Project No.: CIT4-CT-2006-028698

RECON
Reconstituting Democracy in Europe

Integrated Project
Priority: 7 - Citizens and Governance in a Knowledge-Based Society

Deliverable No. 77
Workshop on computational-linguistic text-analytical methods in the social sciences

Due date of deliverable: June 2010
Actual delivery date: 27-28 May 2010

Start date of project: 1 January 2007 Duration: 60 months

Lead contractor for this deliverable:
Partner 5 FUB
Freie Universität Berlin, Germany

Final Version

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)		
Dissemination Level		
PU	Public	
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	X
CO	Confidential, only for members of the consortium (including the Commission Services)	

1. Introduction

RECON Deliverable No. 77 – *Workshop on computational-linguistic text-analytical methods in the social sciences* is part of work package 6, which focuses on the foreign and security policy dimensions of the European Union. It seeks to assess the status and the prospect for democracy within the field of foreign and security policy in relation to the three RECON models. During the months 37 to 54 of the project, WP 6 focuses on assessing and analysing empirical findings, while to some extent continuing to collect data. Empirical data will be assessed with an eye to answer the following two questions: to what extent has a move beyond intergovernmentalism taken place, and what kinds of competences and powers (if any) have been uploaded to the EU level? One part of the research agenda of WP 6 investigates the EU's security and defence policies in media debates on war and military intervention.

2. Presentation of the workshop

With the availability of digital text archives and computer-aided tools for textual analysis, social scientists are able to boost empirical investigation on political communication. The project 'The EU's security and defence policies in media debates on war and military intervention', which is part of RECON and directed by Thomas Risse and Cathleen Kantner at Free University Berlin (FUB), used these new possibilities to extens research across a time-span of 16 years, seven country cases, and four languages (cf. RECON Online Working Paper 2008/19). However, computational advances raise a number of methodological challenges, such as how to assure a balanced and relevant corpus of texts, how to make use of automatic information retrieval without losing out on validity, and how to mix quantitative and qualitative methods.

The workshop 'Computer-aided methods of textual analysis' was organised on 27-28 May and focused on the methodological benefits and challenges brought about by new computational advances. The objective of the workshop was to bring together researchers from computational linguistics and political science, who have dealt with these issues systematically in recent years. The workshop was organised by Cathleen Kantner and Amelie Kutter and financed by RECON and the Jean Monnet Centre of Excellence at FUB. It was designed as an inter-disciplinary forum, in which common methodological interests of computational and corpus linguists working on computer-aided linguistic analysis, on the one hand, and social scientists, using computational and corpus-linguistic tools to advance text analysis, on the other, were discussed. The aim was to strengthen already existing research collaboration in the field while also creating new paths for trans-disciplinary research. The workshop gathered 16 researchers from ten institutions.

Political scientists presented different applications of semi-automated content analysis to issues of political communication, which had been realised in collaboration with computational linguists. Bruno Wüest, Simon Clematide, and Daniel Laupper (University of Zürich) showed how core sentence analysis, which is suited for the exploration of *party positioning*, can be supported by automatically identifying relations between actors and issues. Jacek Kołodziej and Michał Buchowski (Jagiellonian University Krakow) discussed possibilities and limits of exploring *cultural symbols* such as 'democracy' in publicly available collections of texts, using tools like WordStat and Poliqarp. Kutter and Kantner revealed how inductive semantic field analysis can be used for internally valid automated

content analysis that reveals *issue salience* of 'military intervention' and the *distribution of frames* such as 'foreign policy actorness' in press coverage.

A second cluster of presentations dealt with the automatic retrieval of genre- and language-specific information from texts that give a clue about the social context and in which these texts were generated. Fabienne Fritzinger and Ulrich Heid (University of Stuttgart) presented a paper on the extraction of morphological and syntactical information, which allows distinguishing *domain-specific terminology* such as that of intellectual property rights from general language use. Manfred Stede (Potsdam University) introduced the theory and methods guiding the automatic identification of structures of *pro/contra commentaries*. Drawing on the example of a century-spanning corpus of Latin texts, Alexander Mehler, Ulli Waltinger, and Rüdiger Gleim (Universities of Bielefeld and Frankfurt) showed how *change in language use* could be detected by inducing linguistic networks at the level of words, lemmas, and sentences.

A third group of presentations dealt with the performance of particular tools for language-sensitive exploration of social science questions. Manuel Burghardt and Christian Wolff (University Regensburg) presented their usability assessment of web-based tools designed for the annotation of specific text corpora. Brit Helle Aarskog (University of Bergen) introduced the workshop participants to *textUrgy* as a platform for computer-aided inductive analysis. Ulli Waltinger (University of Bielefeld) and Rüdiger Gleim (Frankfurt University) explicated uses of *eHumanities Desktop*, such as the topic-categorisation of documents. Finally, Peter Kolb (Potsdam University) showed how *DISCO* helps building semantic clusters from a specific corpus, e.g. the lexical field of 'missile'.

Across these thematic clusters ran issues of methodological discussion that had been touched upon in Thomas Risse's introductory remarks. The first related to the performance of automatic procedures concerning both standards of internal validity and relevance for social science research questions. The general conclusion was that the mining of dependency relations between words, rather than only of their morphological characteristics, provides important information to the social scientist, which can be used for sorting and analysing the text material according to particular content-related questions. The second strand expounded problems of disciplinary terminology and research focus that need to be bridged in truly interdisciplinary effort. Finally, the necessity and difficulty of contextual interpretation of automatically generated information of texts was discussed.

The contributions from the workshop will be published as an edited collection of papers in the working paper series of the Jean Monnet Centre of Excellence at FUB, at http://www.jmc-berlin.org/new/index.php.

3. Participants

Bukowski, Michal	Jagiellonian University Krakow
Burghardt, Manuel	University of Regensburg
Clematide, Simon	University of Zürich
Fritzinger, Fabienne	University of Stuttgart
Gleim, Rüdiger	Goethe University Frankfurt
Helle Aarskog, Brit	University of Bergen
Kantner, Cathleen	Free University Berlin

Kolb, Peter	University of California, San Francisco
Kolodziej, Jacek	Jagiellonian University Krakow
Kutter, Amelie	Free University Berlin
Laupper, Daniel	University of Zürich
Mehler, Alexander	University of Bielefeld
Risse, Thomas	Free University Berlin
Stede, Manfred	University of Potsdam
Wolff, Christian	University of Regensburg
Wüest, Bruno	University of Zürich

4. Programme

See attachment.

Computer-aided methods of textual analysis

JMCE / RECON Workshop
WP 6 – The Foreign and Security Dimension

Berlin, May 27-28 2010.

The joint workshop of the Jean Monnet Centre of Excellence (JMCE) and RECON is organised by the Otto-Suhr-Institute of Political Science, Free University Berlin, under work package 6 of the RECON project. WP 6 – The Foreign and Sequrity Dimension – aims to assess the status as well as the prospect for democracy within the field of foreign and security policy in Europe.

In the digital era, large bodies of electronic data hold great promise for the exploration of researchers' questions. Social scientists have started to use computer-aided and linguistic methods for the investigation of social and political issues, drawing on large text corpora. In the process, they have invariably been confronted with a host of problems related to the compilation of an issue-specific corpus; the retrieval of content-related information and its frequency-based generalisation; the down-sizing of corpora; the detailed qualitative-linguistic analysis of smaller sub-corpora; and the triangulation of text analytical methods for the qualitative and quantitative exploration of texts. At the same time, computational and corpus linguists have developed and applied new tools for the computational retrieval, annotation, and representation of linguistic information in large text corpora. However, they are faced with the challenge of communicating these innovations to a broader scientific audience that has an increasing interest in these tools of analysis, but often operates with different (contentrelated) categories.

The objective of this workshop is to explore the common methodological interests of computational and corpus linguists working on computer-aided linguistic analysis, on the one hand, and social scientists, who have used computational and corpus-linguistic tools to advance text analysis in the social sciences, on the other. We hope to foster interdisciplinary cooperation focusing on method-related issues:

- Innovations in computational and corpus linguistics: How to retrieve, annotate, and represent linguistic information and how to bring these insights to fruition across disciplines. Typical wording as a key to meaning. Innovative approaches to tracing complex structures of meaning.
- Method-mix: How to combine multiple computer-aided text analytical methods on the threshold between quantitative / quantifying and qualitative analysis and between different text analytical approaches (content analysis, corpus analysis, discourse analysis).

The workshop is designed as an 'authors' workshop.' It provides a forum for the discussion of results gained in ongoing research projects with the aim of compiling these results in a journal special issue on interdisciplinary methods. The workshop is meant to strengthen existing research collaboration in this field and to create new avenues for trans-disciplinary research.

RE/CON

Programme

Thursday, 27 May 2010:

How to make use of computer- and corpus-linguistic methods for large-n text analysis in the social sciences

12:30-13:00 **Arrival, registration and coffee**

13:00-13:15 **Welcome and opening**
Thomas Risse
Free University Berlin

13:15-13:30 **Introduction: How to make use of computer- and corpus-linguistic methods for large-n text analysis in the social sciences**
Amelie Kutter and Cathleen Kantner
Free University Berlin

The preparation of large multi-lingual corpora

13:30-14:00 **Computing in the humanities by means of the eHumanities desktop**
Alexander Mehler and Rüdiger Gleim
University of Bielefeld and Goethe University Frankfurt

14:00-14:15 **Discussion**

14:15-14:45 **How to get rid of the noise in the corpus: cleansing large samples of digital newspaper texts**
Amelie Kutter and Cathleen Kantner
Free University Berlin

14:45-15:00 **Discussion**

15:00-15:30 *Coffee break*

Text analysis between human coding/annotation and automatisation

15:30-16:00 **Toward identifying arguments automatically**
Manfred Stede
University of Postdam

16:00-16:15 **Discussion**

16:15-16:45 **Improving manual annotation with usable software tools**
Manuel Burghardt and Christian Wolff
University of Regensburg

16:45-17:00 **Discussion**

17:00-17:30 **Electoral campaigns, relation mining, and dependency parsing: Extracting semantic network data from Swiss newspaper articles**
Bruno Wüest, Simon Clematide and Daniel Laupper
University of Zürich

17:30-17:45 **Discussion**

17:45-18:00 **Final debate and conclusions**

19:30 **Dinner,** *Jules Verne Restaurant*
Schlüterstraße 61, S-Bhf Savignyplatz

Friday, 28 May 2010

How to analyse millions of texts and still be home for tea? Innovations in computer-aided linguistics and the social sciences

10:00-10:15 *Coffee*

Typical wording as a key to meaning

10:15-10:45 **Avoiding the hammer-and-nails problem. A theory driven approach to computer-aided political keywords analysis**
Michal Bukowski
Jagiellonian University Krakow

10:45-11:15 **Automatic meaning? Large text corpora as the source of social knowledge**
Jacek Kolodziej
Jagiellonian University Krakow

11:15-11:30 **Discussion**

11:30-12:00 **Corpus-based extraction of domain-specific terminology**
Fabienne Fritzinger
University of Stuttgart

12:00-12:15 **Discussion**

12:15-14:00 **Lunch:** *Harnack-Haus*
Ihnestraße 16-20, U-Bhf Thielplatz

Typical wording as a key to meaning – continued

14:00-14:30 **Semantic fields and conceptual categories: exploring the synergy of corpus and content analysis**
Cathleen Kantner and Amelie Kutter
Free University Berlin

14:30-14:45 **Discussion**

14:45-15:00 *Coffee*

Approaches to tracing complex structures of meaning

15:00-15:30 **A text-mining system for content analysis of large text corpora**
Peter Kolb
University of California, San Francisco

15:30-15:45 **Discussion**

15:45-16:15 **Speech act theory and grammar-based text analysis**
Brit Helle Aarskog
University of Bergen

16:15-16:30 **Discussion**

16:30-16:45 *Coffee*

Open discussion on problem-solutions and recommendations and concluding remarks

16:45-17:30 **Presenters and other experts:**
Brit Helle Aarskog, *University of Bergen*
Michal Bukowski, *Jagiellonian University Krakow*
Manuel Burghardt, *University of Regensburg*

Fabienne Fritzinger, *University of Stuttgart*
Rüdiger Gleim, *Goethe University Frankfurt*
Cathleen Kantner, *Free University Berlin*
Peter Kolb, *University of California, San Francisco*
Jacek Kolodziej, *Jagiellonian University Krakow*
Amelie Kutter, *Free University Berlin*
Alexander Mehler, *University of Bielefeld*
Thomas Risse, *Free University Berlin*
Manfred Stede, *University of Postdam*
Christian Wolff, *University of Regensburg*

19:30 **Dinner:** *Anda Lucia, Restaurand & Tapasbar*
Savignyplatz 2, S-Bhf Savignyplatz