ObamaSpeeches.com – Approaches to Building and Processing a Corpus of Political Speeches

Sabine Bartsch, Stefania Degaetano, Tomek Grubba, Nina Petrychka, David Sullivan,
Christoph Tragl, Claudio Weck;

Technische Universität Darmstadt, Institut für Sprach- und Literaturwissenschaft, Hochschulstrasse 1, 64289 Darmstadt; corresponding author: bartsch@linglit.tu-darmstadt.de

Linguistic research questions that require other than available standard corpora confront linguists with a number of issues concerning corpus collection, encoding, annotation and query. This paper reports on a student project aiming at a discourse and register analysis of a corpus of the speeches of US President Barack Obama that entails the four steps just mentioned. The focus of the paper will be on issues of corpus annotation and analysis by means of heterogeneous tools. The principal aim of the project was to learn more about the characteristics of a popular and influential set of political speeches in terms of established register features (Biber 1988, 1995; Conrad & Biber 2001) as well as their discourse structure in terms of topic distribution within speeches, use of cohesive devices (Halliday & Hasan 1976), coherence in terms of rhetorical structure modeled on the basis of the Rhetorical Structure Theory (RST) (Mann & Thompson 1987) and thematic development. To this end, a corpus of 120 speeches by Barack Obama given between 2002 and 2009 was collected from the URL http://www.obamaspeeches.com as a set of html files. The goal was to enrich the corpus with annotations that would allow the qualitative and quantitative analysis of the features just mentioned. A central methodological and didactic aim of the project was to evaluate the interaction between different tools in terms of their usability in a complex linguistic analysis workflow.

A large number of tools are available for corpus processing and annotation at every level of linguistic organization. In the course of their training as corpus linguists, students learn to use different tools for specific research tasks, e.g. the use of a concordancer, different part-of-speech taggers and parsers as well as manual annotation tools. An issue facing more advanced students as well as experienced linguists is the interplay between those tools and their usability in a project workflow. The available tools are of principally two types: there are single-purpose tools that serve one specific annotation purpose and complex tools that serve multiple annotation purposes and are usually configurable in terms of different underlying annotation schemes and applicability to annotations at different levels of linguistic organization. Among the first set of tools are tools for automatic annotation such as part-of-speech taggers and syntactic parsers, but also manual annotation tools for example for the annotation of higher level discourse structure such as rhetorical structure (e.g. RSTTool). The latter set commonly comprises tools that are more versatile in their application to different levels of linguistic organization and typically allow for the development of bespoke annotation schemes. All of these tools require their input data to be in a specific format and produce a specific kind of output format; these formats are nt standardized and differ between different tools. In terms of their interaction with one another these tools fall into principally three types: individual stand-alone tools for manual and automatic annotation, frameworks for the integrated execution of sets of tasks, and processing systems or pipelines. Examples for stand-alone tools are part-of-speech taggers (e.g. the Stanford NLP POS tagger, or Helmut Schmid's widely used Decision Tree Tagger) and manual annotation tools (e.g. Mick O'Donnell's UAM Corpus Tool, MMAX2). As an example of a framework, we use GATE. We furthermore employ Steven Bird et al.'s Natural Language Toolkit (NLTK) as an example of a processing system based on a programming language, Python, which has advanced language processing capabilities. NLTK was selected because it promised to facilitate the integration of various annotation and analysis steps.

Using configurations of these tools, the corpus is encoded in XML according to the standards set out by TEI P5; input data for all tools is derived from this basis. The annotations comprise tokenization, part-of-speech tagging, semantic clusters and named-entity recognition as well as discourse structure in terms of cohesive ties, rhetorical structure and thematic development. These annotations are the basis for quantitative and qualitative analyses of a variety of register and discourse features. The analysis entails profiling each individual text as well as the entire corpus in terms of basic quantitative characteristics and feature distributions (e.g. type / token, features counts and distributions, lexical clusters and chains as well as word clouds for content characterization etc.). The corpus is furthermore characterized in terms of a set of register features according to Biber in order to characterize the speeches in terms of their speaker – recipient distance, informational character and spoken vs. written language characteristics. Various features of discourse structure and organization are explored in order to arrive at a description of the topic development, cohesive ties and text coherence. By means of these analyses, the project seeks to arrive at a conclusive characterization of recurrent patterns in the speeches of Barack Obama that can also be extended to comparative studies of other political speeches. The overall aim of the project is an exploration of the usability and interaction between different tools in a complex linguistic analysis workflow as would be found useful in discourse and register analysis.

References:

Biber, Douglas. 1988. *Variation Across Speech and Writing.* Cambridge University Press.

Biber, Douglas. 1995. Dimensions of register variation: A Cross-Linguistic Comparison. Cambridge University Press.

Bird, Steven & Edward Loper. 2004. NLTK: The Natural Language Toolkit. In: Proceedings of the ACL demonstration session. Barcelona, Association for Computational Linguistics, July 2004. pp 214-217.

Conrad, Susan & Douglas Biber. 2001. *Variation in English: Multi-Dimensional Studies.* Longman.

Dipper, Stefanie, Michael Götze & Manfred Stede. 2004. 'Simple Annotation Tools for Complex Annotation Tasks: an Evaluation',in: Proceedings of the LREC Workshop on XML-based Richly Annotated Corpora, Lisbon, Portugal. pp. 54-62.

Halliday, MAK & Ruqaiya Hasan. 1976. *Cohesion in English.* Longman.

Loper, Edward & Steven Bird. 2002. NLTK: The Natural Language Toolkit. In: Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia, Association for Computational Linguistics. July 2002. pp 62-69.

Mann, William C. & Sandra A. Thompson. 1987. ‚Rhetorical structure theory of text organisation', in: *The Structure of Discourse.* Ablex Publishing Corporation.

ObamaSpeeches.com URL: http://www.ObamaSpeeches.com

TEI P5 URL http://www.tei-c.org/Guidelines/P5/