From sentiments and opinions in texts to positions of political parties

Research project proposal submitted to CAMeRA involving three PhD positions

Contact persons:

Prof. dr. Piek Vossen	Dr. Alan Cienki	Dr. André Krouwel
Dept. of Language & Communication	Dept. of Language & Communication	Dept. of Political Science
Faculty of Arts, VU	Faculty of Arts, VU	Faculty of Social Sciences, VU
p.vossen@let.vu.nl	a.cienki@let.vu.nl	apm.krouwel@fsw.vu.nl

Table of Contents

0 Introduction

The proposed project combines contemporary theories and methods in linguistics and political science to develop an automated research tool for rich text-mining. The transdisciplinary relevance of the project is that a carefully constructed mining tool for language-meaning research can be applied to enhance the *Kieskompas* (Electoral Compass) and prove useful in the social sciences in general. The research will give new insights into the complexity of language use, the linguistic modeling of subjectivity and the representation of this knowledge in a lexicon. It will also shed new light on the complex dimensionality of competition between political parties.

1 Problems addressed

Two main areas of research are thus at the centre of this study: contemporary work in political science and a combination of approaches in linguistics. The following provides background context for each of them.

Background for the project in the political domain of contemporary society

In advanced industrial democracies, such as those found in Western Europe, processes of social mobility and emancipation have left citizens with weaker group loyalties and fewer institutional links to political parties. This has resulted in less traditionally ‘partisan’ voting behaviour (Dalton & Wattenberg 2002). New voter groups (such as younger generations and immigrants) enter the voting population with even less socialisation in party politics (Kitschelt 1997; Dalton 2002). The core vote for traditional parties is declining as voters are more likely to change their party preference from election to election (Dalton & Gray 2003; Dalton 2004; Franklin 1992, 2003), while at the same time the overall number of political parties participating in national elections continues to increase (Webb 2002). With the disappearance of profound ideological differences between the major political parties, voters face an increasingly difficult task of determining what the policy positions of parties are (Krouwel, forthcoming). This is problematic as research also shows that issues and policy considerations have become more important for voters (Franklin 2003; Thomassen et al. 2000).

In order to find out more about the policy positions of parties, voters facing elections in the Netherlands, Belgium, Switzerland, Finland, Germany and other democracies have increasingly made use of party profiling systems since the first paper versions were introduced in the late 1980s (Hooghe & Teepe 2007). With the development of computerised versions, voters can fill out an online questionnaire and subsequently see where they fall in relation to the parties along several dimensions. The accessibility and popularity of party profiling websites has made them very influential in terms of determining voting behaviour as well as affecting media reporting on elections (Boogers 2006; Walgrave & van Aelst 2005). However, the quality of party profilers varies substantially and – much worse – some methods allow political parties to manipulate the tool (van Praag 2007). In order to better inform citizens of the political position of candidates and parties, to improve the quality of media reporting on elections and to curb manipulation by major political actors, a reliable method is needed for party profiling and for determining relative positions of political parties on issues according to salient political dimensions.

The input for such party profiling comes from large scale analyses of various sorts of texts, including media reporting on the political parties, manifestoes and policy statements by the parties published on their web sites, and transcripts of parliamentary debates. The language used is very rich and subtle with respect to how it expresses attitudes, sentiments and opinions towards various political issues. This ranges from lexical choice of positively or negatively loaded words, to degrees of negation, irony, subtle perspective changes or causal associations. A careful analysis of the linguistic encoding of this information is required in order to build automated systems that can learn the positions of political parties from the texts they produce. Since politicians use language in sophisticated ways to manipulate voters, the analysis of political discourse thus represents a major challenge for the design of computational systems to derive opinions from texts (i.e., perform *automated opinion mining*). There is a need for detailed knowledge and more theory on the genres of political texts to define better algorithms for computational analysis. In addition, this knowledge needs to be modelled and stored in some knowledge base or lexicon so that it can be used in automated text analysis. Little is known about the encoding of this knowledge in the lexicon, or about its relation to the semantics and possibilities of combining words and expressions, etc. Modeling this knowledge and acquiring it on a large scale is therefore another problem to be addressed. Finally, if the linguistic ‘machinery’ can adapted for use by computer programs, large databases of relevant texts (*linguistic corpora*) can be analysed automatically to reveal the positions of political parties on different issues. This will require research into the proper balance of statistical and knowledge-rich techniques of text processing, to be evaluated against the objective of party profiling.

Linguistic and lexical analysis of subjectivity, and automated text analysis tools, are active areas of research, both in the field of linguistics and in the area of natural-language-processing. There is an emerging potential for linguistic engineering that could be applied to the political domain as well, as described above. This project aims at bringing these together and to provide a next step in both linguistic modeling and practical exploitation for opinion mining. There are two areas of linguistic research that we want to explore: lexical knowledge modeling and acquisition of fine-grained subjectivity entailments and cognitive/critical discourse analysis of language usage in political text corpora. Both of these areas are discussed in more detail below.

Background for the project in contemporary linguistic research

- **Language technology/ computational lexicology**

Machine learning and statistical techniques have been used for many information and language technology systems in recent years. Also in the area of sentiment and opinion analysis, there are many publications and even commercial systems that use these techniques. Although most of these systems are fast and robust, they have a serious drawback. The level of analysis is relatively shallow and they cannot handle subtle linguistic expressions and structures.

The analysis of attitudes and opinions in political texts has been a core element in content analysis studies within the social sciences for decades, due to the seminal work of political scientists such as Harold Lasswell (Namenwirth & Lasswell 1969) and communication scientists such as Charles Osgood (Osgood, Saporta & Nunnally 1956). Issue positions of parties can be viewed as attitudes of parties towards specific policies or issues. Content analysis aimed at the extraction of the positive, neutral or negative position of specific parties with regard to specific issues belongs to the field of "relational content analysis" (Roberts 1997; Popping 2000), or "semantic network analysis" (Krippendorff 2004). Most automated content analysis studies are based on word co-occurrences, but in recent years attempts have been made to automate semantic network analysis (Van Atteveldt 2008), amongst others by using extensive thesauri and automated grammar parsing (van der Beek, Bouma & Van Noord 2002). (van der Beek, Bouma, & van Noord 2002). Performance indicators (e.g. F1=0.55 [Van Atteveldt et.al. 2008]) for the extraction of issue positions of parties reveal that good progress has been made, but also that much can be improved.

The expression of attitudes and opinions in political text is not trivial. First of all, the association of a value towards themes and properties is a matter of degree rather than an either/or decision (scalar rather than polar). This information should be reflected in the resources used for assigning attitudes and sentiments to text fragments. Secondly, texts are often very rich in terms of modal and epistemic expressions (e.g., verbs like *may*, *could*, *should* and terms concerning perception, belief, knowledge, etc). These can be used to derive a better model for the representation of explicit and implicit attitudes of speakers in text. Beliefs and situations can be confirmed, desired, wished for or negated, denied or rejected in many subtle ways. Furthermore, language can be ironic, sarcastic or involve metaphors as stylistic means to express subtle opinions that are extremely difficult for computer programs to detect. All these aspects have been studied for years in the field of linguistics and pragmatics (for example Aikhenvald 2004; Auwera & Schalley 2004; Billig 1995; van Dijk 1977; Palmer 1986; Rodman 2001), but this knowledge is hardly used in language-technology solutions.

Languages are very rich in terms of lexical choices and expressions for these purposes. Nevertheless, knowledge concerning sentiment and attitudes, epistemic and modal properties in lexicons are usually sparsely encoded and poorly modelled. For example, Esuli and Sebastiani (2005, 2007) encode this information at the level of a *synset* in SentiWordnet. Synsets are sets of synonyms that represent a single concept in Wordnet (Fellbaum 1998). However, synonyms typically tend to vary with respect to pragmatic and attitudinal loading. For example, “maintain” and “defend” are synonyms in Wordnet while they express different degrees of subjective values.

Maks et al. (2008) try to combine existing semantic frameworks as a first descriptive apparatus for adjectives that can serve as a basis for a lexical model for pragmatic and attitudinal entailments. A more complete lexical model would consist of complex and typed knowledge structures that represent generalisations of classes of words and expressions. The classes stand for the structural combinatoric and semantic behaviour with a focus to subjective entailments.

Linguistic knowledge of words and expressions in a lexicon indicates their possible usage in text, i.e., it is an abstraction of their actual usage. Still the actual usage in text is a matter of choice and deliberate strategy. Whether a word ***can*** be used transitively both in passive and active form is a lexical feature that needs to be encoded in the lexicon. However, the observation that certain participants use that word significantly more in passive constructions is a feature that relates it to language use in a specific domain by specific speakers, e.g., in the political domain. Such usage can signal specific positions of the speaker that do not follow from the abstract knowledge of the word. The study of language use in a domain for a communicative strategy is thus a complementary line of research that we will discuss in the next section.

- **Cognitive Linguistics/Critical Discourse Analysis/Corpus Linguistics**

Another approach toward working with large text databases, the field of corpus linguistics, involves the analysis of large collections of machine-readable texts (corpora) in order to reveal phenomena such as widespread patterns of language use. These patterns may be researched for their semantic aspects or other properties.

Other areas of linguistics bring new tools for analysing meaning in language, particularly as expressed in usage, and some of these are now being operationalised to be used with large bodies of texts. Two of them are cognitive linguistics and critical discourse analysis. The field of cognitive linguistics, in which psychologically plausible theories of language structure and use are a priority, bring several decades of research on a number of topics, including the following.

- Cognitive/socio-cultural models as shared structures we implicitly use in reasoning, argumentation and decision-making (Cienki 2007; D’Andrade & Strauss 1992);
- Metaphor as an everyday means of understanding and talking about abstract concepts (Lakoff 1993; Lakoff & Johnson 1980, 1999). Metaphor often plays a key role in the structuring of cognitive/cultural models (Cienki 1999), such as when we talk, think and

make inferences about A NATION *as if it were* A FAMILY.

- Metonomy as a way in which we make reference to things or ideas, often without mentioning them directly (Barcelona 2000; Panther & Radden 1999), as when reporters refer to what *Washington said* without specifying which individual in the US government is the actual source.

All of these can be used to frame a given issue in one way or another, with any given framing serving to highlight and also to hide aspects of the issue according to a certain perspective. Another way of analysing texts is known as critical discourse analysis (CDA). It draws on poststructuralist and critical theory to provide an interpretive means of analysing power structures in social institutions as revealed through the use of language in them and about them. Thus rather than constituting a research method in and of itself, CDA is a perspective toward discourse research aimed toward revealing the underlying dynamics of power positions and relations (see, for example, Fairclough 1995, 2000; Van Dijk 1997; Wodak & Chilton 2005; Wodak & Krzyzanowski 2008). Research in cognitive linguistics (in particular metaphor analysis) and research in CDA have recently begun to be integrated not only with each other (e.g., Hart & Lukeš 2007), but also with computational research techniques (e.g., Baker 2006; Charteris-Black 2004; Deignan 2005).

In light of the above, we now have the potential to:

- solve the text analysis problem using a combination of the computational and interpretive techniques, outlined above, in a grouping of three interrelated PhD projects; and
- compare the results of using these methods with the findings of some of the textual data which was hand-coded by previous researchers; this will allow for fine tuning of our research methods accordingly.

The results of this research will provide a model for the analysis of new texts which have not been coded before for the positions of the political actors on salient political issues and dimensions.

A concomitant benefit of this research is that it will develop a potential method which could be used by future researchers in the social sciences. In addition, while the projects will be based on the analysis of Dutch-language texts, there will also be a pilot study involving analysis of English-language materials. This can serve as a starting point for the development of a language-independent extension of the resulting combination of research methods.

2 Fundamental research questions

The proposed project combines contemporary theories and methods in linguistics, communication studies and political science. It investigates how language is instrumental in shaping meaning and expressing position. Starting out from analysing language and meaning in context, a model will be developed for rich-text mining to be applied in the political domain. In this environment in particular, shaping positions is a communicative process: a dialogue within and between parties, the news media, and the electorate with the specific intention to result in strong and publicly attractive positions. The direct purpose of the project is to enhance the *Kieskompas* and contribute to corpus-linguistic tools for the social sciences in general.

The PhD projects

We propose three PhD projects which, in addition to their independent merits, will work in an integrated fashion to present a three-pronged solution to the question of how to move from words in texts to positions of political actors in a multi-dimensional space.

- The first project (AIO-1) will establish a lexical knowledge base for Dutch and partly for English that models the pragmatic and attitudinal entailments of words and expressions in such a way that it can be used for sentiment and opinion mining.
- The second project (AIO-2) will combine corpus linguistic, cognitive linguistic, and critical methods of discourse analysis as a way of revealing the kinds of framing of different political issues by different parties in the texts as language use.
- The third project (AIO-3) will focus on the actual profiling and positioning of the political parties, and will thus feed into as well as draw on the output of the other two PhD projects.

The fundamental hypotheses/research questions guiding this project are as follows; for the detailed research questions, please see the individual PhD proposals.

AIO-1

Traditional linguistic models use symbolic classes to express properties and behaviour of words in language use. Such traditional representations of linguistic knowledge and their role in language understanding and production have been challenged for many years by the success of machine learning systems. Machine learning systems with shallow features and statistics of language use have been performing better than any complex system of linguistic rules and knowledge. This raises the fundamental issue as to whether or not complex generalizations and knowledge structures exist in our minds at all. We believe however that subjectivity represents a good case for arguing that more complex and generic types of knowledge are involved in natural language processing. Subtle and complex encodings of subjectivity in text require high-level reasoning and inferencing, including empathic reasoning and awareness. The political domain is probably one of the most manipulative and indirect examples of such message framing.

The research thus contributes to the fundamental discussion on the representation of linguistic knowledge as a symbolic system, as a purely statistical system, or possibly in the form of a hybrid level of abstraction in between these systems.

AIO-2

One important goal behind this project will be to investigate how theories from cognitive linguistics about meaning (semantics) as a driving force behind language structure and use can be employed from a critical discourse perspective to complement results obtained computationally from corpus linguistic analysis and to guide the coding of meaning in the computational analysis described above. Cognitive linguistic approaches have been critiqued in the past for overextending claims about individual speakers' language and thought to the behaviour of larger groups of people. Recent work (e.g., Gries & Stefanowitsch 2006) points to ways in which corpus linguistic research can form a bridge to justify claims on the 'supra-individual' level. Thus one hypothesis is that this project will allow us to investigate how the thinking of groups (in this case, political groups) can be modelled. Further, by approaching the findings through critical discourse analysis, a second hypothesis is that we will be able to reveal the implications of the cognitive models being used in terms of the power structures that they covertly invoke while also revealing the dynamics of how they are employed by the political actors concerned.

AIO-3

Research leading up to this project (Kleinnijenhuis & Krouwel 2007; Krouwel, forthcoming) has involved the plotting of political parties within a given country along not just one dimension (such as left- vs. right-wing) but along several, in order to create a multi-dimensional model to better reflect the complexities of the political landscape. The central aim of this project is to develop a more fine-grained model of the dimensionality of competition between political parties. The fundamental questions guiding this study are: (1) which and how many political dimensions underpin party competition and (2) how can these patterns of inter-party differences be extracted from politically relevant texts?

3 Description of the interdisciplinary collaboration

The project will combine research on textual and lexical properties of political text with research on the political positions expressed by the actors in these texts. The detailed and state-of-the-art models that are developed will be formalised as rules and data in such a way that they can be exploited by an existing framework for textual and political analysis. Such a framework is being developed in the *Omstreden Democratie* project of Prof. Kleinnijenhuis and Dr. van Atteveldt. This system can automatically derive a structured database with political facts from a collection of text (see box-4 in Error: Reference source not found below). It will provide an evaluation basis for measuring the contribution of the research in an application setting and comparing this with given state-of-the-art software. The project is thus a unique research collaboration.

The project will combine research on textual and lexical properties of political text with research on the political positions expressed by the actors in these texts. The detailed and state-of-the-art models that are developed will be formalised as rules and data in such a way that they can be exploited by an existing framework for textual and political analysis. Such a framework is the Amsterdam Content Analysis Tookit, which has been developed in recent years as part of the VUBIS/The Network Institute enterprise at the Vrije Universiteit. This toolkit will be developed further in the *Contested Democracy* project of Prof. Kleinnijenhuis and Dr. Van Atteveldt. This system can automatically represent a collection of texts as a semantic network, that can be queried by means of semantic web techniques to arrive at political meaningful textual summaries (e.g. with respect to internal conflict within government coalitions, or with respect to inconsistent means-ends schemes of parties) (Van Atteveldt 2008, chapter 9 – see box-4 in Figure 1 below). The open software environment enables new contributions and new insights to be implemented. The roughly 300,000 coded statements in the system provide an evaluation basis for machine learning and for measuring the contribution of newly proposed improvements in an application setting. The project is thus a unique interdisplinary research collaboration.

Error: Reference source not found then shows an overview of the project. The work of the 3 AIOs is represented by 3 boxes, starting from top to bottom, AIO-1 on Lexical Analysis, AIO-2 on Cognitive, Critical and Corpus Linguistics and AIO-3 on Political Analysis. The AIOs will all work on the same corpus of political texts that is centered in between AIO-2 and AIO-3. The corpus consists of the contributions of each political party in the Netherlands during the post Cold war period of 1990-2010 in three sorts of texts: (1) output of the dominant media outlets (newspaper and television reports, reflecting public debate in society about policy positions of political parties), (2) party manifestoes and web sites (in which they present their 'raw' unmediated policy positions) and (3) party representatives' contributions to parliamentary debates (reflecting the role of the parties in the state). The application of the project is then to automatically extract a database from the text corpus with data on opinions and positions that can be used for tools that support democratic processes. This is symbolized by the broad arrow from unstructured text to a structured database. The other arrows and sub-boxes represent the research and development actions carried out by each AIO to achieve this goal and to build up knowledge. These actions are ordered per AIO, where the first number identifies the AIO and the second sub-number identifies the activity.

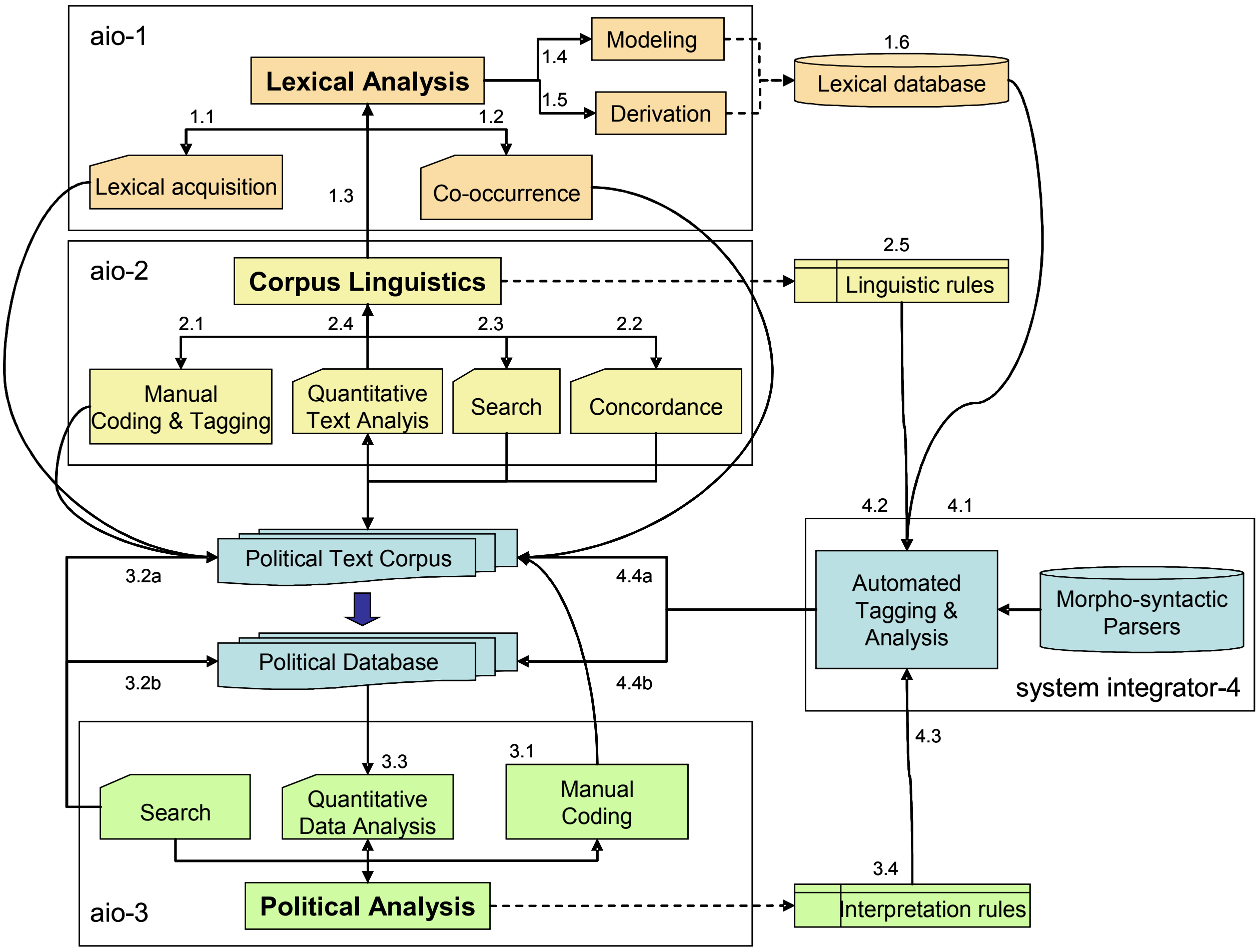

Figure 1: Project overview

The first action carried out by AIO-1 is for example to apply lexical analysis to the text corpus (1.1), whereas AIO-2 carries out manual tagging and coding of the text. Collaboration between the AIOs is based on direct and indirect input. Linguistic corpus analysis of AIO-2 will feed into lexical modelling of AIO-1 but also automatic lexical analysis will benefit from tagging and corpus analysis of AIO-2. On the other hand, the lexical database can be used in the tagging and coding process of both AIO-2 and AIO-3. For example, the names of politicians and their relations to political parties will be stored in the lexical database that is paired to a structured database of participants. AIO-3 will be able to search in the text that is linguistically analysed so that variation in expressing opinions and subjectivity can be handled by a single query. Further, the political coding of text and the database will directly reflect the precision and recall of the corpus analysis and lexical representation to capture this information.

The practical output of the 3 AIOs is symbolized by the output boxes on the right side (1.6, 2.5 and 3.4 respectively). These are formalisations of the knowledge and models by each AIO in the form of rules and a database. These data will be exploited by the fourth component of the project. For this we will use the existing AMCAT text analysis framework, that will be developed further in the NWO *Contested Democracy* project "Short-term media logic and the long-term viability of democracy" by the group of Prof. Kleinnijenhuis (Communication Science) and prof. Van Harmelen (Artificial Intelligence). Both Prof. Vossen and Dr. Krouwel have been participating in this project from its start in 2007. The AMCAT will be developed further also in other projects of The Network Institute (e.g. Dr. Eyal Oren at the Department of Communication Science). The AMCAT enables already the use of standard linguistic knowledge (e.g. thesaurus-matching, POS-tagging, and grammar parsing) for the Dutch language and a database of political players and issues. The system needs to be adapted to incorporate the enriched data. It also has search interfaces, statistical analysis and an evaluation metrics that can be used. Both projects will thus benefit from the collaboration.

The above diagram serves as an overview to show the collaboration and major contributions of the AIOs. Complete details about the activities of each AIO can be found in the full proposals that are attached, with references to this diagram.

4 Global planning

The project has the following global phases:

1. M1-M36:
    a. Studying background literature
2. M1-M6:
    a. Collection of the text corpora
    b. Preparation and first version of the annotated corpus
    c. Definition of the evaluation criteria
    d. Detailed system design
    e. Installation and study of analysis tools
3. M6-M12:
    a. First analysis of the corpus and manual tagging
    b. Cognitive and discourse modelling
    c. Lexical modelling and ontology modelling
    d. Modelling of political data
    e. Running baseline system (version 1) and evaluation on manually annotated data
4. M12-M18:
    a. Formalisation and storage of modelled data
    b. Integration of the improvements to the automatic analysis & second version of the annotated corpus (version 2)
    c. Evaluation of version 2
5. M18-M24:
    a. Second analysis of the corpus

    b. Improvements to the modelling
6. M24-M30:
    a. Formalisation and storage of modelled data
    b. Integration of improvements to the automatic analysis & third version of the annotated corpus (version 3)
    c. Evaluation of version 3
7. M30-M36: Final analysis of the corpus
8. M36-M45: Thesis writing

In the first phase, the corpus will be loaded in the system developed by Atteveldt and Kleinnijenhuis (Kleinnijenhuis, de Ridder & Rietberg 1997; Van Atteveldt 2008; Van Atteveldt et.al. 2008). The text will be automatically processed and tagged with the current system (version 1). Manual tagging will be applied to encode sentiments, opinions and positions in a sample of the text. This will be used for evaluation purposes. The first annotated corpus will not have the results of the deep linguistic and political analysis. The manually tagged sample will be compared to these first automated results that are considered as the baseline.

The corpus will then be analysed by the three AIOs separately, using different technologies and objectives. The analysis can partly be error driven using the manual sample, but will also be based on general empirical techniques and new theoretical models applied to the large corpus.

The evaluations and analysis will as much as possible be exchanged and used across the projects. For example, the insight from the corpus/cognitive/critical linguistic analysis can partly be incorporated in the lexical modelling.

The lexical database can then directly be used in the automatic tagging and processing system (version 2). Also rules and insights from the corpus and political analysis can be used to improve the program, for which implementation algorithms need to be developed. In addition, separate knowledge and theories will be developed by each AIO.

Version 2 of the automated system can be measured against the same manual sample, to measure improvements. The results will again be studied and knowledge and models can be adjusted for version 3 of the system. The final system will then be analysed and evaluated and the results included in the thesis writing.

The aimed start date of the project is September 2008.

Further details on the planning of each AIO are given in the full proposals.

5 Relation to CAMeRA

The project will involve language use and communication by representatives of political parties via speech and writing – (1) in transcribed direct speech (parliamentary debate transcripts), (2) in direct communication from the parties via the medium of their web sites, and (3) as mediated by second parties (newspaper and television). Thus it will involve the analysis of different genres of texts communicated via current popular media. The comparative analyses between these types of texts will help ascertain the relations between the media used and the kinds of political positions expressed.

In turn, the research will link up with the refinement of an existing web-based party-profiling tool (Electoral Compass or *Kieskompas*) for use in the Netherlands, and the further development of such software for other countries: since the method developed to go from text to political position can be extrapolated for use with other languages, it will enable the positioning of parties in other democratic political systems. The application in an electoral compass internet tool will also provide an educational and social benefit from the research.

In addition this research involving the detection of sentiments and viewpoints in linguistic expressions is of more general importance for the development of human-machine interfaces.

6 Relation to LCC

The proposal concerns language, communication, and cognition in a number of respects. (1) The analysis essentially concerns manners of framing of (perspective-taking on) political issues in different ways. As such it would provide a soundly grounded methodology to approach the often subjectively understood notion of framing. (2) The data turned up by mining the linguistic corpora in this project could prove very useful for future research on the use of metaphor in different genres (e.g. here: newspapers and TV news, party web sites, and spoken discourse in parliament), building on the existing NWO-*Vici* program in the Faculteit der Letteren on metaphor in discourse. (3) Furthermore, this kind of semantic modelling in the political domain provides revealing information about cognitive modelling of the relation of oneself to larger social structures, or put another way, of one's own ideas to the ideas of groups with recognised political identities and ideas. (4) Furthermore the research has implications for the study of effective communication by virtue of analysing the lexical and semantic properties of different genres of texts which were all used with the ostensive purpose of persuasive communication. Discourse analysis will contribute by identifying complex interactive strategies typical of persuasive language. Lexical modelling and storage is directly related to research projects (5) run by the Chair of Computational Lexicology (Cornetto and Kyoto).

7 Relation to the Faculty of Social Sciences (FSW)

The project connects with the goals of the *Omstreden Democratie* programme as it will address the essence of a functional democracy: language and communication.

The manifesto texts of the political parties that will comprise part of the corpus of political texts to be analysed have been collected within the framework of the Party Manifesto Research Group (MRG), currently based in FSW at the Vrije Universiteit. The research we will conduct on them will, in turn, form a valuable contribution to this group.

The project relates the work in CAMeRA to the Network Institute, and will strengthen the leading role of the VU in new media, the semantic web, corpus linguistics and semantic network analysis.

8 External collaboration

Collaboration is sought with the Political Mashup project, a web application which facilitates cross-corpus analysis of a large database of political texts collected since 1946 (http://ilps.science.uva.nl/politicalmashup/). Mashup is based at the University of Amsterdam in collaboration with the universities of Tilburg and Groningen.

9 Funding

AIO-1 will be funded by the Faculty of Arts, AIO-2 will be partly funded by LCC and by Camera. AIO-3 will be funded completely by Camera. The project will also provide collaboration with the *Omstreden Democratie* project of Prof. Dr. Jan Kleinnijenhuis.

10 Supervisors

Faculty of Arts: Prof. Dr. Piek Vossen, co-supervisor Dr. Alan Cienki

Faculty of Social Sciences: Prof. Dr. Theo van Tilburg, co-supervisor Dr. André Krouwel

External co-supervisor (AiO2): Dr. Ruth Wodak, Distinguished Professor of Discourse Studies, Department of Linguistics and English Language, Lancaster University, UK
http://www.ling.lancs.ac.uk/profiles/265

11 Proposed external referees

- **Alice Deignan**, Senior Lecturer, University of Leeds, UK
http://www.education.leeds.ac.uk/people/staff.php?staff=31
She is an expert in corpus linguistics and lexicography and is the author of the book *Metaphor and Corpus Linguistics* (2005).

- **Christiane Fellbaum**, Senior Research Scientist, Department of Computer Science, 35 Olden Street Princeton University Princeton, New Jersey 08540-5233, USA.
She is an expert in linguistics and wordnets.

- Mark Franklin, Stein Rokkan Professor of Comparative Politics, European University Institute, Florence, Italy
http://www.eui.eu/Personal/Franklin
He is an expert in electoral studies and heads the research team of the European Election Surveys.
- Eduard Hovy, Ph.D., Information Sciences Institute of the University of Southern California, 4676 Admiralty Way, Marina del Rey, CA 90292-6695, USA.
He is a leading researcher in the field of information extraction and did some breaking research in the field of sentiment and opinion mining.
- **Adam Pease**, Principal Consultant and CEO of Articulate Software. Formerly, he was Director of Knowledge Systems at Teknowledge, where he led a group conducting research and applications in ontology and knowledge based systems.

- **Elena Semino**, Senior Lecturer, Lancaster University, UK
http://www.ling.lancs.ac.uk/profiles/17/
She is an expert in corpus linguistics and also cognitive linguistics, and has experience applying this research to the genres of political speeches and news articles.

- **Stefaan Walgrave**, Professor of Political Science, University Antwerp, BE
http://www.ua.ac.be/main.aspx?c=stefaan.walgrave

He is an expert in political parties, social cleavages, party competition and party positioning. He is also part of a research team that is responsible for the development of a party profiling website (*De Stemtest*).

- **Jan Wiebe**,Professor, Department of Computer Science, Director, Intelligent Systems Program, University of Pittsburgh, PA, USA.
She is a leading researcher in the field of automated sentiment and opinion mining.

12 AIO-1: Lexical model and acquisition for sentiment and opinion analysis in Dutch text

Supervisor: Prof. Dr. Piek Vossen
Vrije Universiteit, Fac. der Letteren, LCC

Introduction

Most language use does not express facts but expresses personal opinions and positions with respect to facts, often even disguised for some communicative or manipulative goal. This is evident in the structure of our language system and in the daily use of language. Nevertheless, most Natural Language Processing systems treat text as factual and naively index it as if it is all true.

More recently, there is increasing interest in the automatic identification and extraction of opinions, emotions, and sentiments in text. This is motivated by the desire to provide tools for information analysis in government, commercial and political domains. The internet offers many new media and communication systems that result in distributed and individualized information and news feeds. Information is thus no longer coming from a few central sources but is more and more rapidly exchanged between small groups and networks that use their own channels. Access to information and the speed of information exchange sometimes leads to turbulent changes in the public opinion. This can have dramatic effects for governments, politicians, marketing and position makers and companies, since they have less control on these opinions and visions of the public and on their role in society. For these groups, it is essential to have precise tools that can handle these rapidly changing and massive amounts of information, spread over many diverse sources. There is a big need for such systems that can also couple sentiments and opinions to themes and to groups and individuals.

- **Language Technology**

Machine learning and statistical techniques have been used for many information and language technology systems in the last years. Researchers from many subareas of Artificial Intelligence and Natural Language Processing have been working on the automatic identification of opinions and related tasks (e.g., Kim and Hovy. (2004, 2006), Liu et al. (2005), Riloff et al. (2003), Turney and Littman (2003), and Yu and Hatzivassiloglou (2003), Takeshi (2008)). Most such work has focused on sentiment or subjectivity classification at the document or sentence level. Document classification tasks include, for example, distinguishing editorials from news articles and classifying reviews as positive or negative (Pang et al., 2004; Yu and Hatzivassiloglou, 2003). A common sentence-level task is to classify sentences as subjective or objective (Yu and Hatzivassiloglou, 2003; Riloff et al., 2003).There are many publications and even commercial systems that use these techniques. Although most of these systems are fast and robust they have a serious drawback. The level of analysis is relatively shallow and they cannot handle subtle linguistic expressions and structures. Typically, the epistemic, attitudinal and deontic implications of language are found in more complex expressions and structures or are hidden in subtle lexical implications. Likewise, these shallow approaches mostly miss the point.

Furthermore, these methods typically require training and test corpora that have been manually annotated. The annotation of such corpora is a labour intensive task. Besides it has been shown that derived systems have a strong dependency on the training corpus. Results degrade enormously when applied to other domains or genres.

- **Linguistic Research**

Research on subjectivity in text has also been done from a linguistic point of view. The expression of attitudes and opinions in text is not trivial. First of all, the association of a value towards themes and properties is scalar rather than polar. This information should be reflected in the resources used for assigning attitudes and sentiments to text fragments. Secondly, texts are often very rich in terms modal, and epistemic expressions, e.g. beliefs and situations can be confirmed, desired, wished for or negated, denied or rejected in many subtle ways. Furthermore, language can be ironic, sarcastic or use metaphors as stylistic means to express subtle opinions that are extremely difficult for computer programs to detect. All these aspects have been studied for years in the field of linguistics and pragmatics (for example Palmer 1986, Billig 1995, Rodman 2001, Auwera & Schalley 2004, Aikhenvald 2004). The same holds for research regarding subjectivity. Currently concepts like 'subjectivity', 'intersubjectivity', 'objectivity' and 'speaker-involvement' have become increasingly popular in cognitive semantic literature (Langacker 1990, Traugott & Dasher 2002). In this respect the role of prepositions, modal verbs, modal adverbs, epistemic adverbs, attitudinal adjectives are studied, focussing on the relations between subjectivity and the semantic and syntactic behaviour of these items. De Smet and Verstraete (2006) distinguish distinct types of subjectivity categories, like 'pragmatic subjectivity' and 'semantic subjectivity'. The drawback of these approaches is that studies are restricted to small and diverse samples of words and phrases and that this knowledge is not aggregated and centrally stored in a lexicon and rules.

If that would be the case, the results of these studies can be exploited to derive a better model for the representation of explicit and implicit attitudes of speakers in text than is currently done in automated sentiment and opinion mining. However, combining these linguistic descriptions in an effective application with all its pitfalls is far from trivial.

- **Combining language technology and linguistic research**

Recently, more and more tools are developed that combine statistical techniques with linguistic insights. The linguistic knowledge is then stored as lexical knowledge. These approaches, however, do not take into account all subtle linguistic nuances of natural language expressions. Data on sentiment and attitudes, epistemic and modal properties in lexical data are usually sparsely encoded and poorly modelled. They mostly just use lists of words or phrases (as in, Osgood et al. (1957), Heise(2001), Subasic and Huettner (2001), Banea (2008). Andrea Esuli and Fabrizio Sebastiani (2006 and 2007) encode this information at the level of a synset in SentiWordnet.

Furthermore, classifications of verbs, adjectives and adverbs in lexicons tend to ignore the logical implications and knowledge claims those words may entail (deontic and epistemic entailments). Although there is a semantic hierarchy of words of *saying* in Wordnet, there is no explicit definition of the subjective implications of these words. Below *assert* we find more specific verbs (hyponyms) such as *claim, reaffirm, confirm, reassert* and also *maintain* and *defend* (and many more) that play different subtle roles in for example political debates. These roles are not encoded in the lexical semantic relations of Wordnet.

Even larger ontologies such as SUMO and MILO (Niles and Pease 2001; Niles and Terry

2004) lack a proper classification of 'loaded' words and expressions. Most of these are simply classified as subjective and even then the encoding is not consistent. A more advanced classification of adjectives is given by Hundsnurscher and Splett (1982), but it also neglects pragmatic aspects of words.

Maks et al. (2008) try to combine these semantic frameworks as a first descriptive apparatus for adjectives that can serve as a basis for a lexical model for pragmatic and attitudinal entailments. The lexical model consists of complex and typed knowledge structures that represent generalizations of classes of words and expressions. The classes stand for the structural combinatoric and semantic behaviour with a focus to subjective entailments. They can be assigned to different synonyms within a synset or to complete synsets in the form of lexical semantic relations or ontological labels. The entailments are modelled through lexical representation layers that relate the subjective implications to the proper structural and semantic properties.

The differentiation of synonyms within synsets and the differentiation of synsets within a lexical semantic network by generic and abstract classes predict their overall usage and communicative effect. In addition to this class-membership, other specific or idiosyncratic properties of these words and expressions can be stored in the lexicon as long as they do not violate the constraints of the class. These classes represent complex knowledge structures with different layers of features, varying from morpho-syntactic, to semantic and pragmatic aspects (compare the typed-feature structures representations in HPSG [Pollard and Sag, 1987, 1994]). As such they are theoretical constructs that represent the most specific and most general typing of sets of words and expressions, i.e. predicting most properties for most cases of words and expressions respectively.

The representation of linguistic knowledge and its role in language understanding and production has been challenged for many years by the success of machine learning systems. We believe however that subjectivity represents a good case for arguing that more complex and generic types of knowledge are involved in natural language processing. The research thus contributes to the fundamental discussion on the representation of linguistic knowledge as a symbolic system, as a purely statistical system, or possibly in the form of a hybrid level of abstraction in between these systems.

It is in this area, - i.e. in the combination of cognitive linguistic and lexical research and language technology tools - that the present research must be situated.

Fundamental Research Question and Hypothesis

Traditional linguistic models use symbolic classes to express properties and behaviour of words in language use. These classes range from very shallow and broad generalization such as nouns and verbs, to very detailed and complex knowledge structures that combine morpho-syntax with semantics and pragmatics. Such traditional representations of linguistic knowledge and their role in language understanding and production have been challenged for many years by the success of machine learning systems. Machine learning systems with shallow features and statistics of language use have been performing better than any complex system of linguistic rules and knowledge. This raises another more fundamental issue whether or not complex generalizations and knowledge structures exist at all in our minds.

We believe however that subjectivity represents a good case for arguing that more complex and generic types of knowledge are involved in natural language processing, leaving aside what is the proper implementation of such as system, as a lexical database or as a corpus. Subtle and complex encodings of subjectivity in text require high-level reasoning and inferencing, including empathic reasoning and awareness. Associative systems used in machine-learning cannot grasp the deeper motivation and interpretation of these textual constructs at these levels. The political domain is probably one of the most manipulative and indirect examples of such message framing.

The research thus contributes to the fundamental discussion on the representation of linguistic knowledge as a symbolic system, as a purely statistical system, or possibly in the form of a hybrid level of abstraction in between these systems.

The hypothesis underlying the research of AIO-1 is that linguistic knowledge on subjectivity can be adequately modeled in a rich semantic lexicon that represents a symbolic abstraction of statistical linguistic experience. Furthermore, this knowledge can be acquired automatically such that it results in a significant improvement in the performance of the current state-of-the-art opinion and sentiment analysis tools.

The hypothesis is validated both by the descriptive adequacy of the lexical types and their relations, and by the empirical validation in a natural processing system that tries to mine opinions and sentiments from text.

Research questions

Given this hypothesis the following research questions arise:

1. What knowledge is involved in the expression of attitudes and opinion in Dutch? The main linguistic research question here is: what is the relation between structural textual and genre properties and the encoded sentiment and position information. A distinction will be made between attitudinal properties, such as positive, negative and neutral on the one hand, and epistemic and modal properties on the other hand: beliefs, opinions, (im)possibilities, assumptions, impressions, statements, assertions, etc. Attitudes can be both implicitly and explicitly expressed and can be related to any segment of text and even complete documents. Epistemic expressions are usually more explicit and related to specific people, groups or institutes that are the subject of the expression. Furthermore, they connect the subjects to specific themes towards they have an opinion.

Sub-questions:

- Which lexical items are markers for subjectivity and sentiment?
- What syntactic and semantic information is relevant in relation to subjectivity?
- What models for text analysis exist?
- What tools for lexical data analysis exist and which are best?

Method and Data:

- Literature on subjectivity, on corpus-related research, on NLP with regard to sentiment and opinion analysis.
- Lexical and linguistic analysis of a small sample of Dutch opiniated text.
- Comparison of existing guidelines for annotating emotions in text
- Comparison of existing annotated Dutch and English corpus data, e.g. the corpus Verdonk (UvA) , the corpus Omstreden Democratie (VUA), the MPQA corpus annotated

for opinions and sentiments, Janyce Wiebe, USA, etc.

2. How can this knowledge be modeled in a rich semantic lexicon? What knowledge is lexical knowledge and how can this be encoded in a lexicon?

Which subjective expressions (i.e. words and phrases being used to express opinions, emotions, evaluations, speculations, etc (Wiebe et al. 2005, Quick et al. 1985) can be encoded in a lexicon. Besides a model and description of the structures, expressions and words that can carry this information, the research will also investigate how this knowledge can be encoded in a rich semantic lexicon. The information can be encoded on various levels and it needs to be modelled properly to make the correct inferences. Lexical modelling also includes relations between semantic classes of words. Antonyms and co-hyponyms can display various degrees and types of oppositions that play a role in expressing sentiments. Hyperonym relations group together many verbs and nouns of assertion that differ in many subtle ways with respect to their attitudinal and deontic implications. A precise description of these types of verbs and nouns is required to achive a proper lexical modelling.

Sub-questions:

- What linguistic knowledge on subjectivity can be encoded in a lexical resource and which not?
- How does lexical subjective knowledge relate to ontological knowledge?
- How is combinatoric information on words related to subjective knowledge?
- Which sentiment lexicons are already available for Dutch or other languages?
- Which sentiment information is already encoded in existing lexicons?
- Which lexicon models do exist for sentiment/subjectivity data?
- Which lexicon model is suitable for the representation of subjectivity/sentiment data?
- Should subjectivity labels be assigned to lexical units, to synonym sets, to ontological classes or to other units?

Method and Data:

- Literature on lexical modeling
- Comparison of existing lexical resources (Framenet, Wordnet, Ontologies, Sentiwordnet, WordnetAffect, Mikrokosmos)
- Implementation of lexicon model for subjectivity information

3. How can this knowledge be acquired from implicit information and other sources?

Recent research (Banea et al. 2008, has shown that shallow sentiment ratings can be derived automatically from seed lists applied to corpora or semantic networks in English. We need to investigate how this can be extended to more complex and subtle representations and applied to Dutch.

Sub-questions:

- What techniques do exist for automatic lexicon building?
- Are they applicable in this case?
- Is the acquisition at synset level , sense level, ontological level or all?
- Can sentiment information be transferred across languages?

Method and Data:

- Literature on automatic lexicon building
- Manual creation of a gold standard Dutch lexicon file for evaluation
- Development and testing of algorithms for Automatic acquirement of complex lexical data, resulting in an automatically acquired sentiment lexicon for Dutch (Lex-1)

- Development and testing of algorithms for transfer of sentiment data from English to Dutch, resulting in a transferred sentiment lexicon for Dutch (lex-2)
- Development and testing of algorithms for transfer of sentiment data from Dutch to English, resulting in a transferred sentiment lexicon for English (lex-3)
- Evaluation of acquired Dutch lexicon (lex-1) against gold standard for Dutch
- Evaluation of transferred Dutch lexicon (lex-2) against gold standard for Dutch
- Evaluation of transferred (from Dutch to) English lexicon (lex-3) against English Sentiwordnet

4. How can this resource be used within an opinion mining tool?

The most effective combination of statistical and machine learning methods versus knowledge-rich and rule-based systems is a topic of research. In addition to these generic methodological issues, automated opinion mining also needs to make a distinction between the Speaker to which the opinion is assigned, the issue or topic on which the Speaker expresses an opinion and the opinion itself. This means that expressions and statements of Speakers need to be analysed to a rather detailed level by some NLP module.

On the one hand, lexical knowledge can be used in case of sparse data or domain-dependence of machine-learning systems. On the other hand, rule-based systems can be extended with probabilities that can be learned from tagged text corpora.

We will test 3 types of systems:

A: a system based on statistics and machine learning
B: a system using only linguistic knowledge and lexical resources
C: a hybrid system based on a combination of the methods used in A and B

Sub-questions

- What techniques exist for automatic opinion and sentiment analysis?
- How can linguistic knowledge that is stored in a lexical semantic database and in linguistic rules, be exploited for automatic analysis of sentiments and opinions?
- How can knowledge-rich rules be combined with statistical methods or supervised machine learning methods that require a tagged corpus?
- Can a more-fine-grained analysis of opinions be achieved by combining these methods?
- Is there evidence that lexical knowledge can make existing sentiment mining systems more robust and less-dependent on the training domain?

Method and Data

- Definition of test data and manual preparation of the test data
- Implementation and evaluation of 3 types of sentiment analysis tools
- Implementation and evaluation of sentiment analysis tools reported in the literature on existing evaluation data

Time Schedule

Phase	Task	Start	End
	Literature research	M0	M24
	Analysis of text data	M0	M12
Lexical analysis	Import of word lists and statistical data	m3	
	Evaluation of baseline, version 1	m6	m7
	Lexical modelling-1	m6	m9
	Lexical acquisition-1	m6	m12
	Evaluation of the version 2	m12	m15

	Evaluation of lexical acquisition-1	m12	m15
	Lexical modelling-2	m15	m24
	Lexical acquisition-2	m15	m24
	Evaluation of the version 3	m24	m30
	Evaluation of lexical acquisition-1	m24	m30
	Thesis	m30	m45

13 AIO-2: Positions in political texts: Integrating Cognitive linguistic, Critical discourse and Corpus linguistic analyses (CCC)

Supervisor: Prof. Dr. Piek Vossen
Daily supervisor: Dr. Alan Cienki
Vrije Universiteit, Fac. der Letteren, LCC
Co-supervisor: Prof. Dr. Ruth Wodak
Lancaster University, UK

Problem addressed

In political communication a major problem faced by researchers is how one can systematically identify the policy positions of political candidates and parties from the written and spoken texts produced by them and written about them. Existing methods usually involve either the coding of texts by hand, or techniques of computational analysis involving word counting. The former method is time-consuming and subject to human errors, a problem which is compounded as the analysis of larger text databases involves coordinately more coders. The latter, computational method is more efficient, but has the downside that one loses the connection between the words counted in the data and the significance of the contexts in which they occurred (Baker 2006).

Meanwhile, in contemporary linguistics there are three areas that developed in the latter half of the twentieth century which can address different aspects of the problems described above. One is cognitive linguistics – a collection of theories and methods which strive for explanations of language structure and use that are based on what is known from cognitive psychology about human thought and reasoning. A second is critical discourse analysis (CDA) – an interpretive means of analysing power structures in social institutions as revealed through the use of language in them and about them. A third is a field of study which handles the analysis of large collections (bodies or *corpora*) of machine-readable texts in order to reveal phenomena such as patterns of language use over many individual speakers or writers; this is known as corpus linguistics.

However these three approaches have largely been followed independently of each other by different groups of linguists. Some researchers have begun to combine two of them, e.g., cognitive linguistics and CDA (Hart & Lukeš 2007), cognitive linguistic research on metaphor with corpus linguistics (Deignan 2005), and CDA research with computational techniques of analysis (Baker 2006). The possibility of integrating all three has just begun to be explored, and so far only in projects with limited scope, such as small-scale metaphor analysis (Charteris-Black 2004). As yet there is no research which provides a methodology for combining all three of these approaches in a comprehensive way. Yet the potential benefits of bringing together their complementary insights are immense. The analysis proposed for this PhD project (AiO 2) will involve integrating corpus, cognitive, and critical linguistic (CCC) approaches to provide input for, and receive input from, the computational techniques (AiO 1) that will help automate a large portion of the analysis. This will be of immense use in analyzing the political texts with the goal of mapping the various political actors' positions relative to each other (AiO 3).

Research questions

Specifically, the research questions to be addressed in this project include:

- How are social issues framed in political discourse in different ways via choice of words? Cognitive linguistic theories will provide the basis for a typology of such frames, taking into its model theories on linguistic and conceptual framing and cognitive semantics (metaphor, metonymy, cognitive/socio-cultural models, language of emotion [appeals to sentiment], etc.; see Cienki 2007). This typology will form one basis for the corpus tagging.
- What do the power dynamics of political discourse entail? Critical Discourse Analysis (CDA) theories will add aspects of the communicative/interactive nature of political texts with an intent to persuade. In addition, a CDA approach will provide ways to distinguish groups in social settings with particular linguistic features related to sentiments and identity (van Dijk 1997; Wodak & Krzyzanowski 2008).
- How can political discourse be categorized to make this knowledge suitable for automated recognition in large corpora of political texts? A typology needs to be worked out to provide qualitative features to the automated analysis model (Baker 2006; Wodak & Chilton 2005).
- What are the flaws in automatic text mining and how can manual analysis—namely using using cognitive semantic and CDA techniques—compensate for such flaws? The model needs to be optimized but we want to take notice of the richness of the texts and add further manual analysis.

Description

This project will focus on the question of how the relatively saliently high or low use of words and phrases by a political party in the texts produced by and about them may reflect that party's particular position towards, and framing of, different issues. The research will involve an integration of the three different approaches within contemporary linguistics described above—corpus linguistic, cognitive linguistic, and critical discourse analytic—in the analysis of a large corpus of Dutch political texts with comparison to a small corpus analysis of English political texts.

The combination of these theories and methods (called here CCC) represents an innovative approach in linguistics which gives this project its own inherent integrity. Within the overarching framework of this larger project on political discourse, it will provide a linguistically-grounded approach to better understand the positions of political parties in relation to each other.

Data

The first part of the research will involve comparing the contributions of each political party in the Netherlands during the post Cold war period of 1990-2010 in three sorts of texts: (1) output of the dominant media outlets (newspaper and television reports, reflecting public debate in society about policy positions of political parties), (2) party manifestoes and web sites (in which they present their 'raw' unmediated policy positions), and (3) party representatives' contributions to parliamentary debates (reflecting the role of the parties in the state). Note that the content of the first two text groups (call them sub-corpora) — the newspaper and TV data and the party manifestoes — have been coded by hand in previous research: the former have been coded for attitudes of actors towards one another and towards salient issues; the latter has been coded for issue saliency and ideological position. However the content of the third text type, the sub-corpus of parliamentary debates, has not been hand-coded. Therefore the second part of the study would involve comparing the party-specific use of language from the sub-corpus about which we know little, that of the parliamentary debates, with the two sub-corpora which have been coded in detail. These inner-corpus and cross-corpus comparisons could also allow for the assessment of variation in policy positions of the actors. In other words, this research could also reveal differing effects of media genre — how parties present different positions more or less saliently in different contexts (written and spoken news media, web sites, and parliamentary speeches). We are in contact with the group based at the University of Amsterdam that is developing the web application called Political Mashup (www.politicalmashup.nl), and this will provide a helpful means of analyzing the text types as composites, rather than having to treat each as a set of individual documents.

In addition a small, focussed comparative study is proposed involving the analysis of a corpus of British political texts, such as a set of speeches by one of the Prime Ministers. The motivation for this is two-fold. First, within the Dutch data, the contributions by the Dutch Prime Minister (PM) have a special status, given the PM's role among the all the political actors represented. Pulling out the contributions of the PMs as a sub-corpus will allow for analysis of this special discourse. Comparison with comparable data from the PM of another country (namely the UK) will allow for more valid generalizations about the contributions of this important participant in the political discourse of a given country. Second, analysis of such a comparable data set in English, a language closely related to Dutch, would serve as an initial test of the potential for extending this research method to another language. Thirdly, the methodology for CDA has been developed from, and primarily applied to, English-language textual data. This also applies to the computational techniques being developed for CDA research. Therefore this project could benefit from the experience of scholars who have done this work with materials in English, namely those at Lancaster University in the UK, with whom we have several contacts, including Prof. Dr. Ruth Wodak, Dr. Veronika Koller and Dr. Elena Semino. Prof. Wodak herself has expressed interest in this project and has agreed to serve as a co-supervisor of this PhD student.

Methods

In order to extract (relative) position on issues and salient conflict dimensions (AiO3's project), the AiO2 project will analyse these three corpora of texts separately according to each political party, and in addition the texts of each type will be analyzed together per party. This will reveal differences between the positions expressed by the parties in the three contexts in which the texts were produced, and also will provide a view of the cumulative position of each party across the different contexts. It will also allow for a comparison of party positions over time and the changing 'policy-distance' of parties along salient political cleavage lines.

Lexical semantic analysis of the texts will be done using concordancing software (such as PhraseContext <http://www.hjkm.dk/PhraseContext/>), which not only provides information on word frequency counts, but also can collect all the instances of the phrasal contexts in which any given word was used. In addition, the software allows for statistical analysis of the unusually high or unusually low frequency of words within any one (sub)corpus compared to the others being analyzed at the same time. This information is provided in terms of the log-likelihood of over- and under-use of words in any sub-corpus (such as a particular party's contributions in these texts) relative to the whole corpus (the discourse produced by all of the parties) [step 2.4 in Figure 1 of the main proposal text]. Using approaches from cognitive linguistics, the words and phrases thus derived will then be coded and analyzed [step 2.1 in Figure 1] for their use of metaphor (as a means of characterizing the abstract in terms of the concrete; Cienki 2008), metonymy (ways of mentioning what the speaker or writer intends indirectly via salient cognitive reference points), and therefore framing of issues (revealing assumptions and models for thinking and reasoning about them; Cienki 2005). Approaching this research from a critical discourse perspective will facilitate the unveiling of the power relations expressed via language. (See Kooi 2007 as a preliminary example of an analysis combining these methods.)

One specific aspect of the study will involve comparing the party-specific use of language from the sub-corpus about which we know little, that of the parliamentary debates, with the two sub-corpora which have been coded in detail, those of the media outlets and the parties' own web sites. Comparing each party's salient use of particular words and phrases (key terms) against the background of the latter two reference corpora would allow for determination of the semantic valence of those key terms in the realm of Dutch politics through computational analysis of the other words with which they co-occur most frequently, known as their collocates (e.g., 'taxes improve the social and physical infrastructure' versus 'taxes inhibit economic growth') [steps 2.2 and 2.3 in Figure 1]. This means there will be a clear and replicable method for determining how specific words and phrases are used by the parties to frame political positions. For example, the contexts of word use provided via the concordancing software can allow for analysis of metaphorical use of key terms, and of correlation of key words with certain syntactic roles (e.g., frequent use as grammatical subjects or objects, or in other words, as *agents* or *patients*). Both of these are known to be framing devices in that they are factors which focus information according to specific perspectives. Analysis of the words/phrases in context will also show the saliently frequent use of semantic associations (also known as semantic prosody; Nelson 2006) with these words within larger discourse structures, further augmenting the interpretive analysis of the texts.

Goals

For the analysis of these political data, the research will reveal the implications of relatively high and low frequency word/phrase use by the different political parties in terms of, for example, the semantic roles played by such words. The critical discourse approach will add the perspective of analyzing the potentially persuasive and even manipulative use of such words. From the perspective of linguistics, an important outcome will be the development of a means for integrating cognitive linguistic theory and corpus linguistic methods with critical discourse analysis.

The results from this “CCC” project will be used in the tagging of the lexical database in the first AiO project on sentiment and opinion in a lexical knowledge base [step 1.3 in Figure 1]; this will then feed into the automated tagging and analysis of the political corpus.

In addition, the CCC approach to text analysis also has potential for future application in other domains in which the relative differences between corpora of argumentative texts is of interest, for example, in the analysis of business communication or of media bias in varous respects.

Preliminary time planning

2008 - Literature study on the key concepts of this research project, in particular recent work on corpus, cognitive, and critical linguistic analyses of political texts - Development and refinement of the CCC methodological framework - Become acquainted with the various relevant departments/research groups (institutionalising cooperation) - Familiarisation with various analytical tools and programmes
2009 - Become familiar with the two existing coded databases/corpora of Dutch political texts (from media outlets and party web sites) - Analyse the two existing coded corpora of Dutch texts using the CCC framework - Compose and analyse a small corpus of English PM texts for comparative study - Participation in and drafting of a paper for a conference on the CCC methodological framework - Publication of an article in refereed journal on the methodological framework or the comparative Dutch-English study
2010 - Refine the analysis of the two pre-coded corpora based on feedback from the AiO 1 project - Conduct the analysis of the third corpus (parliamentary debates) using the refined method - Finalisation of the CCC methodological framework - Publication in international refereed journal (based on empirical findings of CCC method of analysis) - Co-organisation of a conference and/or research group meeting in order to compare findings with other types of research analysing political texts
2011 - Writing of PhD dissertation - Cooperation in edited volume on the basis of the conference

Project supervision

Proposed supervising professor: Prof. Dr. Piek Vossen

Proposed daily supervisor: Dr. Alan Cienki
Dr. Cienki is lecturer in the Dept. of Language and Communication in the Faculty of Arts at the VU; a member of the Language, Cognition, and Communication (LCC) research group; and has been a participant in the Amsterdam Discourse Centre of the UvA. He has served as Director of the Program in Linguistics at Emory University in Atlanta, USA, and there he was also associate professor in the Graduate Institute of the Liberal Arts. Dr. Cienki is co-editor of *Conceptual and Discourse Factors in Linguistic Structure* (Stanford: CSLI, 2001) and of *Metaphor and Gesture* (Amsterdam: Benjamins, 2008), and is on the editorial board of *Cognitive Linguistics* and the executive board of the Researching and Applying Metaphor international association. He is a founding member of the international metaphor research group "Pragglejaz", initiated by Prof. Dr. Gerard Steen, which has developed a reliable method for the identification of metaphorically used words in texts (Pragglejaz Group 2007). Dr. Cienki is also a co-supervisor of the four PhD students in the NWO-*Vici* project

"Metaphor in discourse" in the Faculty of Arts which is supervised by Prof. Steen. The research of the Pragglejaz Group and that of the *Vici* PhD students will both be relevant to this project, and the members of both of those groups and this AiO will benefit from contact with each other.

Proposed co-supervisor: Prof. Dr. Ruth Wodak
Prof. Wodak is Distinguished Professor of Discourse Studies in the Dept. of Linguistics and English Language at Lancaster Universitiy, UK, where she moved from the University of Vienna where she had been Professor of Applied Linguistics. She is one of the founders of the field of CDA and is the author or (co-)editor of such books as *A New Agenda in (Critical) Discourse Analysis*, *Methods of Critical Discourse Analysis*, *The Discursive Construction of National Identity*, *European Union Discourses on Un/employment*, and *Communication in the Public Sphere: Handbook of Applied Linguistics, Vol. 4*. Prof. Wodak is co-editor of the journals *Discourse and Society*, *Critical Discourse Studies*, and the *Journal of Language and Politics*. Further details of her distinguished record can be found at <http://www.ling.lancs.ac.uk/profiles/Ruth-Wodak/>.

14 AIO-3: Party profiling and mapping

Supervisor: Prof. Dr. Theo van Tilburg
Daily supervisor: Dr. André Krouwel
Vrije Universiteit, Faculty of Social Sciences

Problem formulation

In advanced industrial democracies processes of social mobility and emancipation have left modern citizens with weaker group loyalties and fewer institutional links with party-political organisations, resulting in less traditionally ‘partisan’ voting behaviour (Dalton and Wattenberg 2002). New voter groups (particularly younger generations and immigrants) enter the electorate with even less party-political socialisation (Kitschelt 1997; Dalton 2002). The core vote for traditional parties is declining as voters are more likely to change their party preference from election to election (Dalton and Gray 2003; Dalton 2004; Franklin 1992; 2003), while at the same time the overall number of political parties participating in national elections increases (Webb 2002). With the disappearance of profound ideological differences between the major political parties, voters face an increasingly difficult task of determining what the policy positions of parties are (Krouwel, forthcoming). This is problematic as research also shows that issues and policy considerations have become more important for voters (Franklin 2003; Thomassen et al 2000). Against this background of increasingly weaker party attachment of voters, the proliferation of new parties into a more fragmented and multi-faceted political landscape, as well as the abandonment of traditional policy positions by the main political actors, Europe’s voters face a challenging decision problem: Whom should I vote for?

In order to make an informed decision, they need access to high quality and reliable information (Baron, 1988). Meeting the information needs of voters becomes ever more challenging as politically relevant information is increasingly locked up in large, heterogeneous, dispersed and often unstructured (digital) sources. Voters need and seek support in extracting necessary facts from these complex information environments when they are faced with difficult decisions. As a result, several initiatives have emerged using e-technologies for enhancing electoral choices of voters (Gibson et al 2004). The rapid development of information technologies – particularly internet – are already playing a significant role in providing political information, with important effects on the functioning of modern party democracy (Treschel et al. 2003; Frick 2003). Several of these online efforts have focused on developing internet solutions to allow voters to match their own political preferences with issue positions of political parties and candidates.

Online Party Profiling

Web-based party profiling tools have been around for a number of years and come in many guises (Laros 2007). Party profiling tools make voters aware of their own issue positions, their location in the political spectrum and their proximity to the relevant parties or candidates. Most party profilers output a ranked list of parties, based on a linear notion of distance to the voter that uses the system. In contrast, the Dutch profiling site KiesKompas (ElectoralCompass), developed at the VUA, is based on a multi-dimensional space, represented in a two-dimensional format. Most tools compute a distance between the voter and each of the parties for each issue included in the test. A important distinctive feature of the Electoral Compass profiler is that it positions voters and parties in a multi-dimensional space, based on the use of averaged positions for parties and for the individual voter with respect to a set of specific propositions (Kleinnijenhuis and Krouwel 2007). The issues included in the questionnaire are themselves identified using an extensive analysis of party programs, party web sites and other channels of expressions used by parties and politicians. In the interface, a symbol (for example a red pencil) denotes the voter's position, while the parties are represented using their logos. The ElectoralCompass profiling tool does not provide a simple and restrictive voting advice, but will allow users to make a rational analysis and decision, with direct access to information on the positions of parties and candidates with regard to a large number of salient issues. The tool does not entirely predetermine the saliency of issues, but will allow voters to position themselves with regard to meaningful subsets of issues so as to enable them to reduce the complexity of a key political choice: "Which party or candidate is closest to their own position?"

Text analysis and focused position extraction

Before we can compare voter preferences with party positions on issues, we first have to determine what (relative) position parties adopt on each of the issues. This project focuses on the question of how to systematically extract policy positions of political actors (candidates and parties) from texts of the documents and speeches they have produced. A related problem is how to assess the relative position of each political actor on a policy scale or ideological dimension on which parties compete with one another.

Language is one of the main weapons of politicians. The power of words – by which politicians communicate their ideas in speeches, written propaganda and mass media – structures the evaluation of these actors by opponents and citizens. Hence, analysis of political texts is crucial to grasp the competitive interaction between political actors. Content analysis has benefited from the development of computer-based research methods, as larger quantities of text can now be examined. Yet, extracting political meaning from texts has proved a difficult task. Existing methods for large-scale text analysis to extract issue-positions of political actors have commonly been of two types. The first involves an interpretive approach in which texts and discourse are interpreted and hand-coded (Budge et al. 1987; Laver and Budge 1992; Klingemann et al. 1994; Garry 2001). The drawbacks of this methods are the high costs as hand-coding is labour-intensive, expensive to replicate or change, as well as sensitive to errors of interpretation and inter-coder reliability. Another approach is to analyze text as data through word-scoring methods, yet with these type of analyses one has to curb problems of reduction to issue-emphasis, meaningless 'number-crunching' and de-contextualization. Laver, Benoit and Garry (2002; 2003) have developed a new 'probabilistic word scoring' method, which statistically compares patterns of word frequencies within a set of texts of which something is known (reference texts), to analyse word frequencies within texts which have not been subject to interpretative analyses (virgin texts). Such analyses have been employed to estimate policy positions of parties in different polities (Benoit & Laver 2002; Benoit et al. 2005; Laver, Benoit & Sauger 2006). This method also has some shortcomings, particularly with regard to the transformation procedure of the word frequencies into positional codings and the selection of reference documents (Martin & Vanberg 2007). Others have relied on counting the (co-)occurrence of words and phrases in order to study agenda-setting and public awareness of issues (McCombs 2004), but these counting methods of associations and disassociations between actors and issues are unable to pin down patterns political competition. Some studies have tried to tackle this problem by using dictionary-based comparisons (Bara 2001; de Vries et al. 2001; Kleinnijenhuis & Pennings 2001). Yet, the development and testing of dictionaries and codebooks does still require substantial resources in time and effort and the dictionaries are sensitive to linguistic, cultural and political context. Computational linguistic methods have been develop to improve the recognition of actors and issues (Jurafski & Martin 2000), however politics is more than the saliency of issues. The problem remains how to extract positions, how to determine the 'relative position' of the various actors on specific issues and how to relate those issue-positions to more abstract dimensions of political conflict. During recent years, more automated linguistic tools have become available (e.g. lemmatization, POS-tagging, grammar parsing), whereas new developments in computer science and artificial intelligence (description logics, semantic web, network visualization techniques) open new windows as well. Using such techniques, issue positions of parties can be extracted reasonably well from newspaper texts (Van Atteveldt et.al. 2008: F1-score of 0.55), but a still higher precision and recall are required.

Thus far, no comprehensive study has tried to combine all available methods so as to minimise the weaknesses of the individual approaches and maximise the benefits of each of the approaches. This proposed PhD project aims to combine and compare various methods to answer the research question: How can a combination of various methods of large-scale text analysis tackle the problem of going from the language used in the texts to the positioning of the parties along salient political dimensions? The main goal of the project is to develop a automated method of extraction of political positions of actors from relevant texts.

In this focused position extraction, will use corpora of texts that are relevant in terms of political competition. As a crucial case analysis, we need to select those corpora of texts in which parties and candidates will most clearly reveal their political stances. Our assumption is that three corpora will most likely reveal those issue positions: the official party manifestos, the websites of parties and candidates, transcripts of parliamentary sessions, speeches and debates, and the political news coverage in newspapers and on television. Thus, the first part of the research project involves comparing issue positions of political parties since 1989 in three sorts of texts: (1) output of the dominant media outlets (newspaper and television reports, reflecting public debate in society about policy positions of political parties), (2) party manifestoes and web sites (in which they present their 'raw' unmediated policy positions), and (3) party representatives' contributions to parliamentary debates (reflecting the role of the parties in the state). Note that the content of the first two sub-corpora — the newspaper and TV data and the party manifestoes — have been coded by hand in previous research: the former have been coded for attitudes of actors towards one another and towards salient issues; the latter has been coded for issue saliency and ideological position. However the content of the third text type, the sub-corpus of parliamentary debates, has not been hand-coded. Yet, for this text we also have an external anchor since parties vote on issues after they have debated the pro's and con's. From this roll call behaviour, i.e. on tabulating votes cast by elected representatives of parties on specific issues, we will extract dimensions of party competition. Such analyses have been performed on roll call behaviour and often yielded quite stable and surprisingly solid results (Poole & Rosenthal 1991). However, the decision space of parties in parliament is much narrower than the space of possible political opinion and discourse (Potoski & Talbert 2000). The analysis of party manifestos and media content produces a much more varied and unconstrained space. Our analysis of various corpora of texts provides us with a clearer perspective of the structure of party competition. Given this focused analysis, an important challenge is to identify the major and most salient topics and dimensions of party competition, to determine how they develop over time and across multiple sources.

Direction, intensity and stability over time of issue positions

The major challenge of this study is to develop language resources and tools for identifying, extracting, attributing and aggregating the subjective issue position of political parties. The corpora of political texts include subjective (as opposed to factual) information in which parties express their perspective and opinions. From this highly "opinionated" textual data we need to device a method that goes beyond factual content, and can unearth relative issue positions of parties and politicians. An important step in extracting salient issues and dimensions is to explore corpora of text using the template information 'party X says Y about topic Z in source S at time T'. This template can be used to extract and standardized information on three dimensions of political position of actors: the direction (left-right, conservative-progressive, libertarian-authoritarian, green-grey etcetera), intensity (how radical, extreme or moderate are actors) and stability over time (how much do actors shift their position over time).

This part of the study involves comparing the party-specific use of language from the sub-corpus about which we know little, that of the parliamentary debates, with the two sub-corpora which have been coded in detail. These inner-corpus and cross-corpus comparisons could also allow for the assessment of variation in policy positions of the actors. In other words, this research could also reveal differing effects of media genre — how parties present different positions more or less saliently in different contexts (written and spoken news media, web sites, and parliamentary speeches and debates). Identifying and representing the major dimensions of party competition and the extent and manner in which the various salient issues "scale" towards these dimensions is a major goal of this part of the project.

In the field of Information Retrieval (IR) tremendous progress has been made over the past decades, yet most of the applications focused on retrieving entire documents. Yet, in the political domain, users are often interested in more specific, focused information than contained in full documents, for example more fine-grained information that is expressed at the level of words, sentences, sections or paragraphs. In order to extract policy positions from political texts, we will combine methods that use words as data (see Laver, Benoit & Garry 2003; Laver & Garry 2000) as well as methods that use the phrasal context in which the words occur.

Research objectives

The project will develop a semi-automated method to extract, from a focused search is a pre-determined corpus of relevant political texts (party platforms, mass media reports and parliamentary debates), all relevant issues and salient dimensions of political competition. First, an analysis will be performed on how the parties address political issues. How do they frame these issue? How salient is an issue for them (in terms of frequency of addressing it but also the intensity by which it is expressed)? How stable or variable is their stance over time?

We will improve the Electoral Compass party profiler, so that it can represent political choices in a more transparent manner based on personal beliefs of users (saliency and preferences per issue), ideologies of parties (saliency and issue positions per issue) and the structure of beliefs in the electorate as a whole (dimensionality of issue preferences, nature of the decision rule to aggregate distances between parties and the user per issue towards overall distances between parties and the user). Voters will differ in their need to examine explicit issues or the relevant parties to which they want to compare their own preferences. Thus, the main challenge is to present our findings on the position of political actors in a manner that makes sense and is useful to users, that allows them to analyse the positions of the actors and compare these with their own preferences. The end goal, then, is a non-biased, easy-to-use informational tool that allows citizens to compare the positions of political actors in an unbiased, rational and accessible manner.

Research Questions:

- How can salient actors (parties and candidates) be detected?
- How can salient issues be detected and ranked?
- How can relevant texts and text-snippets be detected by search tools from a larger corpus of texts?
- How can a party's position on a single issue be determined in terms of direction, intensity and consistency from texts?
- How can salient dimensions of party competition be extracted from these texts.
- How can issues be linked to these dimensions of party competition?
- How can we derive semi-automated mappings from texts to a multi-dimensional spectrum?
- How can we develop a tool that will support annotators in analyzing politically relevant texts in order to position parties on issues?
- How can we increase inter-annotator agreement, and thereby increase the reliability of the positioning of parties in the multi-dimensional political map?
- How can we link the lexical resources to the search tools to further improve the quality of the multi-dimensional mapping?

Data and methods

The party manifesto texts have been collected within the framework of the Party Manifesto Research Group (MRG), currently based at the Vrije Universiteit. It contains all party platforms of all (relevant) parties in democratic systems. The data on Dutch parties are complete, readily available and accessible in various formats. These manifestoes have been hand coded (sentence by sentence) into more than 60 categories of issues, based on the assumptions of thematic content analyses and saliency-theory: the idea that the ideological character of political parties can be established by the emphasis they put on various policy issues. By transforming manifestoes in this manner, quantitative analyses and comparisons can be performed on the data. These hand coded party platforms have also been used to position parties along various political dimensions, in particular the Left-Right and Progressive-Conservative (or GAL-TAN) dimensions. The second corpus of data – newspaper and TV-news texts for the period 1994-2008, largely from the Netherlands, but also from the UK, Germany and France – have been collected by the Department of Communication Science at the VU using the Amsterdam Content Analysis Toolkit (AMCAT) developed by Van Atteveldt (2008). For many content analysis projects, the texts have been hand-coded into core statements, each specifying the object, subject, issue and relation among these. This corpus has been collected under the assumptions of semantic network-analysis: political language is more than emphasis and salience as political actors position themselves on various issues, adopt positive or negative stances towards other actors or issues, attack opponents and form (policy) coalitions. Subsequently, all these actions are reported in the media, usually accompanied with evaluative comments, and with attributions of success and failure, that shape political momentum byportraying actors as winners and losers. The media-texts have been coded in such a manner that these semantic networks can be revealed – albeit that the manner in which this extraction can be done is open for debate. The third body of date are collected by Dutch parliament, consisting of all contributions made by (representatives of) parties during debates. For the UK, France and The Netherlands these are partly incorporated in the AMCAT database. These texts have not been hand-coded, but have been used in various types of automated content analysis based on co-occurrence (Van Noije 2007; Van Noije et.al. 2008). We do, however, have a small sample of voting behaviour of parties in the period 2002-2007 (collection will continue over the coming years). Thus, while the texts themselves have not been analysed, an element of positioning is present in this corpus. This research project will thus entail a strong quantitative data-analysis component, combined with qualitative research methods.

The preliminary time planning is as follows:

2008 - Collection and reviewing existing body of literature, literature study on the key concepts of this research project, in particular dimensions of party competition; - Elaboration of the research design and conceptualisation and operationalisation of theories/variables, data collection and organisation; - Construction and refinement of the theoretical framework by (re) formulation of central hypotheses of the theories used in this research; - Participation in and drafting a paper for a conference; - Institutionalising cooperation with the various disciplines/departments/research groups - Familiarising with various analytical tools and programmes - Involvement in coding team for EUProfiler
2009 - Further collection and organisation of data and analyses; - Analyses of (full body of) comparative data - Conference paper on the first empirical findings - Publication of an article in refereed journal - Cooperation in research team and analysis of EUProfiler - Analysis of EUProfiler data
2010 - In depth analyses of three corpora and the development of analyses of dimensionality of political conflict - Theoretical evaluation and innovation (finalisation and revision of theoretical framework), critical assessment of the validity of the various theoretical assumptions; - Publication in international refereed journal (based on empirical findings on dimensionality of political conflict); - Organising a conference and/or research group meeting in order to compare findings. - Integration into party profiling website tot test applicability
2011 (possibly integrated into the third year, with the edited volume being shifted to a later stage) - Finalisation of PhD; - Cooperation in edited volume on the basis of the conference

15 References

Aikhenvald, Alexandra Y. (2004) *Evidentiality*. Oxford: Oxford University Press.

Baker, Paul. 2006. *Using Corpora in Discourse Analysis*. London: Continuum.

Bara, Judith. 2001. "Tracking Estimates of Public Opinion and Party Policy Intentions in Britain and the USA." In *Estimating the Policy Positions of Political Actors*, Michael Laver, ed. London: Routledge, 217–36.

Barcelona, Antonio, ed. 2000. *Metaphor and Metonymy at the Crossroads.* Berlin and New York: Mouton de Gruyter.

Baron, J. 1988. *Thinking and Deciding*. Cambridge University Press.

Beineke, Philip, Trevor Hastie, & Shivakumar Vaithyanathan. 2004. The sentimental factor: Improving review classification via human provided information. In *Proceedings of ACL*, pages 263-270.

Billig, Michael. 1995. Discourse, Opinions and Ideologies: A Comment. *Current Issues in Language & Society* Vol. 2, No 2, 1995

Boogers, M. J. G. J. A. 2006. *Enquête bezoekers Stemwijzer*. Tilburg: Universiteit van Tilburg, onderzoeksnotitie Tilburgse school voor Politiek en Bestuur.

Budge, I., H-D. Klingemann, A. Volkens, J. Bara, & E. Tanenbaum, (2001), *Mapping Policy Preferences: Estimates for Parties, Electors, and Governments 1945-1998*. New York: Oxford University Press.

Charteris-Black, J. 2004. *Corpus Approaches to Critical Metaphor Analysis*, Basingstoke: Palgrave Macmillan.

Cienki, A. 1999. Metaphors and cultural models as profiles and bases. In Raymond W. Gibbs, Jr. and Gerard J. Steen, eds. *Metaphor in Cognitive Linguistics*. Amsterdam/ Philadelphia: John Benjamins, 189-203.

Cienki, Alan. 2005. Metaphor in the "Strict Father" and "Nurturant Parent" cognitive models: Theoretical issues raised in an empirical study. *Cognitive Linguistics* 16: 279-312.

Cienki, Alan. 2007. Frames, idealized cognitive models, domains. In Dirk Geeraerts & Hubert Cuyckens (Eds.). *The Oxford Handbook of Cognitive Linguistics*. Oxford: Oxford University Press, 170-187.

Cienki, Alan. 2008. The application of conceptual metaphor theory to political discourse: Methodological questions and some possible solutions. In T. Carver & J. Pikalo (Eds.), *Political Language and Metaphor*. Londen/New York: Routledge, 241-256.

Dalton, R.J. 2002. The Decline of Party Identifications. In Dalton, R. J. and Wattenberg, M. P. *Parties Without Partisans. Political Change in Advanced Industrial Democracies,* Oxford, Oxford University Press.

Dalton, R.J. & M. Gray. 2003. Expanding the electoral market place, in: B.E. Cain, R.J. Dalton, and S.E. Scarrow (eds) *Democracy transformed? Expanding political opportunities in advanced industrial democracies*. Oxford: Oxford University Press.

Dalton, R.J. & M. Gray.2003. Expanding the electoral market place, in: B.E. Cain, R.J. Dalton, and S.E. Scarrow (eds) *Democracy transformed? Expanding political opportunities in advanced industrial democracies*. Oxford: Oxford University Press.

Dalton, R. J. & Wattenberg, M. P. (2002) *Parties Without Partisans. Political Change in Advanced Industrial Democracies*, Oxford, Oxford University Press.

D'Andrade, Roy G. & Claudia Strauss, eds. 1992. *Human motives and cultural models.* Cambridge: Cambridge University Press.

Das, Sanjiv & Mike Chen. 2001. Yahoo! for Amazon: Extracting market sentiment from stock message boards. In *Proceedings of the Asia Pacific Finance Association Annual*

Conference (APFA).

Deignan, Alice. 2005. *Metaphor and Corpus Linguistics.* Amsterdam: John Benjamins.

De Smet, H. & J. Verstraete. 2006. Coming to Terms with Subjectivity. *Cognitive Linguistics* 17 -3, 365-392.

de Vries, Miranda, Daniela Giannetti, & Lucy Mansergh. 2001. "Estimating Policy Positions from the Computer Coding of Political Texts: Results from Italy, The Netherlands and Ireland." In *Estimating the Policy Positions of Political Actors*, Michael Laver, ed. London: Routledge, 193–216.

Esuli, Andrea & Fabrizio Sebastiani. 2005. Determining the semantic orientation of terms through gloss analysis. In *Proceedings of CIKM-05, 14th ACM International Conference on Information and Knowledge Management*, pages 617-624, Bremen, DE.

Esuli, Andrea. & Fabrizio Sebastiani. 2006. SENTIWORDNET: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06)*, pages 417-422, Genoa, IT.

Esuli, Andrea. & Fabrizio Sebastiani. 2007. SENTIWORDNET: A high-coverage lexical resource for opinion mining. Technical Report 2007-TR-02, Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche, Pisa, IT.

Faiclough, Norman. 1995. *Critical Discourse Analysis: The critical study of language.* London: Longman.

Fairclough, Norman. 2000. *New Labour, New Language.* London: Routledge

Feldman, S & Conover, P.J. 1983. Candidates, Issues and Voters: The Role of Inference in Political Perception, *The Journal of Politics*, Vol. 45, No. 4., pp. 810-839.

Fellbaum, Christiane, ed. 1998. *WordNet: An Electronic Lexical Database.* Cambridge, MA: The MIT Press.

Franklin, M. 1992. The decline of cleavage politics. In M. Franklin, T. Mackie & H. Valen (Eds) *Electoral Change. Responses to evolving social and attitudinal structures in Western countries*. Cambridge: Cambridge University Press.

Franklin, M. 2003. *Voter Turnout and the Dynamics of Electoral Competition*. Cambridge/New York: Cambridge University Press.

Frick, M. (2005) Parliaments in the digital age : exploring Latin America. E-democracy center, E-Working papers 2005/1, Geneva.

Gibson, Rachel K., Micheal Margolis, David Resnick & Stephen J. Ward (2003): "Election Campaigning on the WWW in the USA and UK". *Party Politics* 9:47-75.

Granberg, D. & Holmberg, S. (1988) *The Political System Matters: Social Psychology and Voting Behavior in Sweden and the United States*. Cambridge: Cambridge University Press.

Gries, S. & Stefanowitsch, A. 2006. *Corpora in cognitive linguistics.* Berlin: Mouton de Gruyter.

Hart, Christopher & Lukeš, Dominik. 2007. *Cognitive Linguistics in Critical Discourse Analysis: Application and Theory.* Cambridge: Cambridge Scholars Press

Hunston, Susan and Thompson, Geoff (Eds.) 2000. *Evaluation In Text: Authorial Stance And The Construction Of Discourse*. Oxford: Oxford University Press.

Hooghe M. and Teepe, W. 2007. Party Profiles on the web: an analysis of the logfiles of non-partisan interactive political internet sites in the 2003 and 2004 election campaigns in Belgium, *New Media Society*, Vol. 9, pp. 965-985.

Hundschnurscher, F. & J. Splett. 1982. *Semantik der Adjektive im Deutschen: Analyse der semantischen Relationen.* Westdeutscher Verlag.

Jurafsky, D. & Martin J. (2000) *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition.* Cambridge, MA: MIT Press.

Kamps, Jaap, Maarten Marx, Robert J. Mokken, & Maarten De Rijke. 2004. Using WordNet to measure semantic orientation of adjectives. In *Proceedings of LREC-04, 4th International Conference on Language Resources and Evaluation*, volume IV, pages 1115-1118, Lisbon, PT.

Kanayama, Hiroshi & Tetsuya Nasukawa. 2006. Fully automatic lexicon expansion for domain-oriented sentiment analysis. In *Proceedings of ENMLP*, pages 355-363.

Kim, S.-M. & E. Hovy. 2004. Determining the sentiment of opinions. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING'04)*, pages 1367-1373, Geneva, CH.

Kim, Soo-Min & Eduard Hovy. 2006. Automatic identification of pro and con reasons in online reviews. In *Proceedings of COLING/ACL Poster Sessions*, pages 483-490.

Kitschelt, H. 1997. European Party Systems: Continuity and Change. In M. Rhodes, P. Heywood, & V. Wright (eds.), *Developments in West European Politics*. Houndmills: Macmillan.

Kleinnijenhuis, J. et al. 2007a. *Nederland vijfstromenland: de rol van media en stemwijzers bij de verkiezingen in 2006* (The role of media and party profiling websites in the Dutch elections of 2006) Amsterdam: Bert Bakker.

Kleinnijenhuis, J., Van Hoof, A. M. J., Oegema, D. and De Ridder, J. A. 2007. "A test of rivaling hypotheses to explain news effects: news on issue positions of parties, real world developments, support and criticism, and success and failure." *Journal of Communication 57*(2), 366-384.

Kleinnijenhuis, J. & Krouwel, A. 2007. The nature and influence of party profiling websites. Paper presented at the annual meeting of Belgian and Dutch Political Scientists, Antwerpen.

Kleinnijenhuis, J., de Ridder, J. A. & Rietberg, E. M. 1997. Reasoning in economic discourse: an application of the network approach to the Dutch press. In C. W. Roberts (Ed.), *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts*. New York: Erlbaum, pp. 191-207.

Kleinnijenhuis, J. & Pennings, P. 2001. "Measurement of party positions on the basis of party programmes, media coverage and voter perceptions. In M. Laver, (Ed.), *Estimating the Policy Positions of Political Actors.* London: Routledge, pp. 162-182.

Klingemann, H-D, Hofferbert, R., Budge, I. & Keman, H. (1994) *Parties, Policies, and Democracy.* London: Westview Press.

Koppel, Moshe & Jonathan Schler. 2005. The importance of neutral examples for learning sentiment. In *Workshop on the Analysis of Informal and Formal Information Exchange during Negotiations (FINEXIN)*.

Kooi, Marjolein. 2007. Framing the Iran-UK sailor crisis: A corpus-driven approach to Critical Discourse Analysis of news articles of four news sites. Doctoraal thesis. Dept. of CIW, Faculteit der Letteren, VU, Amsterdam.

Krippendorff, K. (2004). *Content Analysis*. Thousand Oaks: Sage.

Krouwel, A. P. 1999. The catch-all party in Western Europe. A study in arrested development. PhD thesis, Vrije Universiteit, Amsterdam.

Krouwel, A. (forthcoming) *Party Transformations in Advanced European Democratic States*. Albany, NY: SUNY Press.

Kushal, Dave, Steve Lawrence, & David M. Pennock. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In Proceedings of WWW, pages 519-528.

Lakoff, G. 1993. The contemporary theory of metaphor. In A. Ortony (Ed.), *Metaphor and thought* (pp. 202–251). Cambridge: Cambridge University Press.

Lakoff, G. & Johnson, M. 1980. *Metaphors We Live By*, Chicago: University of Chicago Press.

Lakoff, G., & Johnson, M. 1999. *Philosophy in the Flesh: The Embodied Mind and its Challenge to Western Thought*, New York: Basic Books.

Langacker, Ronald W. 1985. Observations and speculations on subjectivity. In John Haiman (ed.) *Iconicity in Syntax.* Amsterdam Benjamins, 109-150.

Langacker, Ronald W. 1990. Subjectification. *Cognitive Linguistics* 1, 5-38.

Langacker, Ronald W. 1991 *Foundations of Cognitive Grammar*, vol. 2. *Descriptive Application.* Stanford University Press.

Lasswell, H. D., & Namenwirth, J.Z. (1969). *The Lasswell Value Dictionary*. New Haven: Yale University Press.

Laver, Michael J., Kenneth R. Benoit, & John Garry. 2003. "Extracting Policy Positions from Political Texts Using Words as Data."*American Political Science Review* 97 (2): 311.

Laver, M. & Budge, I. (1992) *Party policy and government coalitions.* London: St. Martin's Press / New York: Macmillan.

LeDuc, L., Niemi, R.G., Norris, P. (Eds). 2002. *Comparing Democracies 2. New Challenges in the study of elections and voting*. London, Sage Publications.

Liu, Hugo, Henry Lieberman, & Ted Selker. 2003. A model of textual affect sensing using real-world knowledge. In *Proceedings of Intelligent User Interfaces (IUI)*, pages 125-132.

Liu, Bing, Minqing Hu, & Junsheng Cheng. 2005. Opinion observer: Analyzing and comparing opinions on the web. In *Proceedings of WWW.*

Maks, I., Vossen, P., Segers, R., & van der Vliet, H. (2008), Adjectives in the Dutch semantic lexical database CORNETTO, LREC 2008, Marrakesh.

McCombs, M. (2004). *Setting the agenda: The mass media and public opinion.* Malden, MA: Blackwell.

Nelson, Mike. 2006. Semantic associations in Business English: A corpus-based analysis. *English for Specific Purposes* 25: 217-234.

Niles, I., & Pease, A. 2001. Towards a Standard Upper Ontology. In: *Proceedings of FOIS 2001*, Ogunquit, Maine, pp. 2-9.

Niles, I. & Terry, A. 2004. The MILO: A general-purpose, mid-level ontology. *Proceedings of the International Conference on Information and Knowledge Engineering*. Las Vegas, Nevada.

Osgood, C. E., Saporta, S., & Nunally, J. C. (1956). Evaluation assertion analysis. *Litera, 3*, 47-102.

Palmer, F. R. 1986. *Mood and Modality*. Cambridge: Cambridge University Press.

Pang, Bo, Lillian Lee, & Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of EMNLP*, pages 79-86.

Pang, Bo & Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of ACL-04, 42nd Meeting of the Association for Computational Linguistics*, pages 271-278, Barcelona, Spain.

Pang, Bo & Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of ACL-05, 43rd Meeting of the Association for Computational Linguistics*, pages 115-124, Ann Arbor, US.

Panther, Klaus-Uwe, and Günter Radden, eds. 1999. *Metonymy in language and thought.* Amsterdam/Philadelphia: John Benjamins.

Pollard, C. & I. A. Sag 1987. *Information-based Syntax and Semantics. Volume I: Fundamentals*. Stanford: CSLI.

Pollard, C. & I. A. Sag 1994. *Head-Driven Phrase Structure Grammar.*. Chicago: U Chicago

Press.

Poole, K.T. & Rosenthal, H. 1991. Patterns of congressional voting. *American Journal of Political Science, 35*(1): 228-278.

Popescu, Ana-Maria & Oren Etzioni. 2005. Extracting product features and opinions from reviews. In *Proceedings of HLT/EMNLP*.

Popkin, S. (1991) *The Reasoning Voter.* Chicago, University of Chicago Press.

Popping, R. (2000). *Computer-assisted text analysis*. London: Sage.

Potoski, M. & Talbert, J. 2000. The dimensional structure of policy outputs: Distributive policy and roll call voting. *Political Research Quarterly, 53*(4):695-710.

Praag, P. van (2007). 'De stemwijzer: hulpmiddel voor de kiezers of instrument van manipulatie?' Amsterdam: Lezing Amsterdamse Academische Club 24-05-2007.

Pragglejaz Group. 2007. "MIP: A method for identifying metaphorically used words in discourse." *Metaphor and Symbol* 22: 1-39.

Qu, Yan, James Shanahan, & Janyce Wiebe, editors. 2004. *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications.* AAAI Press. AAAI technical report SS-04-07.

Quirk, R., S. Greenbaum, G. Leech, and J. Svartvik. 1985. *A Comprehensive Grammar of the English Language*. New York: Longman.

Riloff, Ellen, Janyce Wiebe, & Theresa Wilson. 2003. Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of CONLL-03, 7th Conference on Natural Language Learning*, pages 25-32, Edmonton, CA.

Roberts, C. W. (1997). *Text analysis for the social sciences: methods for drawing statistical inferences from texts and transcripts*. Mahwah, NJ: Erlbaum.

Rodman, Lilita. 2001. You-Attitude: A Linguistic Perspective, *2001 Association for Business Communication, Quarterly*, Vol. 64, No. 4, 9-25.

Shanahan, James G., Yan Qu & Janyce Wiebe. *Computing Attitude And Affect in Text: Theory And Applications (The Information Retrieval Series)*. ISBN:1-4020-4026-1, Springer, 2006.

Takamura, Hiroya, Takashi Inui, & Manabu Okumura. 2005. Extracting emotional polarity of words using spin model. In *Proceedings of ACL-05, 43rd Annual Meeting of the Association for Computational Linguistics*, pages 133-140, Ann Arbor, US.

Takamura, Hiroya, Takashi Inui, & Manabu Okumura. 2006. Latent variable models for semantic orientations of phrases. In *Proceedings of EACL*, pages 201-208.

Takeshi S. Kobayakawa, Jin-Dong Kim and Jun'ichi Tsujii. 2008. An unsupervised method for extracting topic expressions from reviews on TV shows. In *Proceedings of EMOT 2008*, Marrakesh.

Talbert, J. C., & M. Potoski. 2002. Setting the legislative agenda: The dimensional structure of bill cosponsoring and floor voting. *Journal of Politics* 64(3):864–91.

Thomassen, J., Aarts, K. & van der Kolk, H. 2000. *Politieke veranderingen in Nederland 1971-1998: kiezers en de smalle marges van de politiek*, SDU Uitgeverij.

Tong, Richard M. 2001. An operational system for detecting and tracking opinions in on-line discussion. SIGIR Workshop on Operational Text Classification.

Traugott, E. 1995. Subjectification in grammaticalisation. In Stein, D. and S. Wright (eds.), *Subjectivity and Subjectivisation. Linguistic Perspectives.* Cambridge: Cambridge University Press, 31-54.

Traugott, E. and E. Konig. 2002. *Regularity in Semantic Change.* Cambridge: Cambridge University Press.

Turney Peter D., & Michael L. Littman. 2002. Unsupervised learning of semantic orientation from a hundred billion-word corpus. Technical report, National Research Council Canada.

Turney, Peter D. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of ACL-02, 40th Annual Meeting of the Association for Computational Linguistics*, pages 417-424, Philadelphia, US.

Turney, Peter D., & Michael L. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21(4):315-346.

Valitutti, Alessandro, Carlo Strapparava, & Oliviero Stock. 2004. Developing affective lexical resources. *PsychNology Journal*, 2(1):61-83.

van Atteveldt, W. (2008). Semantic Network Analysis: techniques for extracting, representing and querying media content. PhD dissertation. Amsterdam: Vrije Universiteit.

van Atteveldt, W., Kleinnijenhuis, J., Ruigrok, P. C., & Schlobach, S. (2008). Good news or bad news? Conducting sentiment analysis on Dutch text to distinghuish between positive and negative relations. *Journal of Information Technology & Politics, 5*(1), 1-22.

Van der Auwera, Johan & Ewa Schalley. 2004. From optative and subjunctive to irrealis. In Brisard, Frank et al. (Eds) *Seduction, community, speech: A Festschrift for Hermann Parret.* Amsterdam: John Benjamins, 87-96.

Van der Beek, L., Bouma, G., & van Noord, G. (2002). Een brede computationele grammatica voor het Nederlands. *Nederlandse Taalkunde, 7*(4), 353-374.

Van Dijk, Teun A. 1977. *Text and context: explorations in the semantics and pragmatics of discourse.* London: Longman.

Van Dijk, Teun. A. 1997. *Discourse as Social Interaction. Discourse Studies: A multidisciplinary introduction*. London: Sage

Van Noije, L. L. J. (2007). The democratic deficit closer to home: agenda building relations between Parliament and the Press, and the impact of European Integration, in the United Kingdom, the Netherlands and France. PhD dissertation. Amsterdam: Vrije Universiteit.

Van Noije, L. L. J., Kleinnijenhuis, J., & Oegema, D. (2008). Loss of parliamentary control due to mediatization and Europeanization: a longitudinal analysis of agenda-building in the UK and the Netherlands. *British Journal of Political Science, 38*(2).

Van Praag, P. 2007. *De stemwijzer: Hulpmiddel voor kiezers of instrument van manipulatie?* Amsterdam: Lezing Amsterdamse Academische Club, 24-5-2007.

Vegnaduzzo, Stefano. 2004. Acquisition of subjective adjectives with limited resources. In *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*, Stanford, US.

Walgrave, S., & van Aelst, P. 2005.. Much ado about (almost) nothing: Over de electorale effecten van Doe de Stemtest 2004. *Samenleving en Politiek*., 12, 61-72.

Webb, P. (2002) Political parties and democratic control in advanced industrial societies, in: P. Webb, D. Farrell & I. Holliday (eds.), *Political Parties in Advanced Industrial Democracies*. Oxford: Oxford University Press, 438-460.

Whitelaw, Casey, Navendu Garg, & Shlomo Argamon. 2005. Using appraisal groups for sentiment analysis. In *Proceedings of CIKM-05, 14th ACM International Conference on Information and Knowledge Management*, pages 625-631, Bremen, Germany.

Wiebe, J., Wilson, T., & Cardie, C. 2005. *Annotating Expressions of Opinions and Emotions in Language.* Dordrecht: Kluwer Academic Publishers.

Wiebe, J. & R. Mihalcea. 2006. Word sense and subjectivity. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics (ACL'06)*, pages 1065-1072, Sydney, AU.

Wilson, Theresa, Janyce Wiebe, & Rebecca Hwa. 2004. Just how mad are you? Finding strong and weak opinion clauses. In *Proceedings of AAAI-04, 21st Conference of the American Association for Artificial Intelligence*, pages 761-769, San Jose, US.

Wilson, Theresa, Janyce Wiebe, & Paul Hoffmann.2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of HLT/EMNLP*.

Wodak, Ruth & Paul Chilton (Eds) 2005. *A New Agenda in (Critical) Discourse Analysis: Theory, Methodology and Interdisciplinarity.* Amsterdam: Benjamins.

Wodak, Ruth & M. Krzyzanowski (Eds). 2008. *Qualitative Discourse Analysis in Social Sciences.* Palgrave MacMillan.

Yi, J. and W. Niblack. 2005. Sentiment mining in web-fountain. In *Proceedings of the 21st International Conference on Data Engineering*, pages 1073–1083.