of digitising and evaluating the famous English Dialect Dictionary by Joseph Wright (1898-1906). This paper topicalises the value of the electronic version of the dictionary and problems of its complex architecture, as well as the retrieval routine aimed at. The paper is an elaborated version of the Powerpoint presentation delivered at the conference. First of all, I try to prove the great value of Wright's dictionary from the point of view of English studies. On the other hand, given the mixed nature of the participants of the Dagstuhl conference, the paper tackles interface problems typically arising when printed texts are computerised, problems ranging from "normalisation" to aspects of parsing and of the design of the query mask.

Keywords: Corpus linguistics, dialectology, spoken English, history of English, lexicography, computer philology, normalisation, web design

Full Paper: http://drops.dagstuhl.de/opus/volltexte/2007/1052

The Regensburg Diachronic Corpus of Russian. Design Considerations and State of the Art

Roland Meyer (Universität Regensburg, D)

The Diachronic Corpus of Russian currently being compiled at Regensburg University is intended as a corpus for linguistic research. It uses the ACT tool (Ribarov et al. 2004) for document storage and data entry. The talk provides a discussion of design considerations, with special attention on lemmatization and POS tags, and a presentation of the present stage of the corpus.

A Diachronic Corpus of Russian for Linguistic Research: Design and Tools for Construction and Retrieval

Roland Meyer (Universität Regensburg, D)

Diachronic corpora face a number of specific difficulties connected to their design, construction and exploitation. The current paper presents and argues for the solutions to these issues adopted in the Regensburg diachronic corpus of Russian, a source currently being compiled for linguistic research into historical language development, especially in the realm of morphosyntax.

Keywords: Russian corpus diachronic ACT XML relational linguistic annotation

Searching in text databases with nonstandard orthography

Thomas Pilz (Universität Duisburg-Essen, D)

In this paper we present research results of the recent project "Rule based search in text data bases with non-standard orthography". There are numerous steps involved from facsimile to searchable text-document. This paper focuses on techniques to ensure better retrieval results on historical texts with non-standard spellings. Historical documents - especially those in black letter fonts - encourage recognition errors. Adequate preparation of the image sources prior to OCR can successfully reduce the amount of misinterpretation of characters. Furthermore, the application of a search engine with categorized distance measures between user interface and text database can help to enhance retrieval results. Specific metrics cover problems in optical character recognition, transcription and historical spelling variation. With a synoptic view interface the users can be kept completely unaware of the methods applied after their queries.

Keywords: Rule based search, Optical character recognition, spelling variation, edit distance

Extended Abstract: http://drops.dagstuhl.de/opus/volltexte/2007/1053

Tagging historical corpora - the problem of spelling variation

Paul Rayson (Lancaster University, GB)

In this presentation, we explain how we have sought to overcome the spelling issue as part of our work in respect to the development of an historical semantic tagger. The historical tagger is based on an existing semantic tagger, the UCREL Semantic Annotation System, which automatically tags modern English data (spoken and written) with semantic information. More specifically, we will explain how we have developed a VARiant Detector (henceforth VARD) as a means of detecting and normalising spelling variants to their modern equivalent so that USAS can begin to annotate historical data from Shakespeare onwards. We will also demonstrate a new version of VARD.

Keywords: Corpus annotation and retrieval

Joint work of: Rayson, Paul; Archer, Dawn; Baron, Alistair; Smith, Nicholas

Tagging Historical Corpora - the problem of spelling variation

Paul Rayson (Lancaster University, GB)

Spelling issues tend to create relatively minor (though still complex) problems for corpus linguistics, information retrieval and natural language processing tasks that use 'standard' or modern varieties of English.

For example, in corpus annotation, we have to decide how to deal with tokenisation issues such as whether (i) periods represent sentence boundaries or acronyms and (ii) apostrophes represent quote marks or contractions (Grefenstette and Tapanainen, 1994; Grefenstette, 1999). The issue of spelling variation becomes more problematic when utilising corpus linguistic techniques on non-standard varieties of English, not least because variation can be due to differences in spelling habits, transcription or compositing practices, and morpho-syntactic customs, as well as "misspelling". Examples of non-standard varieties include:

- Scottish English1 (Anderson et al., forthcoming), and dialects such as Tyneside English2 (Allen et al., forthcoming)
- Early Modern English (Archer and Rayson, 2004; Culpeper and Kytö, 2005)
- Emerging varieties such as SMS or CMC in weblogs (Ooi et al., 2006)

In the Dagstuhl workshop we focussed on historical corpora. Vast quantities of searchable historical material are being created in electronic form through large digitisation initiatives already underway e.g. Open Content Alliance3, Google Book Search4, and Early English Books Online5. Annotation, typically at the part-of-speech (POS) level, is carried out on modern corpora for linguistic analysis, information retrieval and natural language processing tasks such as named entity extraction. Increasingly researchers wish to carry out similar tasks on historical data (Nissim et al, 2004). However, historical data is considered noisy for tasks such as this. The problems faced when applying corpus annotation tools trained on modern language data to historical texts are the motivation for the research described in this paper.

Previous research has adopted an approach of adding historical variants to the POS tagger lexicon, for example in TreeTagger annotation of GerManC (Durrell et al, 2006), or "back-dating" the lexicon in the Constraint Grammar Parser of English (ENGCG) when annotating the Helsinki corpus (Kytö and Voutilainen, 1995).

Our aim was to develop an historical semantic tagger in order to facilitate similar studies on historical data to those that we had previously been performing on modern data using the USAS semantic analysis system (Rayson et al, 2004). The USAS tool relies on POS tagging as a prerequisite to carrying out semantic disambiguation. Hence we were faced with the task of retraining or back-dating two tools, a POS tagger and a semantic tagger. Our proposed solution incorporates a corpus pre-processor for detecting historical spelling variants and inserting modern equivalents alongside them. This enables retrieval as well as annotation tasks and to some extent avoids the need to retrain each annotation tool that is applied to the corpus. The modern tools can then be applied to the modern spelling equivalents rather than the historical variants, and thereby achieve higher levels of accuracy.

The resulting variant detector tool (VARD) employs a number of techniques derived from spell-checking tools as we wished to evaluate their applicability to historical data. The current version of the tool uses known-variant lists, SoundEx, edit distance and letter replacement heuristics to match Early Modern English variants with modern forms. The techniques are combined using a scoring mechanism to enable preferred candidates to be selected using likelihood values. The current known-variant lists and letter replacement rules are manually created. In a cross-language study with English and German texts we found that similar techniques could be used to derive letter replacement heuristics from corpus examples (Pilz et al, forthcoming). Our experiments show that VARD can successfully deal with:

- Apostrophes signalling missing letter(s) or sound(s): 'fore ("before"), hee'l ("he will"),
- Irregular apostrophe usage: again'st ("against"), whil'st ("whilst")
- Contracted forms: 'tis("it is"), thats ("that is"), youle ("you will"), t'anticipate ("to anticipate")
- Hyphenated forms: acquain-tance ("acquaintance")
- Variation due to different use of graphs: <v>, <u>, <i>, <y>: aboue ("above"), abyde ("abide")
- Doubling of vowels and consonants -e.g. <-oo-><-ll>: triviall ("trivial")

By direct comparison, variants that are not in the modern lexicon are easy to identify, however, our studies show that a significant portion of variants cannot be discovered this way. Inconsistencies in the use of the genitive, and 'then' appearing instead of 'than' or vice versa require contextual information to be used in their detection. We will outline our approach to resolving this problem, by the use of contextually-sensitive template rules that contain lexical, grammatical and semantic information.

Footnotes

1. http://www.scottishcorpus.ac.uk/
2. http://www.ncl.ac.uk/necte/
3. http://www.opencontentalliance.org/
4. http://books.google.com/
5. http://eebo.chadwyck.com/home

Keywords: Corpus annotation, spelling variation, historical variants

Joint work of: Rayson, Paul; Archer, Dawn; Baron, Alistair; Smith, Nicholas

Extended Abstract: http://drops.dagstuhl.de/opus/volltexte/2007/1055

TextGrid – Grid-based philological infrastructure

Thorsten Vitt (TU Darmstadt, D)

TextGrid is a project to provide an infrastructure for the collaborative edition, publication, annotation and analysis of texts. It is part of the German D-Grid project, which builds a generic Grid infrastructure for the sciences and humanities in Germany.

TextGrid aims at the integration of future and existing projects in literary and linguistic computing. This includes integrated search and retrieval facilities over all edition and corpus projects using our infrastructure, but also linking with supplementary material like dictionaries of the corresponding dialect and period.

As part of the infrastructure, TextGrid will provide an integrated set of tools and services for the creation and work with digital editions and corpora. We are open for cooperation with related projects, and we commit ourselves to open standards, open source software, and an open architecture.

Keywords: EHumanities, Digital Humanities, Grid computing, tools, services, digital editions, corpora

Joint work of: Vitt, Thorsten; Jannidis, Fotis

Guideline: Multiple Hierarchies

Andreas Witt (Universität Tübingen, D)

As the title of the Dagstuhl Seminar “Digital Historical Corpora - Architecture, Annotation, and Retrieval” already suggests, corpus architecture and corpus annotation is an important topic for representing (historical) texts. Especially the limitation of SGML-based markup languages to tree structured annotations raises a special problems when dealing with manuscripts: How is it possible to represent overlap. This problem was addressed by the Text Encoding Initiative (TEI) and by several scholars. This text gives an overview of several techniques for handling the overlap problem.

Keywords: XML, Overlapping Markup

Full Paper: http://drops.dagstuhl.de/opus/volltexte/2007/1040

Design and Applications of Polimatth - a Small Parallel Diachronic Bible Corpus of Polish

Amir Zeldes (HU Berlin, D)

These two presentations give a brief introduction to Polimatth, a small parallel corpus containing the Gospel of Matthew from two Polish Bible translations: the Biblia Gdañska (1606) and the Biblia Warszawska (1975). The first presentation deals with problems that were encountered in the process of setting up the corpus and the solutions that were chosen for them, with a special focus on the morphological annotation employed in the corpus and its applications. The second presentation describes a study conducted on the corpus using techniques from the field of example based machine translation.

Keywords: Polish, Parallel Corpora, Corpus, Diachronic, Historical, Bible, EBMT, Example Based Machine Translation