Workshop “Making effective use of metadata of historical texts and corpora”

7-8 September 2017

List of talks and abstracts

**Jörg Knappen & Katrin Menzel** (Saarland University), *Enrichment of and metadata production for the Royal Society Corpus*

Abstract: tba

---

**Peter Fankhauser** (Institut für Deutsche Sprache, Mannheim), *Visual correlation for exploring paradigmatic language change*

Abstract: Paradigmatic language change occurs when paradigmatically related words with similar usage context rise or fall together. We introduce an approach to explore such paradigmatic change in diachronic corpora by visually correlating two factors: Frequency change and distributional semantics of words. Frequency change is visualized by means of color derived from the slope of a logistic growth curve fitted to the frequency trend. Semantics of words is visualized by positioning them in two dimensions such that words with similar usage contexts are positioned closely together. As a result we get islands of paradigmatically related words with similar color that can act as a guide for exploring language change.

---

**Louianne Ferlier** (Royal Society, London), *The Royal Society Journal Collection: unlocking 300 years of scientific periodicals*

Abstract: tba

---

**Stefania Degaetano-Ortlieb** (Saarland University), *Linguistic profiles of social variables in diachrony*

Abstract: tba
Julie Weeds & Justyna Robinson (University of Sussex), *Distributions of concepts in the Old Bailey Voices Corpus*

Abstract: Among several challenges facing corpus linguistics, two are of particular importance to a historical linguist. The first challenge is to find ways of interrogating conceptual content in large data sets. Previously mainly analyses of grammatical information and basic lexical co-occurrences have been carried out and it is still not clear how to best access conceptual content. The other challenge is to include metadata in analysing and explaining observed linguistic patterns. So far, the few available meta-linguistically-tagged corpora provided insights into variation across text types and time. However, analysis at a more fine-grained level such as age, gender, social class of speaker/author has been hardly possible because of lack of availability of this information.

In the current presentation, we address the above-mentioned challenges by analysing functional (i.e. grammatical and sociolinguistic) distribution of concepts in the Old Bailey Proceedings as represented in the Old Bailey Voices Corpus (OBVC).¹

The OBVC is a unique database because it contains wealth of sociolinguistics information on the context of speech and demographics of a speaker. Additionally, the OBVC represents a historically real, yet linguistically-controlled dataset restricted to one genre. Since the OBVC takes a consistent generic form (the trial); and since judicial speech aims at maximal transparency by minimising ambiguity, the OBVC is well suited for testing methods of automatic concept identification. In this presentation we showcase our approach to developing a method to explore intra- and extralinguistic relationships between concepts. For example, do men and women use different concepts in the context of a trial? How do these concepts function in relation to each other? After presenting several case studies, we conclude by outlining paths to further application and development of the proposed method.

¹ **OBV** is derived from two sources: the [Old Bailey Corpus (version 2) (OBC)](http://fedora.clarin-d.uni-saarland.de/oldbailey/) and the [Old Bailey Online (OBO)](http://www.oldbaileyonline.org). It contains data from all **single defendant trials** (21023 defendants) in 227 sessions of the Old Bailey Proceedings between 1780 and 1880 which have had linguistic markup added by the Old Bailey Corpus project.

The dataset has been created in order to explore the [Voices of Authority](https://www.digitalpanopticon.org/?page_id=221) research theme of the [Digital Panopticon](http://www.digitalpanopticon.org) project.

---

Justyna Robinson (University of Sussex), *Identifying and analysing meanings and discourses in 55,000 early English books*

Abstract: This presentation outlines the background, premises, and recent developments of the AHRC-funded project, ‘The Linguistic DNA of Modern Western Thought’. The Linguistic DNA project (LDNA) is an AHRC-funded collaborative project between the universities of Sheffield, Glasgow, and Sussex. The project focuses on designing automatic processes to investigate the emergence and development of concepts in pre-1800 CE print. Employing Early English Books Online, manually-transcribed through the Text Creation Partnership (EEBO-TCP), the project is developing and refining a processing pipeline which assembles groupings of words bound together by their use in discourse. In order to uncover the conceptual history of modern thought we relate the project to traditional work in conceptual and semantic history and define our object of study as the discursive concept, a category of meaning encoded linguistically as a cluster of expressions that co-occur in discourse.

This paper discusses results from a branch of the project which is investigating the lexical semantic relationships encountered in the analysis of discursive concepts. This is done by comparing co-occurrence
data for those semantic categories which show unexpected changes in their size as evidenced by The Historical Thesaurus of English. The outputs of the LDNA processor are here employed to uncover historical dependencies and socio-linguistic relations that are not at all obvious to the 21st century historian or a sociolinguist. In this way, Linguistic DNA tools provide a rather exciting prospect for historical sociolinguistic research that is not constrained by the worldview of modern reader.

Magnus Huber (Justus-Liebig-Universität Gießen), tba

Martin Wynne (Bodleian Libraries, University of Oxford), Forty years of the Oxford Text Archive: reflections on repositories, corpora, and research infrastructure

Abstract: The current deluge of historical data in digital form presents both opportunities and challenges. Five years ago I wrote:

"The emergence of fast and high capacity networks, a deluge of data, and web service APIs mean that it is increasingly possible to imagine and build distributed architectures for scholarly services, where data, tools, computing resources, and the outputs of annotation and analysis live in different parts of the network but can be brought together virtually in the user’s desktop environment."
http://blogs.it.ox.ac.uk/martinw/2012/04/06/silos-or-fishtanks/

This was part of a vision of a research environment where digital technologies allow researchers not only greater ease of access to data and software, but where new types of research become possible. Such a vision was, and remains, key to the mission of the Oxford Text Archive (OTA), and to CLARIN. Reflecting not only on these past five years, but also on experiences over forty years with the OTA, and more than ten years with CLARIN, I will examine how much progress has been made towards the vision of a connected digital ecosystem, considering resources including Eighteenth Century Collections Online (ECCO), Electronic Enlightenment, Cultures of Knowledge, the Newton Project, the Oxford Dictionary of National Biography, and Wikidata.

Susanne Haaf (Berlin-Brandenburgische Akademie der Wissenschaften), Deutsches Textarchiv (German Text Archive): Digitization, standardization and community involvement for a living and growing historical corpus

Abstract: Since 2007 the Deutsches Textarchiv project (German Text Archive, DTA) has been working on creating the basis for a reference corpus of the Historical New High German language of 3 centuries (ca. 1600--1900). Primary objectives were to ensure interoperability by usage of standardized formats and guidelines, high quality benchmarks for transcription and annotation, transparency through extensive documentation, community involvement in various steps of the corpus lifecycle and the assurance of free reuse. The basis for the digitized corpus texts was twofold: About 1600 historical works were digitized
from scratch by applying the double-keying method. Additionally, about the same amount of digitized historical documents were gathered from different sources, e.g. edition projects, individual scholars, or Internet text collections. The latter had to be adjusted to the transcription and encoding standards of the DTA which included some automatic and (depending on the primary quality and format) a rather considerable amount of manual effort. The current presentation provides an overview of the workflows, guidelines and corpus characteristics of the DTA and their respective implications. It reflects on the specifics of historical texts that have to be considered, on standardization and harmonization issues as well as on benefits and challenges of resource-reuse. The important factor of addressing and involving the scholarly community will also be discussed.