SAARLAND UNIVERSITY

IDeaL
SFB 1102

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

7-8 September 2017

List of talks and abstracts

Stefan Fischer, Jörg Knappen & Katrin Menzel (Saarland University), *Metadata production and enrichment for the Royal Society Corpus*

Our research goal is to gain empirically-based insights into the diachronic development of written scientific English from the middle of the 17th century to the present. A corpus of scientific writing from the Philosophical Transactions and Proceedings of the Royal Society of London for the time period of 1665-1869 has been compiled that contains around 30 million running words and approximately 10.000 documents. Considerable effort has gone into making the Royal Society Corpus (RSC) a high-quality resource (e.g. OCR correction, improvement of part-of-speech tagging), and improving the quality of the corpus at all levels is a continuously ongoing process.

We have investigated the hypothesis of increasing linguistic densification, which predicts the emergence of denser, less redundant encodings that optimize efficiency in communication in scientific writing. We found an overall higher information density of later productions compared to earlier ones as well as typical linguistic features involved in change that point to greater encoding density over time (Degaetano-Ortlieb & Teich 2016; Degaetano-Ortlieb et al. 2016).

We want to make use of recently collected metadata and systematically add more metadata in order to determine which register variables in terms of field, tenor and mode of discourse can best explain linguistic variation in our data.

In our talk, we focus on the types of metadata that are already available for the RSC (e.g. Knappen et al. 2017). Various metadata have been collected or generated. For instance, topic modelling has been applied to automatically assign topics (approximation of scientific disciplines) to the documents (Fankhauser et al. 2016). We will discuss which data could be used directly or still needs further processing to address our current research questions and which metadata still have to be created or imported into the dataset.

References:

Degaetano-Ortlieb, S., Kermes, H., Khamis, A. and Teich, E. (2016). An Information-Theoretic Approach to Modeling Diachronic Change in Scientific English. Selected Papers from Varieng – From Data to Evidence (d2e), Helsinki, Finnland . In Suhr, C., Nevalainen, T. and Taavitsainen, I., eds. Selected Papers from Varieng – From Data to Evidence (d2e). Helsinki, Finnland. (to be printed at Brill)

Degaetano-Ortlieb, S. and Teich, E. (2016). Information-based Modeling of Diachronic Linguistic Change: From Typicality to Productivity. In Reiter, N., Alex, B. and Zervanou, K. A., eds. Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH). Berlin, Germany, pages, 165-173. ACL.

Fankhauser, P., Knappen, J. and Teich, E. (2016). Topical Diversification over Time in the Royal Society Corpus Proceedings of DH 2016. Krakow, Poland.

Knappen, J., Fischer S., Kermes, H., Teich, E. (2017). The Making of the Royal Society Corpus. In Bouma, G. and Adesam, Y. eds. Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language, Gothenburg. , Sweden, 133, p. 7-11.

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.

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

Abstract: In 2015, the Royal Society celebrated the 350th anniversary of its first publication, The Philosophical Transactions. To unlock the great archive of science contained in the pages of its numerous publications, the academy of science for the UK and the Commonwealth engaged itself in an ambitious project to digitise its full collection of printed journals from 1665 to 1996, when its publications became digitally born. This paper will introduce this project and the specific ways in which it addressed the challenge of creating a corpus fit for a community of researchers with varied interests. The metadata created will therefore be discussed at length and questions of linked data, identifiers and standards will figure prominently.

Stefania Degaetano-Ortlieb (Saarland University), *Unfolding linguistic variation across social variables and time: a data-driven approach*

Abstract: We present a data-driven approach to study linguistic variation according to social variables (henceforth SV) and their interaction (such as gender and social class). Besides sociolinguistic studies on linguistic variation according to SVs (e.g., Weinreich et al. 1968, Bernstein 1971, Eckert 1989, Milroy and Milroy 1985), recently within the NLP community, computational approaches have gained prominence due to increased (a) data availability of social media and (b) awareness of the importance of linguistic variation according to SVs (see e.g., Eisenstein 2015, Danescu-Niculescu-Mizil et al. 2013, and Nguyen et al. 2017 for an overview).

We aim to investigate possible linguistic profiles of SVs, their interaction and diachronic development. For this, we use the Old Bailey Corpus (OBC; Huber et al. 2012), a diachronic corpus of the Proceedings of the Old Bailey Court (from 1720-1913) annotated with linguistic (e.g., tokens, parts of speech) and social information (e.g., age, gender, social class of the speaker based on the HISCO standard).

For the detection of sociolinguistic patterns of variation at different linguistic levels (e.g., lexical, syntactic), we use Kullback-Leibler Divergence (KLD; Fankhauser et al. 2014), allowing us to capture differences between two probability distributions based on linguistic units (e.g. words, parts of speech) and inspect features contributing to these differences. Thus, we are able to combine macro- and micro-analytic perspectives as we obtain knowledge on (1) the amount of difference of two distributions (e.g. male vs. female) in bits and (2) which features are typical of the one or the other distribution (i.e. typical for male or female).

Rather than considering SVs in isolation, we consider variation based on interactions between SVs. Considering, for example, differences between male and female based on social class, the biggest difference is found between male of higher vs. female of lower class (0.23 bits by KLD). This difference is stable over time. Differences of social class for the same gender show greater difference between female than male of higher vs. lower class (female: 0.112 bits, male: 0.08 bits). Inspecting features contributing to these differences, e.g., female of lower class are distinguished by self-reference (e.g. I) and lexis covering roles (e.g. master, servant, mistress) and locations (e.g. kitchen, home, stairs), while female of higher class use prepositions (e.g., of, for, in), conjunctions (e.g., and, but) and determiners (e.g., a, this), reflecting a more elaborate style.

In the talk, we present our methodology as well as selected analyses of linguistic variation of social variables across time.

References:

Bernstein B. 1971, Class, Code and Control: Volume 1 Theoretical Studies towards a Sociology of Language, Routledge Taylor & Francis Group, London and New York.

Danescu-Niculescu-Mizil C., Sudhof M., Jurafsky D., Leskovec J. and Potts C. 2013, A Computational Approach to Politeness with Application to Social Factors, in Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, volume 1, Sofia, Bulgaria, pp. 250-259.

Eckert P. 1989, Social Categories and Identity in the High School, Teachers College Press, New York.

Eisenstein J. 2015, Written Dialect Variation in Online Social Media, in Boberg C., Nerbonne J. and Watt D. (eds.), Handbook of Dialectology, Wiley.

Fankhauser P., Knappen J. and Teich E. 2014, Exploring and Visualizing Variation in Language Resources, in Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, pp. 4125-4128.

Huber M., Nissel M., Maiwald P. and Widlitzki B. 2012, The Old Bailey Corpus. Spoken English in the 18th and 19th centuries. www.uni-giessen.de/oldbaileycorpus

Milroy J. and Milroy L. 1985, Linguistic Change, Social Network and Speaker Innovation, in “Journal of Linguistics” 21 [2], pp. 339-384.

Nguyen D., Dogruöz A.S., Rosé C.P. and de Jong F. 2016, Computational Sociolinguistics: A Survey, in “Computational Linguistics” 42 [3], pp. 537-593.

Weinreich U., Labov W. and Herzog M.I. 1968, Empirical foundations for a theory of language change, in Lehmann W.P. and Malkiel Y. (eds.), Directions for Historical Linguistics: A Symposium, pp. 95-188, University of Texas Press, Austin.

Julie Weeds[*], Justyna Robinson[*] & Fraser Dallachy[#] ([*]University of Sussex, [#] University of Glasgow**)**,

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), *Sociolinguistic annotation in the Old Bailey Corpus*

Abstract: 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.