ObamaSpeeches.com: Building and Processing a Corpus of Political Speeches A student project

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

Institut für Sprach- und Literaturwissenschaft, Hochschulstrasse 1, 64289 Darmstadt, URL: http://www.linglit.tu-darmstadt.de

TECHNISCHE UNIVERSITÄT DARMSTADT

1. Introduction

This poster presents a student project aiming at integrating annotation tools for a discourse analysis of a corpus of speeches by US President Barack Obama. The project entails corpus collection, encoding, annotation and query.

The linguistic aim of the project is to learn more about the characteristics of a set of political speeches in terms of established register features (Biber 1988, 1995) as well as their discourse structure in terms of topic development within speeches, use of cohesive devices (Halliday & Hasan 1976), rhetorical structure modeled on the basis of Rhetorical Structure Theory (RST) (Mann & Thompson 1987) and thematic development (Matthiessen 1995, Halliday 2004).

Issues are the interplay between different tools in light of heterogeneous data formats, and the integration of automatic annotation procedures as pre-processing steps for manual annotation tasks.

Aim: development of a processing chain that allows the linguist to explore the relevant properties of the corpus at different levels of linguistic organization.

Approach: integration of automatic and manual annotation tasks by means of NLTK.

2. The ObamaSpeeches Corpus (OSC)

120 speeches by US-President Barack Obama
Time span: 2002-2009.
Source: www.ObamaSpeeches.com
Source format: html

Derived formats for linguistic processing:

- plaintext
- html
- XML (TEI P5)
- GATE data store

3. Methods: Multi-level corpus annotation

Annotation requirements:

- Corpus metadata
- Tokenization
- Part of speech tagging
- Cohesive chains
- Rhetorical structure
- Thematic structure

Data format: multi-layer standoff

Tools explored:

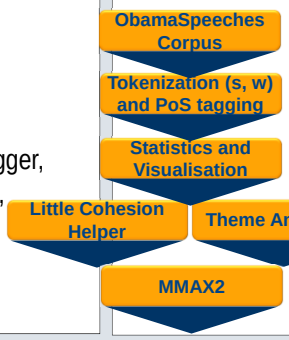
- Stand alone tools (Decision Tree Tagger, Theme Annotator, UAM Corpus Tool, MMAX2, etc.)
- GATE
- Natural Language Toolkit (NLTK)

3. Multi-level corpus annotation (ctd.)

Tool	Feature evaluation
Stand alone tools	+ powerful, reliable - tool integration - heterogeneous data
GATE	+ well-integrated - usability (new tools) - stability + homogeneous data
NLTK	+ powerful, flexible + tool integration + usability + homogeneity possible

4. Adopted approach

The Natural Language Toolkit (NLTK) is used as a basis for an implementation of automatic annotation steps whose output is prepared for further manual processing with MMAX2.



NLTK enables the integration of many standard annotation tools (e.g. the Punct-Tokenizer, Unigram tagger) as well as an api to resources such as WordNet.

4. Automatic support for manual annotation: The Little Cohesion Helper

As an example module developed with NLTK, the Little Cohesion Helper is presented here. Based on the NLTK / Python interface to WordNet, the Little Cohesion Helper (LCH) (Weck, Tragl 2009), this tool was developed to automatically identify and annotate cohesive ties in free text and prepare the output for further manual processing.

MMAX2 is the tools of choice for the annotation of cohesion, a task that has previously be shown to be amenable to automatic support on the basis of resources such as WordNet (Teich, Fankhauser 2006).

LCH integrates all pre-processing steps such as tokenization, pos-tagging with cohesion annotation .

LCH produces as its output an MMAX2 project that allows further manual processing (see Figure x.x). It produces statistics on different types of cohesive relations, distance of relations and chain length (see Fig. x.x).

4. The Little Cohesion Helper (ctd.)

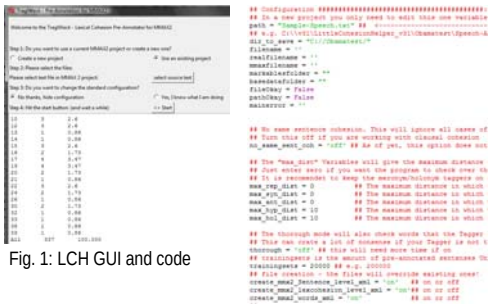


Fig. 1: LCH GUI and code

LCH can be used on the command line or Python's IDLE or through a GUI.

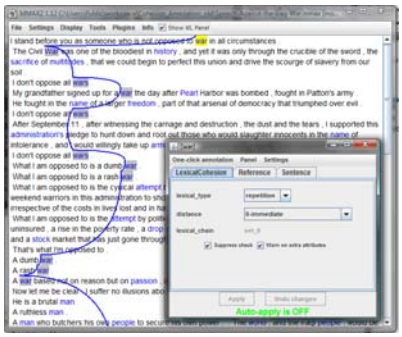


Fig. 2: MMAX project by LCH

The user can select all types of cohesive ties described in Halliday & Hasan (1976) for identification.

5. Additional features and future work

NLTK is also used for basic text statistics and visualizations thereof e.g. as a wordcloud.

Thematic structure annotated automatically by means of the Theme Annotator (Schwarz et al. 2008) can also be integrated into MMAX2 projects.

Query of the data currently proceeds by the MMAX2 query & statistics facilities. In the future, ANNIS2 will be employed to hold the data and allow for more advanced query.

References

AnnoLab,: http://www.annolab.org
ObamaSpeeches.com URL:http://www.obamaspeeches.com
MMAX2 URL: http://mmax2.sourceforge.net/
NLTK. URL: http://www.nltk.org
Steven Bird, Ewan Klein, and Edward Loper. 2009. Natural Language Processing with Python. Analyzing Text with the Natural Language Toolkit. O'Reilly.
Halliday, MAK, Ruquayia Hasan. 1976. Cohesion in English. Harlow: Longman.
PAULA Interchange Format for Linguistic Annotations, URL: http://www.sfb632.uni-potsdam.de/~d1/paula/doc/