**BANSKI — Search**

<0 figures>

**Search for Needles in DH Haystacks Continued: Zooming In with Corpus Query Tools—Tutorial Proposal for DH2015**

Bański, P., Diewald, N., Hanl, M., Kupietz, M. and Witt, A.

The proposed tutorial builds on the success of the ‘Looking for Needles in DH Haystacks’ tutorial that we taught at DH2013 in Lincoln.

Judging by the yearly increase in the number of DH projects and presentations related either to retrieving documents or records on the basis of their metadata, or to extracting structured information from the often unstructured or semistructured data commonly found in the humanities, there is a growing need for a pill-sized introduction package that we will be happy to offer at DH2015. For this purpose, we have prepared a ‘reloaded’ version of the successful 2013 proposal—a half-day tutorial, focused on textual data and more oriented towards the actual tool use.

The rapid development of the discipline (or, more precisely, disciplines) known as digital humanities has resulted in the ever wider accessibility of digitization methods and, consequently, the steadily growing amount of digitized and interlinked data. However, as in many other disciplines that have followed a similar pattern of development, it turns out that, while the amount of information is growing, the methods for quick, easy, and successful retrieval of that information are either not yet established or not yet sufficiently widespread.

In view of the massive amount of available data, an average DH scholar is confronted with the task of finding a needle in a haystack: while, seemingly, everything is there structured, interlinked, and ready to be used, and while well-known query mechanisms exist and have been used for years in other disciplines, the fundamental questions still concern the best way to formulate the particular research questions, the method most appropriate to the task at hand, or a friendly tool that would provide the relevant results in the desired format and without too steep a learning curve.

The tutorial is going to present state-of-the-art methods in querying textual data, with a focus on use cases commonly found in digital humanities, or envisioned for the near future of this expanding field. It is offered by a team of researchers and coders dealing directly with markup languages, corpus linguistics, and query systems architecture—our most recent project, KorAP, involved building an open-source corpus analysis system able to process immense haystacks of textual data by offering the user a variety of tools both for selecting the documents of interest into so-called virtual collections, and for searching within those collections.

This is not meant to be a tutorial only for linguists, however: we intend to provide an opportunity for the participants to learn and practice how to carry over some well-known methods and techniques from linguistic research, where they have been used for years, onto the broader area of digital humanities, where queries target much more than merely the linguistic properties of texts, but should also address their structural, statistical, and formal properties. We shall focus primarily on the search in metadata, nonannotated data, and structured annotated data (especially TEI-encoded), and to show how to ‘zoom in’ on text collections in order to discover or confirm the existence of interesting data patterns.

Part of the way to ensure closer cooperation among DH researchers may be to provide them with a common language in which they can specify questions asked of a variety of datasets in a variety of structures. The tutorial shall present to the participants one way to address that issue, currently developed within ISO TC37 SC4 ‘Language resource management’, where two of the presenters lead the project *Corpus Query Lingua Franca*.

**Outline**

Main issues addressed by the tutorial:

• What should a text query system for DH in the 21st century look like?

• What kinds of queries should a query system be able to deal with?

• How to efficiently use metadata information?

• How to characterise a modern query language? How to make it fulfil the users’ information needs?

• How should a text corpus be structured in the future?

List of topics (some of them may receive only cursory attention; much depends on the composition of the audience and the demand):

• Forms of digital text.

• Information that can be associated with text and ways of structuring it.

• Annotation formats (HTML, TEI, others), their pros and cons.

• Text corpora (written vs. spoken language, approximated spoken language, aligned data streams).

• Simple full-text search, search with regular expressions.

• Direct search in XML data (XQuery, XPath)—BaseX or Exist.

• Corpus analysis systems (deployed locally or used remotely)—e.g., Corpus, Workbench, Textométrie, KorAP.

• Corpus query languages: why so many? With brief illustrations of use cases.

• Corpus Query Lingua Franca: an attempt to rule them all . . .

**Contact Information**

Piotr Banski <banski@idsmannheim.de>

Nils Diewald <diewald@ids.mannheim.de>

Michael Hanl <hanl@ids.mannheim.de>

Marc Kupietz <kupietz@ids.mannheim.de>

Andreas Witt <witt@idsmannheim.de>

tel. +49 621 1581 410

fax +49 621 1581 200

Postal address:

Institut für Deutsche Sprache

R5, 6-13

68-161 Mannheim

Germany

**Background Information**

*Piotr Bański* is a senior researcher at the Institut für Deutsche Sprache in Mannheim, where he is the project manager of the ‘Corpus Analysis Platform of the Next Generation’ (KorAP), a project financed by the Leibniz Association (Leibniz-Gemeinschaft). He is also a guest lecturer in digital humanities at the University of Warsaw. He served as an elected member of the TEI Technical Council for term 2011–2012 and since 2010 has been involved in the work of the ISO TC37 SC4 committee for Language Resource Management. His latest project within the scope of ISO is work on *Corpus Query Lingua Franca*, within TC37 SC4 Working Group 6. His current interests focus mostly on text encoding, markup languages, as well as the creation and use of robust language resources.

*Nils Diewald* is a research associate at the Institute for German Language (IDS) in Mannheim, currently working as a software developer in the KobRA project (Korpus-basierte Recherche und Analyse mit Hilfe von Data-Mining) of the BMBF (Federal Ministry of Education and Research). Before that, he worked in the KorAP project together with Piotr Bański. He received a B.A. in German philology and text technology and an M.A. in linguistics (with a focus on computational linguistics) from Bielefeld University.

*Michael Hanl* is a research associate at the Institute for the German Language (IDS) in Mannheim. While being enrolled in the M.A. Computational Linguistics programme at Darmstadt University in 2011, he started work at the IDS as a student assistant for Elexiko, the online dictionary, in the field of stand-off annotation evaluation. He joined KorAP in 2013 as a fulltime software developer and is responsible for application- and data security. Additionally Michael Hanl received a B.A. in International Communication and Translation from Hildesheim University.

After finishing his master’s in linguistics in 1997 at Bielefeld University, *Marc Kupietz* worked in different research projects in the area of psycholinguistics, cognitive science, and neural network modelling of human language processing and from 2000 on also in the areas of text technology and information management. One year after receiving his Ph.D. from Bielefeld University in 2003, he started working at the Institut für Deutsche Sprache in Mannheim, where he became responsible for the German Reference Corpus DeReKo, and since 2012 is head of the Corpus Linguistics Programme Area. Marc is co-editor of the series Corpus Linguistics and Interdisciplinary Perspectives on Language (CLIP), and his main research interests concern empirically grounded linguistics and cognitive science, philosophy of science, as well as corpus and text technology.

After graduating from Bielefeld University in 1996, *Andreas Witt* started at this university as a researcher and instructor in text technology. He was heavily involved in the establishment of the Magister and BA programmes in text technology at Bielefeld Universität in 1999 and 2002, respectively. After completing his Ph.D. in 2002 he became an assistant lecturer with the text technology group in Bielefeld. In 2006 he moved to Tübingen University, where he was involved in a project on Sustainability of Linguistic Resources and in projects on the interoperability of language data. Since 2009 he has been a senior researcher at the Institut für Deutsche Sprache (Institute for the German Language) in Mannheim. Andreas is a member of numerous research organizations, including the TEI Special Interest Group ‘TEI for Linguists’. His major research interests deal with questions of the use and limitations of markup languages for the linguistic description of language data.

**Description of Target Audience and Expected Number of Participants**

We expect up to ca. 35 people.

We intend the tutorial for general DH audience: variety is a virtue in this case, because we want to address actual use cases, some of which will surely come from the participants themselves.

**Special Requirements for Technical Support**

Video projector, Internet connection, whiteboard or flip chart, sound system depending on the room’s acoustics.