

Semantic Network Analysis

Techniques for Extracting, Representing,
and Querying Media Content

Wouter van Atteveldt

Reading committee: prof.dr. Enrico Motta
dr. Gertjan van Noord
prof.dr. Guus Schreiber
prof.dr. Klaus Schönbach
prof.dr. Philip A. Schrodtt



© Wouter van Atteveldt 2008

You are allowed to copy and distribute this book in whole or in part and to make a derived work under the terms of the Creative Commons Attribution-Noncommercial 3.0 Netherlands License.

(<http://creativecommons.org/licenses/by-nc/3.0/nl/>)

An electronic version of this book is available from

<http://vanatteveldt.com/dissertation>.

This book can be purchased from <http://www.amazon.com/>.



SIKS Dissertation Series No. 2008-30

The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

Published by BookSurge Publishers, Charleston SC

ISBN: 1-4392-1136-1

VRIJE UNIVERSITEIT

Semantic Network Analysis
Techniques for Extracting, Representing, and Querying Media Content

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. L.M. Bouter,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de faculteit der Exacte Wetenschappen
op vrijdag 14 november 2008 om 13.45 uur
in de aula van de universiteit,
De Boelelaan 1105

door

Wouter Hendrik van Atteveldt

geboren te Texel

promotoren: prof.dr. F.A.H. van Harmelen
prof.dr. J. Kleinnijenhuis
copromotor: dr. K.S. Schlobach

Preface

Undoubtedly, this thesis contains many inaccuracies and omissions. One of the most glaring is the single author name on the cover: Although I doubt any Ph.D. thesis is really the sole work of its defender, especially an interdisciplinary work such as this thesis is the product of many hours of talking and collaborating. I started this thesis with little knowledge of both Communication Science and Knowledge Representation, and I am very grateful for the patient explanations and pointers from my supervisors and other colleagues.

I think I've been very lucky with my supervisors: I will not quickly forget the 5 AM e-mails from Jan when I was in Turkey working on one of our papers last year; or the hours spent in front of the whiteboard with Stefan while working on modal logic; or the innocent sounding questions Frank always asked to keep me on track both substantively and in terms of planning. Between the three of them, I think I've received incredible support both on the content of my work and on the procedure and planning needed to get me here. I especially appreciate the way how it was always stressed that, after everybody had their say, it was my thesis, and my responsibility to decide what I wanted to study, how I wanted to write it down, and when I wanted to finish it. This gave me the confidence to write down and defend this thesis with my name on it, even though the work is neither finished, nor perfect, nor solely mine.

According to Frank, the only thing worse than having no desk is having two desks, but I am very happy that I was a member of both the Communication Science and Knowledge Representation groups. The KR group (both on the AI and BI side of the invisible line) was and is a dynamic group with a lot of room for discussion and learning. Apart from

my supervisors, I especially appreciate the long talks, about work or otherwise, with Mark, Laura, Michel, and Willem. On the other side of the tram line, I was very lucky to start my PhD just after Rens and Lonneke started theirs, and I fondly remember the hours we spent looking at data and models together. Although Dirk wisely turned down my request to become co-promotor he was always there to talk about work but especially about non-work. He taught me to think about why I do the things I do, and to concentrate on doing the things that really matter. Anita showed a surprising side to her character when she was stuck in Amsterdam during a storm and we drank champagne for her birthday and played Catan until 2 AM. I am also very happy we stole back Janet from the UvA for the Contested Democracy project; we can always use more Klaverjas players to join us to the ICA conferences. I really look forward to continuing my collaboration and friendship with all these colleagues.

I would also like to use the opportunity to thank my professors from the University College Utrecht and University of Edinburgh, especially Maarten Prak, who showed me how university education should be done, Mark Steedman and Greame Ritchie, who got me hooked on natural languages, and of course Miles Osborne, Claire Grover, Bonnie Webber and all the others.

The ACLU Lawyer Clarence Darrow once said that the first half of our life is ruined by our parents, and the second half by our children. In contrast, I feel that my childhood has both been very pleasant and helped me to value thinking, knowledge, and discussing. I have always felt my parents to support me in whatever I did, and I believe that I am very lucky to have had such a wonderful family. Bas, I sometimes miss the early days of 2AT, exploring the game called running a company, and I am glad that things are going so well with the company. I am also thrilled that we finally got the catamaran working, and I hope we will have a very windy summer next year. Nienke, I am looking forward to our first publication together, using text boxes and arrows between them to explain the human condition, and I look forward to more roof terrace parties if you still find Amsterdam liveable after your stay in New York.

If you compare the above list with my list of co-authors, one name is conspicuously lacking. I would probably end up on the couch if I would list Nel among my colleagues and collaborators, even though we did spend a lot of time working, discussing, and writing together. However, that is completely insignificant compared to her contribution to my real life. Since meeting her I've learned invaluable lessons on people, emotions, and insecurity, and I feel that I've become a much better person over the last four years, or at least a better dressed person.

Contents

| | | |
|----------|---|-----------|
| 1 | Introduction | 1 |
| 1.1 | Introduction | 2 |
| 1.2 | Research Question | 6 |
| 1.3 | Domain and Data | 7 |
| 1.4 | Contributions | 8 |
| 1.5 | Thesis Outline | 9 |
| | | |
| I | Background | 11 |
| | | |
| 2 | Content Analysis | 13 |
| 2.1 | Introduction | 14 |
| 2.2 | Content Analysis in Communication Science | 18 |
| 2.3 | Semantic Network Analysis | 23 |
| 2.4 | The NET method | 29 |
| 2.5 | The 2006 Dutch parliamentary elections | 33 |
| 2.6 | Computer Content Analysis | 35 |
| 2.7 | Conclusion | 39 |
| | | |
| 3 | Natural Language Processing | 41 |
| 3.1 | Introduction | 42 |
| 3.2 | The Preprocessing Pipeline | 43 |
| 3.3 | Thesauri | 48 |
| 3.4 | Evaluation metrics | 49 |
| 3.5 | Conclusion | 50 |

| | | |
|------------|---|------------|
| 4 | Knowledge Representation and the Semantic Web | 51 |
| 4.1 | Introduction | 52 |
| 4.2 | The Semantic Web | 54 |
| 4.3 | The Semantic Web as a Knowledge Representation frame- work | 60 |
| 4.4 | Conclusion | 61 |
| II | Extracting Semantic Networks | 63 |
| 5 | Extracting Associative Frames using Co-occurrence | 65 |
| 5.1 | Introduction | 66 |
| 5.2 | Frames as Associations | 68 |
| 5.3 | A Probabilistic Model of Associative Framing | 71 |
| 5.4 | Use Case: Terrorism in the News | 75 |
| 5.5 | Conclusion | 88 |
| 6 | Using Syntax to find Semantic Source, Subject, Object | 91 |
| 6.1 | Introduction | 92 |
| 6.2 | Determining Semantic Roles using Syntax Patterns | 93 |
| 6.3 | Determining Validity | 99 |
| 6.4 | Results | 105 |
| 6.5 | Error Components Analysis | 111 |
| 6.6 | Discussion / Conclusion | 114 |
| 7 | Determining valence using Sentiment Analysis | 117 |
| 7.1 | Introduction | 118 |
| 7.2 | Polarity in Political Communication | 120 |
| 7.3 | Task: Classifying NET relations | 121 |
| 7.4 | Sentiment Analysis | 122 |
| 7.5 | Method | 124 |
| 7.6 | Results | 130 |
| 7.7 | Validation | 135 |
| 7.8 | Conclusion | 141 |
| III | Reasoning with Media Data | 143 |
| 8 | Using RDF to store Semantic Network Data | 145 |
| 8.1 | Introduction | 146 |
| 8.2 | Representing Media Data: Statements about Statements | 147 |
| 8.3 | Representing Political Background Knowledge | 152 |
| 8.4 | A political ontology | 156 |
| 8.5 | Using OWL for a richer ontology | 162 |

| | | |
|-----------|---|------------|
| 8.6 | Conclusions | 164 |
| 9 | Querying, Analysing, and Visualising Semantic Network Data | 165 |
| 9.1 | Introduction | 166 |
| 9.2 | Querying the Semantic Network | 167 |
| 9.3 | Searching the News | 168 |
| 9.4 | Using the system: Parties in the news | 171 |
| 9.5 | Conclusion | 175 |
| | | |
| IV | System Description | 177 |
| | | |
| 10 | The AmCAT Infrastructure | 179 |
| 10.1 | Introduction | 180 |
| 10.2 | The AmCAT Navigator and Database | 181 |
| 10.3 | The iNet Coding Program | 192 |
| 10.4 | Conclusion | 202 |
| | | |
| 11 | Discussion and Conclusion | 203 |
| | | |
| | Bibliography | 215 |
| | | |
| | Samenvatting (Dutch Summary) | 231 |

CHAPTER 1

Introduction

'No-campaign Wilders is a circus for bodyguards and media'

(Wilders' tourNEE is circus voor lijfwachten en media; de Volkskrant, May 17, 2005)

'Approach in media increases cynicism of citizens'

(Aanpak in media wakkert cynisme van burgers aan; Trouw, June 12, 1999)

'Against populism; media more dangerous than Muslims'

(Tegen het populisme; media gevaarlijker dan moslims; NRC Handelsblad, March 17, 2007)

1.1 Introduction

Does newspaper coverage of immigration topics increase polarisation? Is the incessant reporting of opinion polls during campaigns a self-fulfilling prophecy? Are rightist leaders portrayed as extremists and racists? Does negative and strategic coverage of politics lead to cynical voters? How does the pattern of conflict between parties influence voter choice?

These questions have three things in common. First, they are all questions that are highly relevant to the current debates in our society. Second, in order to answer these questions, it is necessary to measure specific aspects of media coverage such as attention for issues, the evaluation of politicians by the media and other politicians, and the tone and framing of messages. Third, they will not be answered in this thesis. Rather, this thesis will describe a number of methods and techniques that enable social scientists to answer these and similar questions.

Let us look more closely at these commonalities. The questions posed above are currently relevant to society, and similar questions have been asked for a long time. As early as 1893, Speed conducted an analysis of the shift from 'serious news' to sports and scandals in New York newspapers (quoted by Krippendorff, 2004, p.55). More recently, the media were accused of: blindly accepting White House assertions regarding Weapons of Mass Destruction in Iraq; the demonisation of Pim Fortuyn before his assassination; creating a platform for right-wing extremists by devoting too much coverage to their provocations; and acting as judge and jury by 'solving' the disappearance of Natalee Holloway on prime-time television using investigation methods that the police are prohibited from using. In general, there is strong public and scientific interest in the functioning of the media and their effects on the audience and democracy, as proved by the creation of a research theme 'Contested Democracy' by the Dutch Science Foundation (NWO, 2006). Most political news and almost all foreign news reaches citizens exclusively through the media. Following Lippmann's argument that "the way in which the world is imagined determines at any particular moment what men will do," (1922,p.6), this means that the media are vital in determining how citizens view the world and hence how they act.

In order to investigate these interactions between media, politics, and public, a communication scientist has to be able to systematically describe the relevant aspects of the media. This emphatically does not mean quantifying everything that happens in the media. In communication science, the purpose of analysing the content of (media) messages is to learn more about the interaction of those messages with their social context: what did the sender mean with the message, and why did he

send it? How did the receiver interpret the message, and what effects will it have on his opinions or behaviour? Such substantive research questions on the relation between text and social context guide and focus the measurement in a top-down fashion.

As stated in the third commonality, this thesis is about enabling media research rather than executing it. Referred to in the social sciences as a methodological study, the focus of this thesis is on investigating, creating, and validating measurement tools and methods, rather than testing hypotheses. Since validation is an important part of developing a useful method, where possible techniques will be tested in terms of accuracy or reliability by comparing the results of the automatic extraction with manual codings. Validity is tested by conducting or reproducing substantive analyses using automatic Content Analysis. The ultimate test for each technique is whether its output and performance are sufficient to (help) answer the communication scientific question.

In the social sciences, Content Analysis is the general name for the methodology and techniques to analyse the content of (media) messages (Holsti, 1969; Krippendorff, 2004). As will be described in chapter 2, the purpose of Content Analysis is to determine the value of one or more theoretically interesting variables based on message content. The word ‘message,’ here, is broadly defined, including newspaper articles, parliamentary debates, forum postings, television programs, propaganda leaflets, and personal e-mails. Messages, being sets of symbols, only have a meaning within the context of their use by their sender or receiver. Hence, the purpose of Content Analysis is to infer relevant aspects of what a message means in its context, where the communication research question determines both the relevance and the correct context.

Most current Content Analysis is conducted by using human coders to classify a number of documents following a pre-defined categorisation scheme. This technique is called Thematic Content Analysis, and has been used successfully in a large number of studies. Unfortunately, it has a number of drawbacks. One obvious drawback is the fact that human coding is expensive, and human coders need to be extensively trained to achieve reliable coding. A second problem is that the classification scheme generally closely matches the concepts used in the research question. This means that the Content Analysis is more or less *ad hoc*, making the data gathered in these analyses unsuitable for answering other research questions or for refining the research question. Also, it is often difficult to combine multiple data sets for forming one large data set due to differences in operationalisation.

An alternative Content Analysis method is called Semantic Network Analysis or Relational Content Analysis (Krippendorff, 2004; Roberts,

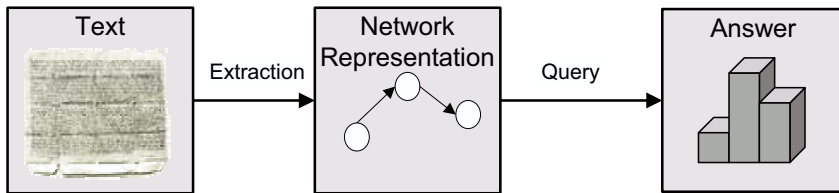


Figure 1.1: Semantic Network Analysis

1997; Popping, 2000). Figure 1.1 is a very schematic representation of Semantic Network Analysis. Rather than directly coding the messages to answer the research question, Semantic Network Analysis first represents the content of the messages as a network of objects, for example the network of positive and negative relations between politicians. This network representation is then queried to answer the research question. This querying can use the concrete objects as extracted from the messages, or aggregate these objects to more abstract actors and issue categories, and query the resulting high-level network. For example, for determining the criticism of coalition parties by opposition parties, one would aggregate all political actors to their respective parties, and then query the relations between the parties that belong to the opposition and those that belong to the coalition. This separation of extraction and querying means that the network representation extracted for one study can be used to answer different research questions, as long as the actors, issues, and relations needed to answer these questions are present in the extracted network. This solves the problem of the tight coupling between research question and measurement present in thematic Content Analysis.

However, Semantic Network Analysis does not solve the problem of expensive human coding: Extracting the network of relations is probably even more difficult than categorising text fragments. Although human coding of text using Semantic Network Analysis is possible and has yielded good results, it is expensive and error-prone due to the complexity of the coding. An advantage is that because the abstract concepts used in the research question are decoupled from the objects to be measured, these objects can be closer to the text than in thematic Content Analysis, thereby narrowing the semantic gap between words and meaning. This should make it easier to automate the coding process in a generalisable way. Nonetheless, a lot of automatic Semantic Network Analysis is currently restricted to extracting and analysing co-occurrence networks of words (e.g. Diesner and Carley, 2004; Corman et al., 2002); a notable exception is the work by Philip Schrodtt and colleagues on ex-

tracting and analysing patterns of conflict and cooperation between international actors, although that is limited to specific relations between specific actors (Schrodt, 2001; Schrodt et al., 2005).

Due to the decoupling of extraction and querying, data obtained from Semantic Network Analysis lends itself to being combined, shared, and used flexibly. Unfortunately, there are no standard ways to clearly define the meaning of the nodes in a network and how they relate to the more abstract concepts used in the research question. This makes it difficult to reuse or combine data in practice, because the vocabulary used in the network needs to be changed and aligned manually to make sure the networks are compatible with each other and with the new research question. Moreover, there is no standard to define patterns on these networks — the concepts we need to measure to answer the research question — in such a way that other scientists can easily understand, use, and refine these patterns. These problems prevent scaling Semantic Network Analysis from the use in single or related studies to creating large archives of analysed media material.

This thesis describes a number of techniques to overcome the limitations described above, and shows how these techniques are immediately useful for communication research. In particular, this thesis investigates using techniques from two fields of Artificial Intelligence: Computational Linguistics, and Knowledge Representation.

As described in chapter 3, Computational Linguistics has seen drastic increases in computer storage and processing power in recent decades, leading to the development of many linguistic tools and techniques. Examples of such tools are robust syntactic parsers for English and Dutch; the availability of functioning Anaphora Resolution systems; and the use of statistical and corpus techniques to improve the analysis of subjectivity in Sentiment Analysis. Although these techniques will not answer social scientific research questions by themselves, they allow us to start with linguistically analysed text rather than with raw text, allowing us to concentrate more on the semantics and less on the surface patterns of language. Content Analysis and linguistic analysis should be seen as complementary rather than competing: linguists are interested in unravelling the structure and meaning of language, and Content Analysts are interested in answering social science questions, possibly using the structure and meaning exposed by linguists.

In order to alleviate the problems of combining, sharing, and querying the Semantic Networks, we turn to the field of Knowledge Representation. As will be described in chapter 4, Knowledge Representation deals with the formal representation of information such as the background knowledge used for aggregating the concrete textual ob-

jects to the abstract concepts used in a research question: For example, which politicians belong to which party and which issues are contained in which issue category. A barrier to sharing data is that the aggregation step from concrete, textual indicator ('Bush') to theoretical concepts (U.S. president) is often made implicitly, either by the coder or by the researcher. Knowledge Representation allows us to formally represent the background knowledge needed for this step. This makes it easier to combine and share data, since it is clear what a concept means, and heterogeneous data can be combined by aligning the involved ontologies. Moreover, the data can be used more flexibly since different research questions can be answered using the same data set by using different ontologies or aggregating on different aspects of the same ontology. By formalising both the ontology of background knowledge and the extracted media data into a combined Semantic Network, it is possible to define the concepts from the research question as (complex) patterns on this network, making all steps from concrete objects to abstract concepts to answers to the research question transparent and explicit.

Roberts (1997, p.147) called for Content Analysis to be something 'other than counting words.' In general, Content Analysis can go beyond manually counting words in three ways: (1) Extracting abstract theoretical concepts rather than word frequencies (2) Extracting structures of relationships between concepts rather than atomic concepts; and (3) Using the computers to conduct the extraction rather than relying on human coders. Often, a step forward in one field is combined with a step backwards in another: Studies such as Roberts (1997) or Valkenburg et al. (1999) identify complex and abstract concepts, but do so manually rather than automatically. Carley (1997) and Schrodtt (2001) automatically extract networks, but limit themselves to relations between literal words and concrete actors. In this thesis, it is argued that we can and should move forward in all three ways simultaneously. This is accomplished by separating the *extraction phase*, where relations between concrete objects are (automatically or manually) extracted from a message; and a *construction phase*, where the complex and abstract variables used in communication science theory are constructed based on the extracted relations.

1.2 Research Question

This thesis investigates whether it is possible to utilise techniques from Natural Language Processing and Knowledge Representation to improve two aspects of Semantic Network Analysis: extracting Semantic Networks from text; and representing and querying these networks to an-

swer social science research questions. In terms of extraction, it looks at automating the recognition of concepts, and the identification and classification of the semantic relations between these concepts from text. This part draws on techniques from Computational Linguistics, such as anaphora resolution, grammatical analysis, and sentiment analysis. This leads to the following research questions:

RQ1 Can we automate the extraction of Semantic Networks from text in a way that is useful for Social Science?

RQ1a Can we recognise the occurrence of specific actors and issues in text in a way that is useful for answering social science research questions?

RQ1b Can we automatically determine the semantic relations between these actors and issues?

RQ1c Can we automatically determine the valence (positive, negative) of these semantic relations?

The second aspect is the representation and querying of media data and background knowledge. The goal of this representation is to make it easier to combine and analyse media data, by formalising the link between the concrete objects in the extracted networks and the abstract objects used in the research question. The goal of the querying is to make it easier to answer research questions by defining patterns or queries on top of the combined network of media data and background knowledge. This part uses techniques from Knowledge Representation. The second research question reads as follows:

RQ2 Can we represent Semantic Network data in a formal manner and query that representation to obtain the information needed for answering Social Science research questions?

RQ2a Can we formally represent Semantic Network data and background knowledge connected with that data in a way that allows the reuse and combination of data for different social science research questions?

RQ2b Can we query these represented networks in a flexible way to answer different social science research questions?

1.3 Domain and Data

Each substantive chapter is based on a different data set and uses a different methodology. A common denominator across all chapters, however,

is that they are based on the data obtained from earlier Semantic Network Analysis performed on Dutch political newspaper articles. Theoretically, none of the techniques presented in this thesis are specific to a single language or medium, and they can and have been used on parliamentary debates, survey questions, and television broadcasts. Practically, however, this choice does have a profound impact.

The choice for focusing on newspaper data is mainly pragmatic: the existing annotated corpus is mainly derived from newspapers; newspaper articles are grammatically correct and written according to fairly strict style rules; and analysing text is easier to automate than analysing images and sound. This makes creating analysis tools easier, but also makes them more useful, as the more available raw material there is, the more useful a tool for automatic analysis of this material is. Moreover, the tools and techniques presented here can be reconfigured and retrained to work on different genres, such as debates or television transcripts.

The choice for investigating the Dutch language is also pragmatic: the corpus of existing material analysed using Semantic Network Analysis consists almost exclusively of Dutch newspaper articles. It is not an indefensible choice, however, as Dutch has traditionally received quite a lot of attention from linguists, and many tools such as thesauri, Part-of-Speech taggers, and parsers are available for Dutch.¹ Presumably, overall performance for English would have been higher due to the better quality of linguistic tools available, but this would also mean that it is more difficult to assess how well the same techniques would have performed on languages that have received less attention from the linguistic community. This is especially relevant for political research, as the English language is only native to a handful of countries, many of which share a number of political features such as a two-party system. For internationally comparative political communication research, it is generally insufficient to include only English-language material. If Natural Language Processing techniques can be successfully used to analyse the content of Dutch text, it can almost certainly be used for English text, and probably for other languages such as French or German.

1.4 Contributions

The techniques presented in this thesis take a step towards solving two problems related to Semantic Network Analysis: First, it leverages recent advances in Computational Linguistics to expand the possibilities of extracting Semantic Networks automatically. Specifically, it uses the syn-

¹See section 3.2 on page 43

tactic analysis of sentences to distinguish between the source, agent, and patient of a proposition. Additionally, it uses techniques from Sentiment Analysis to determine whether the proposition is positive or negative. These techniques are validated by comparing the extracted Networks to those manually extracted, showing that they are immediately useful for social scientific analysis.

Second, it uses techniques from the field of Knowledge Representation techniques called the Semantic Web to represent the Semantic Networks. By formally representing both the relations between the concrete objects expressed in the message and their relation with the more abstract concepts used in social science research questions, it facilitates the reuse, sharing, and combination of Semantic Network data. Moreover, by operationalising the social science research question as a formal query over the Semantic Networks, it makes it easier to analyse these networks and to publish, adapt, criticise, and refine operationalisations.

Taken together, these advances represent an important step forward for Semantic Network Analysis. The techniques presented here potentially allow the combination of Semantic Network data from different research groups, dealing with different countries, different media, and different time periods. These Networks can be extracted automatically, if the source material and desired accuracy permit, or manually, or a combination of the two. Moreover, these data sets can be shared and combined to create large heterogeneous data sets that can be queried to answer various research questions. Such data sets can provide a strong stimulus for communication research, as they allow large international and/or longitudinal studies without incurring the enormous costs of gathering the needed data. Moreover, analysing the same data from different theoretical perspectives or operationalisations can give more insight into the actual social processes than individual studies, as differences in findings can not be caused by artifacts in the data or unconscious differences in the extraction.

1.5 Thesis Outline

The organisation of this thesis closely follows the research questions. Part I will provide some background knowledge on the fields of Content Analysis (chapter 2), Natural Language Processing (chapter 3), and Knowledge Representation (chapter 4). These chapters are meant for readers who are not proficient in these fields and can be safely passed over by others, with the possible exception of section 2.3, in which Semantic Network Analysis is defined.

Part II will provide the answer to the first research question on au-

tomatically extracting Semantic Networks from text. Each chapter answers one of the specific questions defined above: Chapter 5 will discuss extracting and interpreting the (co-)occurrence of actors and issues. Chapter 6 will describe a way of using syntactic analysis to extract semantic relations from text. Chapter 7 describes determining the valence of relations using Machine Learning techniques.

Part III answers the second research question on representing and querying the extracted Semantic Network. Chapter 8 describes the possibilities and limitations of using formalisms from the field Knowledge Representation called the Semantic Web to store Semantic Network data. Chapter 9 shows how this representation can be used to extract the information needed for answering social scientific research questions by discussing a number of use cases and showing the query needed to answer these questions.

In the last substantive part, chapter 10 provides an overview of the AmCAT system and infrastructure that has been developed to use the techniques described in the previous parts to conduct Semantic Analysis and store, combine, and query the results.

Part I

Background

CHAPTER 2

Content Analysis

'Frequently the messages have meaning; that is, they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem.'

(Claude Shannon, A Mathematical Theory of Communication (1948, p.379))

The topic of this thesis is the extraction, representation, and querying of textual content. In the social sciences, Content Analysis is the name given to the techniques used to systematically study the content of messages. Hence, Content Analysis methodology is an important aspect of this thesis. This chapter gives a brief description of what Content Analysis is and how it is used in communication science. Moreover, it presents Semantic Network Analysis as a Content Analysis method and explains the NET Semantic Network method.

2.1 Introduction

This thesis investigates whether it is possible to enhance Content Analysis using techniques from Natural Language Processing and Knowledge Representation. Part I gives the background knowledge that is assumed in the descriptions of the techniques in parts II and III. Chapters 3 and 4 provide information on Natural Language Processing and Knowledge Representation, respectively, while the present chapter describes Content Analysis. This section will review some of the definitions of Content Analysis. Section 2.2 shows some studies that have been performed to give an overview of the questions Content Analysis is used to answer and how these questions are answered. Section 2.3 discusses some of the problems with Content Analysis, and how Semantic Network Analysis can alleviate these problems. A description of the NET method, the Semantic Network Analysis method used in this thesis, is given in section 2.4. Section 2.5 gives a very brief description of the 2006 Dutch election campaign coverage that is used as a case study in chapters 6 – 9. The final section reviews some of the existing computer programs for conducting Content Analysis.

Communication science is the field of social science dedicated to studying communication. Simply put, communication is the sending of a message by an actor and the receiving and interpreting of that message by another actor. Information Theory gives a mathematical definition of communication as a source coding a message as a set of symbols, which are transferred through a (noisy) channel, and received and decoded by the receiver (Shannon and Weaver, 1948). While Information Theory is interested in the mathematical aspects of message coding and decoding, communication science is interested in the interactions between the message and its social context: Why did the source send this message? What does the message tell us about the society in which it was produced? How does the receiver interpret and process the message? What influence does the message have on the receiver? Lasswell summarised this in his famous quote as “who says what to whom, how, and with what effect?” (1948, p.37). Holsti (1969, p.24) added ‘Why?’ to this question to stress the validity of investigating the mechanisms that cause a source to send a certain message. Figure 2.1 visualises these questions in the communication channel. It should be noted that the view of communication as a flow of information is not undisputed: Carey (1989) calls this the ‘transmission view’ of communication and argues that the ‘ritual view’, communication as a community-building shared experience, should not be neglected. However, studying the ritual view of communication also requires understanding the content of the exchanged messages, even if

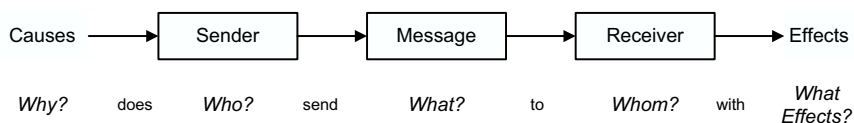


Figure 2.1: The Communication Channel

the focus is on the symbolism and cultural or ritualistic meaning rather than the transmitted information.

In order to make inferences about the social context of a message, we need to be able to measure the relevant aspects of the message. Content Analysis is the name for the methods and techniques used to make these measurements. Since Content Analysis methodology is the central topic of this thesis, we shall take some time to consider the definition of Content Analysis and its relation to communication science and other fields. Figure 2.2 gives a number of definitions from various authors. Although these definitions differ in a number of specifics, they all agree on two requirements for Content Analysis as a scientific tool: validity and relevance.

Validity Holsti (1969), Neuendorf (2002), and Berelson (1952) all explicitly require Content Analysis to be objective and systematic, meaning that every step has to be conducted following explicit rules that are clear enough to avoid subjective interpretation and selection bias. Krippendorff (2004) uses the standard methodological terms replicable and valid, arguing that these terms are more well-defined and imply objectivity and systematicness: it is impossible to conduct a

Content Analysis is any technique for making inferences by objectively and systematically identifying specified characteristics of messages (Holsti 1969, p.14; Stone et al. 1966, p.5)

Content Analysis may be briefly defined as the systematic, objective, quantitative analysis of message characteristics (Neuendorf, 2002, p. 1)

Content Analysis is a research technique for making replicable and valid inferences from texts (or other meaningful matter) to the contexts of their use (Krippendorff, 2004, p.18)

Content Analysis is a research technique for the objective, systematic, and quantitative description of the manifest content of communication (Berelson, 1952, p.18)

Content Analysis [is] any methodological measurement applied to text for social science purposes [...] any systematic reduction of a flow of text [...] to a standard set of statistically manipulable symbols representing the presence, the intensity, or the frequency of some characteristics relevant to social science (Markoff et al., 1975, p.5)

Figure 2.2: Definitions of Content Analysis

replicable analysis without using systematic and well-defined rules, and an analysis that allows for biased and subjective interpretation cannot be valid using the normal definition of that term in research methodology. All these authors agree that Content Analysis needs to adhere to the rules of scientific measurement, “making a claim to participation in a scientific enterprise [rather than] a literary or journalistic enterprise” (Markoff et al., 1975, p.7).

Theoretical Relevance Content Analysis deals with extracting certain characteristics of the message that are “theoretically relevant” (Holsti, 1969, p.5). As stated above, communication science is interested in the interaction between a message and its (social) context, so Content Analysis is employed to measure the aspects of the message that are relevant to those interactions, based on theory or hypotheses regarding the context. The fact that a word occurs with a certain frequency is not by itself interesting to the content analyst, unless it is seen as an indicator of a relevant theoretical construct. This top-down approach, where the theory regarding the message context determines the aspects of a message that are interesting, is what sets Content Analysis apart from most linguistic analysis, which is primarily interested in the structure and content of the message itself.

An important argument made by Holsti (1969) and Krippendorff (2004) is that Content Analysis is a tool for making inferences about the message context rather than just measuring aspects of the message content. As Krippendorff argues, a message by itself does not have meaning: it is a set of symbols. The meaning or semantics of a message is the connection between these symbols and the things they refer to (their denotation). Since each receiver or sender of a message can interpret the message differently, it is important to realize that a message only has a meaning in the “context of its use” (Krippendorff, 2004, p.33). As the communication research question determines which aspects of a message are interesting, it also defines the context in which the message is to be interpreted. The task of Content Analysis is to infer the relevant meaning in that context from the symbols in the message.

This is illustrated in figure 2.3, a combination of two figures from Krippendorff (2004). The oval on the left-hand side represents the *context* of the message as conceptualised by the researcher. The texts are located in this content to stress the fact that the texts are read within that context. The research question with its possible answers is located in the top of this oval. Content Analysis assumes that there is some sort of stable correlation between the features of the message and the possible answers to the research question. For example, the correlation could be as simple as assuming that the frequency with which the word Bush is used is

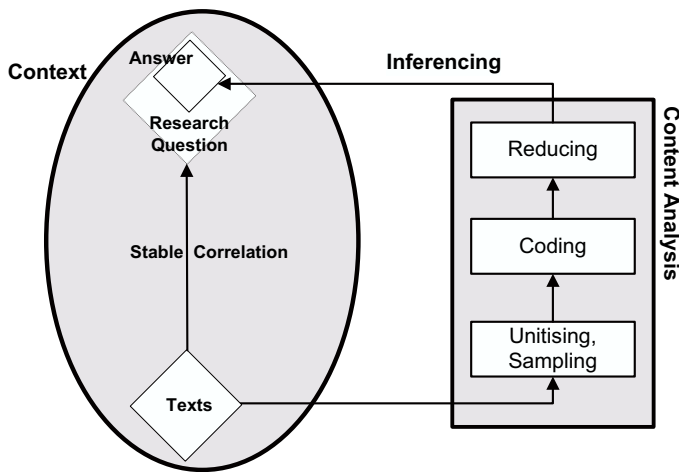


Figure 2.3: Krippendorff's Framework and Components of Content Analysis

(Combined and edited from Krippendorff 2004, figures 2.1 (p.30) and 4.2 (p.86))

correlated with the visibility of the politician Bush in those texts. This correlation guides the Content Analysis, represented by the rectangle on the right-hand side. The first step in Content Analysis is the unitising and sampling of the messages. The resulting units of measurement are coded manually or automatically, and combined or *reduced* by aggregation or statistical techniques to form the units of analysis. Finally, the answer to the research question is inferred from the results at the level of analysis.

In addition to the similarities noted above, there are also a number of differences between the various definitions of Content Analysis, especially between quantitative and qualitative analysis, and between strictly using the manifest content of messages or also using the latent content. A full discussion of these differences is beyond the scope of this thesis, so we will restrict ourselves to briefly describing the positions in these debates.

Quantitative vs. Qualitative The first question is whether Content Analysis should be restricted to *quantitative* analysis, assigning numbers to predefined measures, or whether it can be used *qualitatively*, by creating the units and categories while reading it in what Krippendorff (2004, p.303) calls an *interactive-hermeneutic* process. In this process, the meaning of the text as conceived by the researcher is iteratively refined or constructed in a bottom-up fashion, rather than

using predefined categories. Berelson (1952) explicitly opts for the former by including the term quantitative in his definition, and Neuen-dorf (2002, p.14) calls qualitative analysis infeasible. Krippendorff (2004, p.87–89), on the other hand, argues that qualitative analysis can be conducted validly and repeatably, and stresses the importance of triangulation.

Manifest vs. Latent The second debate is whether only the *manifest* or literal content of messages can be used or whether the (presumably human) coder is allowed to use his or her interpretation of the *latent* content. Here again, Berelson (1952) explicitly uses the word manifest in his definition, while Krippendorff (2004) focuses on the ability of human readers to ‘read between the lines’ in an intersubjective way. More importantly, he argues that the manifest–latent dichotomy assumes that messages contain a manifest meaning which can be extracted, while he thinks meaning only exists within the context of a message, as discussed above.

2.2 Content Analysis in Communication Science

In order to get a feeling for the use of Content Analysis in communication science, this section reviews a number of influential theories in communication science, explicitly describing what information was needed from the analysed content and how that information was obtained. This section will start with research focussed on simple aspects of messages — the visibility of issues or evaluative tone of the language — and move towards studies that combine these approaches, studying both issues and actors and the salience and evaluative tone of their descriptions and of relations between them.

2.2.1 *Agenda Setting and Agenda Building*

A successful line of research within communication science is that of Agenda Setting. The intuition behind it is that “the press may not be successful much of the time in telling people what to think, but it is stunningly successful in telling people what to think about” (Cohen, 1963, p.13). This intuition was empirically tested by McCombs and Shaw (1972). This study hypothesised a *transfer of salience* from the message to the receiver: issues that are visible in the media will become more salient, and hence more accessible, to the receiver. To test this hypothesis, they measured issues visibility as their media variable, and the importance of issues according to the audience as the variable in the social context with

which the message interacts. Issue visibility was measured by categorising a number of articles and television news items as being about one of 15 issues and news categories, distinguishing between major and minor articles based on length and prominence within the newspaper or news broadcast. From this categorisation they created an ordering of issues based on visibility for each time period. The Agenda Setting hypothesis was confirmed by a rank-order correlation between these two variables. A large number of later studies confirmed this effect in a variety of settings (e.g. Schoenbach, 1982; Dearing and Rogers, 1996; McCombs, 2004).

While Agenda Setting studies the effect of the media agenda on the public agenda, Agenda Building studies the factors determining the media agenda, in other words the 'why?' question from figure 2.1. Looking at characteristics of the real-world event covered in a news item, Galtung and Ruge (1965) identified a number of news factors that correlate with coverage, such as (cultural) proximity and negativity. To test these factors, they selected a number of events, counted how much coverage these events received, and correlated this with the news factors of the event. Harcup and O'Neill (2001) expand on this theory using a slightly different methodology: they also determine the news factors of the event on an analysis of the coverage, using coders to code the event an article is about and which news factors this event exhibits based on the description. In this latter study, Content Analysis is used to infer something from the perception of the real-world event by the message sender, and to infer the frequency and prominence of the messages about that event.

Second-level Agenda Setting expands the notion of Agenda Setting to the salience of attributes of actors or issues, using both affective attributes such as evaluations, and substantive attributes such as issues associated with an actor (McCombs et al., 2000). This line of research assumes that the salience of the association between the main object and an attribute is increased by media visibility of that attribute of the object. Kioussis et al. (2006) investigate first and second-level Agenda Setting and Building in the 2002 Florida Gubernatorial campaign. They perform a Content Analysis of both press releases and newspaper articles, having coders identify the occurrence of a set of main issue categories, and of clear statements of a number of candidate attributes, such as issue positions, integrity, and positive and negative evaluations. By correlating these occurrences in the newspaper articles, press releases, and a public opinion survey, Kioussis et al. (2006) managed to show a clear evidence for both Agenda Setting and Agenda Building, and somewhat mixed evidence on Second Level Agenda Setting and Building.

Related to Agenda Setting are studies into the transfer of salience between actors other than the media and the audience. For example, Van Noije (2007) studies the interaction of the issue agendas of media and

parliament by comparing the occurrence of issues in newspapers with the transcripts of parliamentary debates in three countries. In general, transfer of salience studies can be seen as an instance of what Monge and Contractor (2003) call contagion theories: theories that assume the transfer of ideas or attitudes along communication networks. In the mass communication model, this communication network consists of communication directed from a small set of media sources to a large number of unconnected media consumers, meaning that contagion flows from the media to the audience. If one assumes a more complicated network, such as bidirectional links between politics and the media, then contagion can occur in both directions, in line with the results of Van Noije (2007) and others.

2.2.2 *Evaluation*

Agenda Setting and related studies focus on the visibility or salience of issues and actors in texts or other messages. Other researchers focus on the evaluation of issues and actors: rather than looking at how often something is mentioned, they are interested in *how* it is mentioned. A famous example of this work is the Coefficient of Imbalance presented by Janis and Fadner (1943). They tried to work out whether a body of communication, for example the newspaper coverage of an issue, can be seen as balanced. For this, they counted all favourable, unfavourable, neutral, and irrelevant messages, and defined a formula to determine the balance of the coverage based on ten criteria of imbalance, for example that the coefficient must be zero if there is an equal amount of favourable and unfavourable news, and that it has to decrease in absolute terms if the proportion of neutral news increases.

Osgood et al. (1967) developed three primary dimensions of evaluation by conducting a factor analysis of the connotations of a large number of words according to test subjects. From this factor analysis there emerged three orthogonal dimensions which he called evaluation (good/evil), potency (strong/weak), and activity (active/passive). Holsti (1964) developed dictionaries for these categories in the General Inquirer program (see section 2.6.1, below), and used it for example for analysing political communication on the onset of the first World War and the Cuban missile crisis (Holsti et al., 1964a,b).

Fan (1985, 1988) also created lists of positive and negative terms, but rather than using them in a (normative) measure of imbalance, he used these to determine the spread of evaluative of 'ideas' (evaluations of actors and issues). Subsequently, these evaluations were used in a sophisticated time series model called the 'ideodynamic model' to predict the spread of such ideas within public opinion, using the metaphor of con-

tagious diseases. Results from studies using this method include that it is easy for an idea to spread to many people if most people are not yet 'infected' with either that idea or its opposite; and that messages opposing an idea can cause a large number of people to lose that idea if it has a large following (Fan, 1996; Kleinnijenhuis and Fan, 1999; Shah et al., 2002; Watts et al., 1999; Jasperson and Fan, 2004). In these and other studies, Fan showed how the systematic and large scale analysis of evaluative content can be used to predict public opinion on an aggregate level.

Evaluative Assertion Analysis, developed by Osgood et al. (1956), considers the evaluation of specific concepts by their association with and dissociation from each other and of common meaning terms: if an actor is associated with inflation, and inflation is considered to have a common negative meaning, then that actor is also evaluated negatively.

2.2.3 *The Horse Race*

In the coverage of elections, the media often pay attention to how well a candidate or party is doing, by reporting opinion polls or describing that a strategy is working or not, or the amount of donations a candidate receives. Such news is sometimes called *horse race* news (Patterson, 1993), evoking the image of a commentator on the race track continuously describing which horse is ahead of the pack. In communication science, the *bandwagon hypothesis* states that if an actor is portrayed as successful, her popularity or reputation will increase, causing more people to vote for her. The reverse hypothesis, that the loser gets sympathy or protest votes, is called the underdog effect (Simon, 1954; Patterson, 1993). To investigate the effect of the news coverage of the horse race, Farnsworth and Lichter (2006) correlate the occurrence of horse-race statements about a politician with his standing in the subsequent poll in the 2004 New Hampshire presidential primary. Occurrences of positive and negative horse-race evaluations were counted in 152 television news stories instructing the coders to look for phrases such as 'gaining ground' or 'slipping' (p.57). Their study confirmed the predominance of the horse race: "horse-race coverage dominated the more substantive evaluations for all candidates, and horse-race assessments were more significant in predicting changes in public opinion" (p.59). In a study of the effects of various news types in election campaigns, Kleinnijenhuis et al. (2007b) found that the effects of horse race news was stronger than the effects of news about issues, conflicts, or real world issue developments.

2.2.4 *Issue Positions*

The position of political actors in the political spectrum is decided by their issue positions. Issue positions are important in understanding coalition-forming between parties (e.g. Budge and Laver, 1986), and spatial voting theory explains voting behaviour in terms of the relation between the (perceived) issue position of voters and politicians (Downs, 1957; Rabinowitz and MacDonald, 1989).

An important project in measuring issue positions is the Manifesto Project (Budge et al., 2001). In this project, coders assigned sentences from party manifestoes from a large number of countries and elections to 56 issue categories in 7 domains. Although 11 of the investigated issues contain a positive and negative category (such as for or against expanding the welfare state), the focus of this project is on the relative emphasis on the different issues rather than their positions on these issues (cf. Laver and Garry, 2000, p.620). In contrast, the Party Change Project (Harmel et al., 1995) coded manifestoes on 19 issues on a scale of -5 to +5. In an attempt to automate the coding of issue positions, Laver et al. (2003) present a system for scoring texts with unknown texts using word scores derived automatically from texts with known issue positions.

2.2.5 *Conflict and Cooperation Between Actors*

In both politics and international studies, patterns of conflicts and cooperation between actors are interesting data. In international studies, the seminal COPDAB project (CONflict and Peace DATA Bank; Azar, 1980) aims to create a data bank of texts coded for conflicts between national actors. Within this project, Philip Schrodtt and colleagues developed the KEDS/TABARI program (discussed below) to automatically code actor–event–actor triples from newswire articles (Schrodtt and Gerner, 1994; Schrodtt, 2001; Schrodtt et al., 2005). These data are used to predict outbreak of conflict in for example the Middle East, functioning as an ‘early warning system.’ In political news, a form of conflict between politicians is negative campaigns. Ansolabehere et al. (1994) analyse the tone of campaign coverage of 34 U.S. Senate elections in 1992, assigning each campaign a score from -1 through +1 after a reading of the newspaper articles. Their study showed that negative campaigns lead to lower voter turnout.

2.2.6 *Framing*

Many recent studies in communication science focus on how actors or issues are *framed* in messages, see for example the recent special issue on

Framing of the Journal of Communication (2007/1). There is little consensus on the exact definition or theoretical underpinnings of framing, Scheufele (2000, p.103) stating that 'framing is characterised by theoretical and empirical vagueness' and D'Angelo (2002) calling it a 'multi-paradigmatic research program'. A frequently used definition is that of Entman (1993, p.52): "To frame is to select some aspects of a perceived reality and make them more salient in a communicating text, in such a way as to promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation", pointing out that such frames can be present in the sender, message, receiver, and the culture.

Valkenburg et al. (1999) experimentally study the effect of news frames on receivers' thoughts and recall. Each subject reads a newspaper article about either crime or the introduction of the euro, which was manipulated to be framed using either a conflict frame, a human interest frame, a responsibility frame, or an economic consequences frame. These subjects were then asked to give their opinion on the issue in their own words, and a Content Analysis was performed on the resulting answers: coders were asked to answer a number of questions for each frame, such as 'Is there mention of the costs or the degree of expense involved?' (p.561). The answers to these questions were averaged to determine the intensity of the frame in the respondent's answer. Their results showed that the frames in subject responses correlated with the frames in the stimulus article, and moreover that the human interest news frame negatively impacts recall.

2.3 Semantic Network Analysis

Most of the Content Analysis studies described above were performed using what Roberts (1997) defined as Thematic Content Analysis: the messages to analyse are unitised, and each unit is coded on one or more variables. The theoretical variables are then constructed by 'reducing' the individual codes through combination and aggregation. For example, in the Agenda Setting studies they measure whether an article is about one of their issues, which is aggregated to construct the ranked issue list (the theoretical variable). In the Frame Setting study of Valkenburg et al. (1999) described above, the scores for the indicator questions are summed and aggregated, and this aggregate score is the occurrence of the frame. This approach is easy to use and works well, as shown by the successful studies quoted above.

However, by closely tying the measurement-level variables to the analysis-level constructs, the data produced by such studies is *ad hoc* in

the literal sense of the word — the data is produced for answering one research question. Using the same data for answering other research questions, or even refining the answer to the same question by modifying the definition, is generally not possible. This problem is a consequence of the semantic gap between the symbols of the text and the semantics of the data language used to construct the theoretical variables. In particular, there are two gaps between the message and the data language: an abstraction gap and a complexity gap.

The *abstraction gap* refers to the fact that the words in a text or other message generally refer to concrete actors and issues, while the researcher is interested in higher-level concepts. The coders have to bridge this gap by receiving instruction in what each concept means and then judging whether a concrete actor or issue is included in that concept. For example, suppose that we are interested in news about ‘peace missions,’ and a newspaper article writes that “Marijnissen criticised the plans to continue the mission in Uruzgan.” This poses two problems. The first is that the coder has to know whether the mission in Uruzgan is a peace mission or a ‘fighting’ mission, placing a burden of interpretation on the coder. The second problem is that it is impossible to reuse this data for other purposes, for example to extract the attention to Defence topics, or to opposition politicians.

The *complexity gap* is that non-structured categories or variables are used to measure the occurrence of a complex phenomenon. For example, Valkenburg et al. (1999) measure the occurrence of an economic consequences frame by having coders answer questions such as “Is there a reference to economic consequences of pursuing or not pursuing a course of action?” In a text, such phenomena will generally be described using a number of propositions connecting the relevant actors and issues, for example stating that “the costs for the introduction of the euro will be paid by the taxpayer.” Since the data output by thematic Content Analysis like the question above is not complex (in the sense that it consists of measurements with no internal structure), the coder has the burden of converting the complex text to a single score (or set of scores). This causes the same two problems as the *abstraction* discussed above: the coders have to do more interpretation, and the data cannot be reused for different purposes, since the rich, structured information contained in the text was reduced to an unstructured variable. If we are interested in whether the source is expressing euroskepticism, for example, the fact that a message is framed in terms of economic consequences will not help much.

These two problems — coder interpretation and inflexible data — were also identified by Markoff et al. (1975), who required Content Analysis data to be “*semantically* as close as possible to the contents of the

original documents” and can be “recode[d] electronically in accordance with the coding refinements and modifications that emerge” during the research (p.3). This is a necessary condition for combining or reusing different data sets, as otherwise the data can only be combined if the underlying research questions are identical to each other and to the current research question. Content Analysis, either manual or automatic, is always a difficult and expensive process, yet large international and/or longitudinal data sets are needed to answer many interesting research questions. Consequently, it is difficult to overstate the importance of being able to create shared data archives by combining and reusing data.

After eliciting these requirements, Markoff et al. (1975) describe a ‘Representational’ Content Analysis method loosely based on Evaluative Assertion Analysis (Osgood, 1959; Osgood et al., 1956). In Evaluative Assertion Analysis, texts are translated into Attitude Objects and Common Meaning terms which are connected by Connectors. *Attitude Objects* are objects about which people can evaluate differently, in other words be in favour of or against, such as privatising health care. Common Meaning terms are terms about which people generally agree about, such as ‘atrocities’. In text, these objects are connected using Connectors, leading to object-connector-object triples that together form a network of objects. Following Krippendorff (2004), we call this type of Content Analysis *Semantic Network Analysis*; other names are Map Analysis (Carley, 1993) and Relational Content Analysis (Roberts, 1997).

Semantic Network Analysis deals with the *abstraction gap* by using objects that are as close to the text as possible in the extraction, and aggregating these along a hierarchy in the querying step. The choice of which objects to use, an *ontological commitment* in the terms of Chapter 4, determines to an extent the interpretation of the text, or in other words makes a choice of *context* for the Content Analysis. However, this context is kept as broad as possible in the extraction phase. The choice of how to aggregate the concrete objects to the abstract theoretical concepts is a much stronger commitment to a context, but this step is reversible, making the data useful for other analyses using other contexts compatible with the initial choice of objects.

Similarly, the *complexity gap* is dealt with by postponing the irreversible reduction of complex structures into unstructured variables to the queries used to answer the research question. This way, the coders map the rich structure of text to the relational structure of the network representation. If required by the research question, the queries map this relational structure to a set of unstructured variables, for example the frequency of specific patterns of one or more relations.

Figure 2.4 is an adaptation of figure 2.3, showing how Semantic Network Analysis fits into general Content Analysis methodology as de-

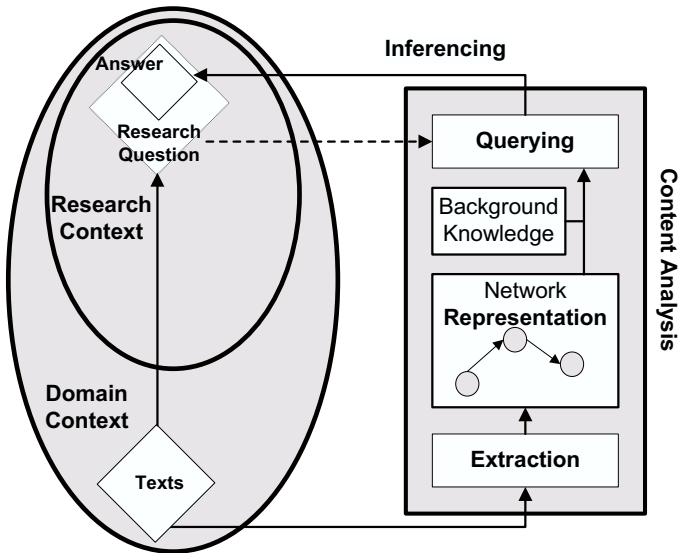


Figure 2.4: Semantic Network Analysis within Krippendorff's framework

scribed in Krippendorff (2004). The Content Analysis box on the right-hand side represents the Semantic Network Analysis process. The first step is a manual or automatic *extraction* of the relations between objects expressed by the text. This yields a *network representation* of these texts. This network representation is combined with relevant *background knowledge* about the objects in the network. This background knowledge contains both dictionary knowledge such as 'a politician is a person,' and encyclopaedic knowledge such as the fact that Bill Clinton was president from 1993 to 2001. This combined data set of media data and background knowledge is then *queried* to answer the research question.

On the left-hand side, there are two contexts rather than the single context shown in figure 2.3. The general *domain context* contains the broad assumptions necessary for extracting the network of objects, specifically the choice of objects and relations that are extracted, and the textual features that are used to extract those relations manually or automatically. The inner oval represents the more specific *research context*, containing the assumptions made to infer an answer to the research question. These assumptions include the way objects are categorised using the background knowledge, the operationalisation of the research question in terms of patterns in the combined data set of media data and background knowledge, and the interpretation of these patterns in terms

of the social context of the message.

The advantage of Semantic Network Analysis for sharing and reusing data can be explained in terms of these contexts: Since the extraction of the Semantic Network depends only on the domain context, that network can be used to answer any research question whose specific research context is contained in that domain context. In other words: to share data between researchers, their choice of which concrete objects to measure has to be compatible, but they can differ in how these objects are aggregated or how the network is queried to answer the research question. The advantage of Semantic Network Analysis for coding can also be seen in these terms: Since coding relies only on minimal assumptions about choice of vocabulary and relations, coders are not burdened with interpreting categorisation schemes or reducing complex textual phenomena to single variables. Automatic extraction is not necessarily easier, but since the extraction method is linked to the domain context rather than to the specific context of a research question, the method is more general and does not need to be retrained or reconstructed for different research questions as long as the assumptions made in the domain context remain valid.

The separation between extraction in the domain context and querying in the research context also allows us to position Semantic Network Analysis in the manifest–latent and quantitative–qualitative debates. In the former, since the extraction of the network should be done as close to the text as possible, it should be restricted to the manifest content. In the querying, however, latent content can be exposed by inferences based on the manifest content (such as transitivity of relations), by inference based on background knowledge (such as relations between objects that are assumed to be general, e.g. that inflation is bad for the economy), or by causal inferences assumed in the query operationalisation of the research question (e.g. inferring motives of sources from their messages). In the quantitative–qualitative debate, the main difference between the two positions is that quantitative analysis requires codes with a definition and meaning that is fixed before reading the text, while qualitative analysis builds the code structure and meaning in an interactive process while (re)reading the text. In the extraction phase, the domain context fixes the vocabulary and the meaning of the contained objects. However, the aggregation to more abstract objects and interpretation of patterns of relations happens in the research context, and this can be conducted in an interactive fashion, by conducting queries, inspecting the results and the original texts on which these results are based, and refining the definitions of the concepts and patterns used in the query. This is compatible with the deductive approach to qualitative coding as described by Miles and Huberman (1994). In this approach, the text is first coded using *de-*

scriptive codes, similar to the extraction of a representational network. The text is then recoded using *interpretive* codes, which place a layer of meaning on the original codes, and can be seen as an interactive parallel to the categorisation of objects using background knowledge. Finally, the text is coded using *pattern* codes to indicate interesting patterns and interaction, which is similar to the querying of the network representation. As noted by Pleijter (2006), many qualitative Content Analysis studies would benefit from a more rigorous description of the used analysis procedure, so it would be very interesting to explore whether Semantic Network Analysis can play a role in increasing that rigour without unduly limiting the hermeneutic analysis.

For Semantic Network Analysis to be possible, it is required that the concept to be constructed can be expressed as a pattern on the extracted network. Although the exact operationalisation is beyond the scope of this thesis, it is plausible that the constructs mentioned above can be expressed as patterns on a network: studying Agenda Setting requires determining the visibility of issues, which can be done by counting the the total number of links connecting the issue node to other nodes in the network. Evaluations of actors and issues can be operationalised as the connections of those actors and issues with the Ideal or with positive values. Second-level Agenda Setting uses the visibility of associations between objects and their attributes, which is the frequency of links between these objects and attributes. Horse Race news is represented by the positive and negative links connecting the real world with the studied actors. Issue positions look at the evaluations of issues in texts from a political source such as manifestoes, or at relations between actors and issues in other texts. Political conflict is naturally operationalised as negative relations between political actors. Frames are defined by Entman (1993) as salience of selected attributes which are also operationalisable as the edges between the framed issues and these attributes. The specific frames as defined in Valkenburg et al. (1999) can be seen as patterns of relations between the relevant objects. For example, the economic consequences frame discussed above can be operationalised as the frequency and valence of causal links from the framed issue to objects representing economic issues. These observations make it plausible that networks are a good way of representing media messages; and successful applications of Semantic Network Analysis point in the same direction (Kleinnijenhuis et al., 2007b,a; Roberts, 1989; Carley, 1997).

2.4 The NET method

Chapters 6 – 9 use an existing Semantic Network data set to develop and test techniques for automatically extracting, representing, and querying media content. This data set, the analysis of the Dutch 2006 election campaign, is coded using the the Network analysis of Evaluative Texts (NET) method. This method is a Semantic Network Analysis method developed by Van Cuilenburg et al. (1986), and extensively used for analysing political campaign coverage (Kleinnijenhuis et al., 1995, 2007b,a), corporate news and reputation (Meijer and Kleinnijenhuis, 2006), coverage of the Bosnian War (Ruigrok, 2005), and economic discourse (Kleinnijenhuis et al., 1997). This section will describe this method, distinguishing between the *extraction* and *querying* steps described above.

Extraction Similar to Osgood et al.’s (1956) Evaluative Assertion Analysis, NET Analysis divides a text into propositions called Nuclear Statements. Each Nuclear Statement consists of a *subject*, a *predicate*, and an *object*. The predicate connects the subject and object by association or dissociation: If ‘*Bos criticises Balkenende*,’ Bos is dissociated from Balkenende, while ‘*Bos being in favour of universal health care*’ associates him with the issue of universal health care. This connection is quantified as the *quality* of the connection, ranging from -1 (maximal dissociation) to $+1$ (maximal association). NET further distinguishes between different kinds of connection, in particular between affinity, action, causation, and equivalence. Affinity is a statement about what an actor would like to do or to see happen; action is a statement about an actor consciously doing something; causation is an actor or issue causing something to happen without conscious effort; and equivalence is the statement that two objects are equal or comparable (e.g. ‘*Universal Health Care is communistic*.’)

The Subject and Object of the statement are drawn from a set of meaning objects called an *ontology*.¹ Fully described in section 8.4, this ontology can be seen as a hierarchy that describes that Balkenende is a Prime Minister, which is an Executive Politician, which is an Actor. Similarly, Universal Health Care is a Health Care issue, which is an Issue. In Evaluative Assertion Analysis, crime is considered a Common Meaning term since everybody is against crime. In NET, such valence issues are treated as normal issues during the extraction step, but can be distinguished in the querying step. For purely evaluative statements, there is a special object called *ideal*, which represents the positive Ideal. Thus, a sentence like ‘*Saddam is evil*’, is coded as a nuclear statement dissociating Saddam from the ideal by calling him evil. Another special object, *reality*, is used

¹See section 4.1 on page 52

when something happens without a specified cause. Thus, *'Inflation risen to 3%'* would be coded as reality causing inflation, while if it specified that unemployment caused inflation to rise, reality would not be used since the cause of the rise is known. Thus, *'High unemployment caused rising inflation'* would be coded as reality causing unemployment and unemployment causing inflation.

News items such as newspaper articles often contain literal or paraphrased quotes from actors. In NET, all statements contain one or more (nested) sources. Each additional source effectively creates a separate network within the network of the containing source: this represents the world of the quoted source according to the newspaper or other message source. Occasionally, a newspaper will refer to another newspaper quoting an actor, in which case there is a double nesting.

Often, the complex sentences used in newspaper articles will contain many such triples. A fictional sentence like "The VVD voted against the proposed tax cuts as they fear the resulting deficit will increase inflation" contains an action (the VVD voting against a proposal), and two causal statements made by a source (according to the VVD, the proposal will increase the deficit, and the deficit will increase inflation). Since NET is limited to dyadic predicates (relations between single subjects and objects) and does not allow predicates to be used as subject or object of other predicates, not all information contained in some sentences can be represented. For example, in "The CDA rejected the vote of the VVD against the proposal as a populist gesture" there is a negative relation between the CDA and the relation representing the vote of the VVD. Since this relation cannot be used as a subject or object, this sentence would be represented as the CDA being against the VVD, since that is the relation that comes closest to the meaning of the sentence in its political context. Similarly, if "the CDA and VVD disagreed on the tax cut," it is obvious that there are negative relations between the VVD and CDA, but it is not generally possible to code the cause of the argument. These limitations are the consequence of remaining within a graph representation, as using relations within relations or relations between more than two objects would result in more complex structures than graphs. Staying within a graph representation is an important advantage as graphs are well-understood mathematically and many methods from graph theory, logic, and social network analysis can be applied to graphs. However, if a research question requires more complicated relations, it is possible to augment the NET method, for example by creating an *angle* or indirect object field to represent the topic of disagreement, or creating special objects to represent certain relations about which there is discussion, such as "Wilders making the film *Fitna*" in the recent debate on that film. Such additions can help answer the research question but are difficult to

interpret in a general graph-centred manner. If one makes sure that the additions strictly add information (i.e. the basic subject–predicate–object relation is still valid) the resulting network will still be useful for general-purpose analysis by ignoring these additions.

Consider the example article in Figure 2.5, which is a (translated) description of a debate between the various party leaders in the Dutch 2006 parliamentary elections, focussing on the clash between Wouter Bos (PvdA / Social Democrats) and Jan Peter Balkenende (CDA / Christian Democrats), the main contenders for the position of Prime Minister. The headline is coded as a reciprocal negative relation between the political blocks Left and Right. The first sentence of the lead is more complicated: The main message is that Balkenende and Bos are fighting, but it is also stated what they are fighting about: the issues Poverty and Health Care. This is coded as two reciprocal negative relations between the two politicians, with the two issues in the respective *angles* of the two relations. In the next sentence, Balkenende states that Bos is scaring people, which is coded as Bos acting against the Dutch citizens with Balkenende as source. The final sentence expresses two relations: according to Bos, investing more money would be good for Health Care, and Bos wants to invest money in Health Care, here coded as an affinitive (issue position) relation between Bos and Health Care Investments. All statements also have the newspaper *De Telegraaf* as implicit primary source.

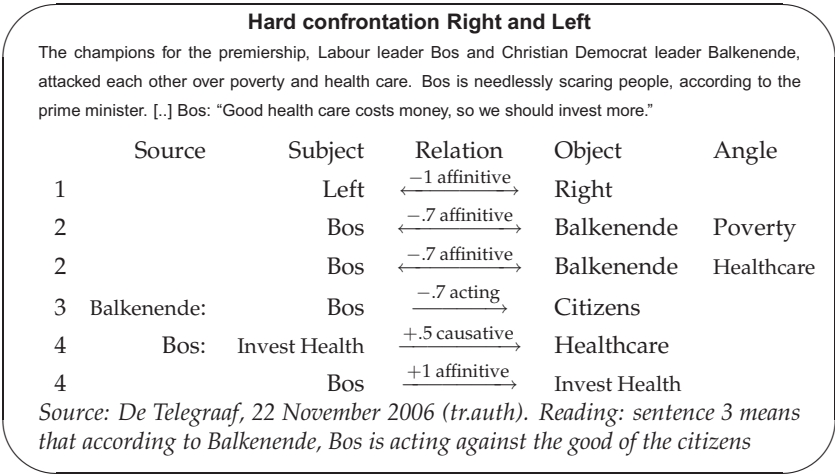


Figure 2.5: Example article with NET coding

Querying The procedure outlined above yields a network of actors and issues, or rather a set of networks for each article and nested source. The queries that are used to answer the research question will not generally be posed at the level of relations between concrete objects in single articles, but rather at the level of (patterns of) relations between abstract categories in collections of articles. The NET method does not specify how the network should be queried, but it does offer a number of ways to aggregate and enrich the extracted network. These methods will be discussed here.

A first step in enriching the network is combining networks and aggregating the nodes and edges. The unit of analysis is often a collection of individual messages, such as all articles in a newspaper in one week. The networks representing these messages are combined into one network for each unit of analysis. Subsequently, the hierarchy is used to aggregate all actors and issues to the concepts at the level of detail required to answer the research question. For example, all members of a political party will be combined with that party into a single node, and all issues under *Leftist Issues* will be similarly combined with that issue. Object (nodes) in the network are now often connected by multiple statements (edges), which can be combined into single edges by calculating the frequency, average quality, and diversity and ambiguity of the quality, i.e. the variance within and between relations (Van Cuilenburg et al., 1986, p.91–92). This yields a network per unit of analysis, consisting of objects at the level of abstraction required for the research question and with a single relation between each pair of objects.

Optionally, depending on the research question, *transitivity* may be applied to the links in the network to infer indirect links: if Bos is against inflation, and inflation causes economic growth to slow, then we can infer that Bos is probably in favour of economic growth using the intuition of ‘the enemy of my enemy is my friend.’ This intuition is a generalisation of Balance Theory and Cognitive Dissonance theory (Festinger, 1957; Heider, 1946) to semi-cycles rather than triangles (cf. Wasserman and Faust, 1994, p.220–233), and to links with both intensity and valence rather than just valence. Such an indirect link is called a *chain*, and the combination of all chains between two objects is called a *bundle* (Van Cuilenburg et al., 1986, p.92–93). Mixed messages may exist: for example, a politician can be in favour of cutting taxes and against budget deficits, while the article states that cutting taxes will increase the deficit. Such mixed messages will lead to a bundle between the politician and the deficit with a high diversity. This can signal an interesting phenomenon, possibly indicating cognitive dissonance or a desire of the source to portray the politician as inconsistent.

Another inference is on the interaction between the sub-networks of

sources and their containing networks. *Source Elimination* allows the lifting of statements from the source network to the containing network if the source is considered neutral, such as in a newspaper article quoting an expert or a ‘well-informed source.’ This assumes that the original message source quoted the source as a way of presenting his or her own opinion. *Judgement Extrapolation* transforms a quoted evaluative statement into an affinity between the source of the quote and the evaluated object: an actor stating that something is bad implies that the actor is against it. This can be combined with transitivity to allow inferring an indirect affinity by an indirectly expressed evaluation: if Bos states that the Government is spending too much, which causes inflation, which is bad, it is implied that Bos is against the government. If inflation is a negative valence issue, the latter part (inflation is bad) may even be left implicit, since being in favour of a negative valence issue implies being associated with the negative ideal. Transitivity and quoted network interaction can be considered ways to uncover the latent meaning of texts.

2.5 The 2006 Dutch parliamentary elections

As stated above, a large part of the substantive work in this thesis has been conducted using the data set created in the analysis of the Dutch 2006 elections. This section gives a brief overview of these elections, see Kleinnijenhuis et al. (2007a) for a more comprehensive description.

Dutch parliamentary elections took place on 22 November 2006 with the Dutch voters causing an upheaval in the political arena of The Hague. Instead of the expected titanic struggle between the leaders of the Christian Democrats (CDA; see table 2.1 for an overview of the main parties and their leaders) and the Social Democrats (PvdA), voters deserted es-

Table 2.1: Main parties in the 2006 elections

| Party | Description | Leader |
|-------|---------------------------------|----------------------|
| SP | Socialist Party | Jan Marijnissen |
| PvdA | Labour, Social Democrats | Wouter Bos |
| CDA | Christian Democrats | Jan-Peter Balkenende |
| VVD | Conservative Liberals | Mark Rutte* |
| PPV | Anti-immigration, extreme right | Geert Wilders |

* Rita Verdonk was the runner-up in the highly contested primary election for the VVD party leader. She ran a largely separate campaign and she, and the strained relations between her and Mark Rutte, received considerable media attention.

established parties, especially the PvdA and the conservative Liberal Party (VVD), in favour of more outspoken parties on both the left and right wings of the political spectrum. This is illustrated in figure 2.6, which clearly shows the increase of the SP and VVD at the expense of especially the more centrist PvdA and CDA. As shown by (Kleinnijenhuis et al., 2007a) these short-term dynamics of voter preferences are highly influenced by preceding news coverage.

In the months before the election, the news coverage in the Netherlands was filled with the upcoming elections and the campaign preceding the vote. Based on early opinion polls, the media speculated on a ‘titanic struggle’ between the two largest parties, the CDA of incumbent Prime Minister Jan Peter Balkenende and Wouter Bos, the leader of the largest opposition party, the PvdA. This struggle, however, failed to materialise. On the contrary, in September the incumbent government presented the 2007 budget full of good news (“After the bitter comes the sweet”, De Telegraaf, 20 September 2006) and the CDA became more often presented as a successful party responsible for economic growth, while the PvdA went into a downward spiral with Bos presented as a ‘flip-flop.’ The other incumbent party, the VVD, could not profit from this success as it was burdened down with internal struggles. After the internal party election between two candidates representing the liberal and conservative wings of the party, the party was unable to form a unified front. In the shadow of this battle and the problems of the PvdA, the smaller parties seized the opportunity to present themselves as a more outspoken alternative: The right-wing Party for Freedom (PVV) presented itself as an alternative for the right-wing liberals, while the Socialist Party (SP) portrayed itself as a stronger and more outspoken alternative for the PvdA. Rather than the political landscape of two large parties as speculated on initially, this resulted in a parliament with five relatively strong parties, which made it very difficult to form a coalition government as even the two largest parties (CDA and PvdA) together did not constitute a majority.

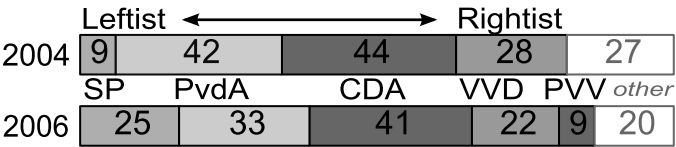


Figure 2.6: The political spectrum before and after the 2006 elections

2.6 Computer Content Analysis

The sections above gave a general overview of Content Analysis methodology and in particular of Semantic Network Analysis and the NET method that is used extensively in this thesis. Part II of this thesis describes methods to automate the measurement part of NET Content Analysis. This section will give a brief overview of some of the Automatic Content Analysis methods currently in use. See Krippendorff (2004, ch.12) and Poppinga (2000, p.185–203) for a more complete overview.

2.6.1 *The Dictionary Approach: General Inquirer and others*

One of the oldest approaches to automatic content analysis is what (Krippendorff, 2004, p.283) calls the dictionary approach: a computer program contains a dictionary of words corresponding to a certain meaning category, and the program counts the occurrence of these words in the input texts.

The General Inquirer is one of the oldest dictionary based computer programs and as such deserves special mention in this section. First presented by Stone et al. (1966), it has been used and developed until the present day. At its heart, the General Inquirer is a very simple program that maps words to categories using a dictionary of words for each category. It also has a disambiguation system for words used in multiple categories and does limited stemming in the sense of matching root words to inflected forms.

Although the General Inquirer can process any kind of dictionary, it is mainly used with its built-in dictionary. This dictionary is the result of the ongoing development of the General Inquirer, containing, among others, the categories of Osgood et al. (1967) and the Lasswell value dictionary (Namenwirth and Weber, 1987). In total, this dictionary has 182 categories ranging from a dozen to 2291 entries.

As surveyed by Alexa and Züll (1999), there are currently a great number of largely incompatible computer programs for dictionary-based coding, many of which also offer explorative analysis such as concordances, Keyword-In-Context (KWIC) listings, and frequency analysis². For example, TextQuest³, (developed by Harald Klein), TextPack (Mohler and Züll, 1998), and VBPro (Miller, 1995) all offer various word frequency, concordance, KWIC, and dictionary-based analysis features. Diction similarly offers a set of custom dictionaries developed for capturing the “tone of a verbal message” in five master variables: Certainty, Activity, Optimism, Realism, and Commonality (Hart, 1985, 2001).

²See also <http://www.textanalysis.info/> for an overview

³<http://www.textquest.de/eindex.html>

Some statistical packages now also contain dictionary-based automatic coding modules, for example for coding open-ended survey questions. WordStat⁴ is a module in the SIMSTAT package, and allows both dictionary based coding and Keyword-in-Context (KWIC) functionality. SPSS offers two modules, a dictionary-based system called LexiQuest Categorize, and a more inductive relation-extraction system called LexiQuest Mine⁵.

2.6.2 Qualitative Text Analysis: *Atlas.ti* and others

There are a number of software packages for Qualitative content analysis, aimed at assisting manual coding rather than replacing the manual coder. For this reason, these software packages are called Computer Aided Qualitative Data Analysis Software (CAQDAS) programs. Lewins and Silver (2007) describe and compares a number of these CAQDAS packages

Atlas.ti is a typical and well-known CAQDAS program. It combines document and codebook management with coding facilities, allowing the user to add documents, define codes and relations between codes, and assign these codes to parts of the added documents. Aimed at qualitative content analysis, Atlas.ti does not fix the unit of measurements at sentences or paragraphs, but allows the user to select a piece of text, called a quote, and code it using a new or existing code. Atlas.ti also allows the user to create relations between codes, including is-a relations. This creates a network of quotes, assigned codes, and relations between these codes. However, this network is not used for formal inference such as aggregation or querying such as done in Semantic Network Analysis as described above. Rather, it is a tool to allow a researcher to explore the quotes and build a theory of the meaning of the text by looking for quotes with related codes. The link between meaning and text being very important in qualitative analysis, Atlas.ti makes it easy to view and navigate to the quotes belonging to codes and to find related quotes and codes.

Other CAQDAS programs are MaxQDA⁶, NVivo⁷, and Kwalitan⁸. All programs allow coding of selected segments of text and searching for text using these codes. These packages also allow the creation of a code hierarchy similar to the network structure of Atlas.ti, and in Kwalitan and MaxQDA this hierarchy can be used functionally to find and

⁴www.provalisresearch.com/wordstat/wordstat.html

⁵http://www.spss.com/predictive_text_analytics/

⁶<http://www.maxqda.com>

⁷<http://www.qsrinternational.com>

⁸<http://www.kwalitan.net>

manipulate all segments coded with a branch of the code hierarchy. Due to the importance of iteratively going from text to code and back, all packages have advanced functionality for viewing existing codes in text and retrieving the text segments associated with a code.

2.6.3 *KEDS / TABARI*

The Kansas Events Data System (KEDS)⁹ is the name of the program that was developed in the late 1980's to automatically code international events data, such as one country making a proposal to another country (Schrodt et al., 2005; Schrodt, 2001; Schrodt and Gerner, 1994). KEDS has since been superseded by TABARI (Text Analysis by Augmented Replacement Instructions), written in C++ and released under the GPL open source license. The data model assumed by KEDS and TABARI is simple: an event set is a set of (Subject, Event, Object) triples, where the Subject and Objects are taken from a list of actors and the events are generally taken from a standardised events typology such as WEIS (McClelland, 1976) or COPDAB (Azar, 1982). International events are conceived as either positive or negative acts, or statements of one actor towards another.

Whereas NET is used primarily to code the variety of overlapping and partially contradictory reports in the media and in the political arena, event analysis is aimed ultimately at assessing the positive or negative nature of real-world events on the basis of a large variety of media reports about them.

TABARI focusses on relations between actors: its aim is not to reveal relations with issues or assess performance or morality. At the heart of the assignment of either positive or negative relations in KEDS is a pre-defined categorisation of statements and acts, starting from a total war between actors, towards various negative and positive statements, and ending with the complete unification of actors.

The event coding program, TABARI, processes individual sentences. Preprocessing removes punctuation and looks up all words in the supplied coding dictionaries. Verbs and nouns are disambiguated based on articles and capitalisation patterns. There are a set of rules to process anaphoras and to handle compound phrases and clauses and subordinate clauses and to remove non-restrictive clauses.

Subsequently, the program tries to match verbs in the sentences to phrases from the dictionary. If such a phrase is found, the program finds the corresponding subject and object using surface pattern matching: If the dictionary phrase specifies actor locations, these are used; if not, the first actor in the sentence and first actor following the verb are used, or

⁹<http://www.ku.edu/~keds>

(if no actor follows the verb) the first non-subject actor in the sentence. Compound noun phrases are expanded into multiple triples.

2.6.4 Map Analysis: Automap and ORA

Map Analysis is the name of a method of Relational Content Analysis based on the pioneering work on Cognitive Maps by Axelrod (1976) and developed at Carnegie Mellon University from the 1980's (Carley, 1986, 1993, 1997). Based on this work, an XML data format called DyNetML (Tsvetovat et al., 2003) and automatic relational content analysis program Automap (Diesner and Carley, 2004)¹⁰ were developed.

Whereas NET makes an attempt to code the nature of relationships between nodes in great detail, Map Analysis succinctly reveals whether, or how frequently, a relationship between two nodes exists, thereby shifting its focus towards the nature of the nodes in the network and the nature of the network of relationships. Whereas NET had its origins in mass communication, and KEDS/TABARI in International Relations, Map Analysis is rooted in Organisational Communication and Sociology and in Social Network Theory (Doreian et al., 2004; Wasserman and Faust, 1994).

In DyNetML, nodes are divided into a number of different node types, such as actors, tasks, and knowledge. A Dynamic Network consists of a set of relations, where each relation can be between two node sets of the same or different types. Although these relations can be an arbitrary number, the representation is generally used for dichotomous or weighted relations. Nodes can have arbitrary string-valued attributes.

Automap is a proximity based network extraction tool. It identifies nodes combining a delete list and a generalisation thesaurus, which assigns a category to each word in the thesaurus, and a meta-matrix thesaurus that assigns a node type to each category. Relations are coded using proximity with options for counting only recognised words and for using linguistic units rather than a fixed window size.

ORA (Organisational Risk Analyser) can be seen as a companion program of Automap: whereas Automap conducts the extraction of networks, ORA is used for analysing these networks. ORA calculates a number of Social Network Analysis and has special features for analysing networks of and between actors, skills, and tasks, and for simulating changes to those networks.

A related program is Crawdad (<http://www.crawdadtech.com>), based on a text analysis method called Centring Resonance Analysis (Corman et al., 2002), which creates a network of nouns linked by their co-occurrence in noun phrases, and uses the Social Network Analysis

¹⁰<http://www.casos.cs.cmu.edu/projects/automap/>

metric centrality to determine the relative importance of the nouns in these networks.

2.7 Conclusion

This chapter presented an overview of the field of Content Analysis, starting from a number of definitions and applications of Content Analysis in the literature. It defined Semantic Network Analysis as a Content Analysis technique and placed it within the general Content Analysis methodology. It also described the NET method, which is the Semantic Network Analysis method that serves as the basis for most of the techniques presented later in this thesis, and the Dutch 2006 election campaign coverage that provided much of the data used. Finally, it reviewed a number of existing computer programs for content analysis, both for fully automatic analysis and for assisting manual coding. This provides the conceptual starting point for the methods and tools presented in parts II – IV.

CHAPTER 3

Natural Language Processing

‘Unfortunately, or luckily, no language is tyrannically consistent. All grammars leak’
(Andrew Sapir, *Language: an Introduction to the Study of Speech* (1921, p. 39))

The first Research Question of this thesis deals with automatically extracting Semantic Networks. Automatically extracting networks from text requires the processing of text by the computer, which is the topic of study in the field of Natural Language Processing. This chapter reviews some of the terms and processes used in Natural Language Processing, providing the background knowledge for chapters 5 – 7.

3.1 Introduction

The first Research Question of this thesis deals with automatically extracting Semantic Networks from text. For this, we make use of techniques from Natural Language Processing (NLP). NLP is a field of Artificial Intelligence that is concerned with extracting the structure and meaning of natural language by the computer. This chapter briefly reviews some of the terms and techniques used in that field.

Linguistics have a long history of trying to unravel the structure of language, and since the invention of electronic computers attempts have been made to use the computer to apply the linguistic findings automatically, using formal grammars or simple probabilistic approaches (Jurafsky and Martin, 2000, p.11-13). Although important advances were made, it was found that it is difficult to scale the early rule-based systems up to the full complexity of language (cf. Manning and Schütze, 2002, p.4–6). Recent decades have seen an enormous increase in available computer power and storage, and the availability of digital texts, allowing the development of Corpus Linguistics and Statistical NLP. These techniques use large annotated or unannotated corpora to refine the hand-crafted rules and discover new patterns automatically, allowing the techniques to scale up to natural language.

Although Content Analysis and Natural Language Processing both deal with extracting meaning from language, they should be seen as complementary rather than competing. The goal of Content Analysis (and hence Semantic Network Analysis) is to answer a research question dealing with an interaction between a message and the social context of that message. NLP, on the other hand, deals with unravelling the structure of the message itself. Hence, progress in NLP makes the task of Content Analysis easier, as the gap between the theoretical concepts used in Communication Research and the extracted linguistic structure is presumably smaller than that with the raw text, allowing the content analyst to concentrate on the link with the social context rather than with the capriciousness of text.

The next section describes the ‘preprocessing pipeline,’ a sequence of linguistic processing techniques that are commonly used to preprocess text prior to more sophisticated analysis. Section 3.3 briefly describes the use of thesauri in linguistic processing. Section 3.4 defines some performance metrics commonly employed in NLP that are used in some of the later chapters.

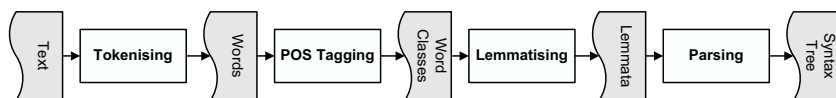


Figure 3.1: The Linguistic Preprocessing Pipeline

3.2 The Preprocessing Pipeline

In the automated linguistic processing of natural language, a number of techniques are designed to be used sequentially, meaning that the output of the first module is the input of the second. Hence, the sequence of these techniques is often called a *processing pipeline*. The techniques presented in part II assume that some or all of these preprocessing steps have been taken.

Figure 3.1 visualises a typical preprocessing pipeline, the white rectangles representing the preprocessing modules described here and the grey documents representing the input and output data. Not all techniques described in part II require a complete run through the pipeline, since some techniques might only require lemmatising while other techniques require their input to be fully parsed. The remainder of this section describes the components of the pipeline. For each component, it describes what the component does, why the output is useful for Content Analysis, and it briefly describes what techniques are used. To illustrate the working of each component, (parts of) the following example sentence will be used:

Senator Barack Obama broke forcefully on Tuesday with his former pastor, the Rev. Jeremiah A. Wright Jr., in an effort to curtail a drama of race, values, patriotism and betrayal that has enveloped his presidential candidacy at a critical juncture. At a news conference ... (Obama's Break With Ex-Pastor Sets Sharp Shift in Tone, *The New York Times*, April 30, 2008)

Most of the techniques presented in this paper are developed for the Dutch language and hence rely on Dutch NLP techniques. Fortunately, the structure of English and Dutch is similar enough to demonstrate the general working of the various preprocessing techniques in English without being distracted by translations.

3.2.1 Tokenisation and Sentence Boundary Detection

Example Input

his former pastor, the Rev. Jeremiah A. Wright Jr., in an effort ... at a critical juncture. At a news conference ...

Example Output

(Sentence 1) his former pastor , the Rev. Jeremiah A.
Wright Jr. , in an effort ... at a critical juncture .
(Sentence 2) At a news conference ...

The first two components, which are lumped together in this description, transform the text from a list of characters to a list of sentences and tokens (words). In Dutch, as in English, this task is not very difficult because we use spaces to delineate words, and full stops to mark sentence boundaries. There are some complications, however, as full stops are also used in abbreviations and punctuation is a separate token but generally written after a word without a space in between. In the example sentence, the system has to understand that the comma after *Jr* and the full stop after *juncture* are punctuation tokens, while the full stops after *Rev*, *A* and *Jr* are part of these words and not a separate punctuation token. Similarly, the sentence boundary detection has to detect that the full stop after *Rev* does not end a sentence, even though the next word is capitalised.

Apart from being a prerequisite for further processing, tokenisation can be useful for Content Analysis because one often searches for specific words rather than character strings. Sentence boundaries can be useful for co-occurrence analysis as performed in chapter 5: two concepts occurring in the same sentence can be more informative than the two concepts occurring within a fixed number of words or characters.

Generally, regular expressions (simple character patterns) and lists of known titles and abbreviations are often sufficient to conduct this task, although Machine Learning Systems are also used (e.g. Grefenstette and Tapanainen, 1994; Reynar and Ratnaparkhi, 1997).

3.2.2 Part of Speech (POS) Tagging

Example Input

Senator Barack Obama broke forcefully
on Tuesday with his former pastor

Example Output

Senator/NNP Barack/NNP Obama/NNP broke/VBD forcefully/RB
on/IN Tuesday/NNP with/IN his/PRP\$ former/JJ pastor/NN

In teaching children the structure of their native language, one of the first things taught is often the difference between the Parts of Speech such as verbs, nouns, and adjectives. Since its part of speech determines to a great extent how a word can be conjugated and used in sentences, this is

also a very important step in Natural Language Processing (cf. Fellbaum, 1998, chs.1–4).

In the example sentence above, *Senator Barack Obama* and *Tuesday* are tagged as names or proper nouns (NNP). *Broke* is tagged as a past tense verb (VBD), and *on* and *with* are prepositions (IN). *Forcefully* is tagged as adverb (RB), *former* is an adjective (JJ), *his* is a possessive personal pronoun (PRP\$), and *pastor* is a singular noun (NN).

The choice of which tagset to use and how to label and define the tags is not fixed: a number of tag systems exist with different levels of detail. The tag set used above is called the Penn Treebank tag set and has 48 tags (Marcus et al., 2004), while the Dutch WOTAN tag set has 233 very detailed tags based on 13 base tags (Berghmans, 1994).

Knowing the Parts of Speech of text is useful in Content Analysis for two reasons: First, it can reduce ambiguity problems by searching for words with a specific part of speech, for example only accepting the word ‘Bush’ if it is a proper name or the word ‘code’ if it is a verb. A second usage is more inductive: if we are interested in compiling a list of positive and negative adjectives, a POS-tagged corpus can be used to extract all adjectives occurring with a specified minimum frequency. This can help reduce the coverage problem associated with word lists.

Tagging software is generally data driven: the tagging rules are learned automatically based on a (large) corpus of example text that has been tagged manually. Various methods exist for learning these rules, but they generally contain a lexical database of possible tags per word, disambiguation based on tags assigned to previous words, and a system for handling unknown words (cf. Jurafsky and Martin, 2000, ch.8).

3.2.3 Lemmatisation

Example Input

```
Obama/NNP broke/VBD forcefully/RB  
with/IN his/PRP$ former/JJ pastor/NN
```

Example Output

```
Obama/NNP/Obama broke/VBD/break forcefully/RB/forceful  
with/IN/with his/PRP$/he former/JJ/former pastor/NN/pastor
```

In Dutch as in English, the base form of a word can be conjugated into multiple forms to express attributes such as the gender, number, and person of a noun or verb. For example, the word *broke* is the past tense form of the verb *to break*. The *lemma* of a word is the head word under which it would be listed in a dictionary, generally the infinitive form of a verb and the singular form of nouns. *Lemmatisation* is the process of

reducing the words in a text to their lemmas. In the example sentence, the verb *broke* is lemmatised to *break*; *forcefully* is lemmatised to *forceful*, and *his* becomes *he*. The remaining words are identical to their lemma and remain the same.

Especially in highly conjugated languages, lemmatisation reduces the number of different forms in which the same word appears.¹ Since in Content Analysis one is generally interested in the meaning of a word rather than its syntactic function, this reduces the amount of synonyms that have to be included in searching for a concept. Although using techniques such as wildcard searches can also help here, this generally does not help for irregular conjugations such as strong verbs or irregular plurals, and can accidentally include words with the same prefix.

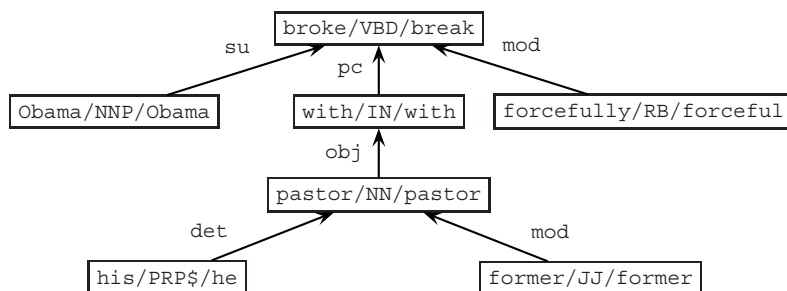
Programs that conduct lemmatisation generally use a lexicon or list of possible lemmas for each surface form, and a component for guessing unknown words, for example by stripping common word endings. Although a lemmatiser could use the raw text as input, it often uses the Part-of-Speech tags to determine which conjugations are possible.

3.2.4 Parsing

Example Input

Obama/NNP/Obama broke/VBD/break forcefully/RB/forceful
with/IN/with his/PRP\$/he former/JJ/former pastor/NN/pastor

Example Output



The last step in the pipeline shown in figure 3.1 is that of syntactic parsing. Where POS-tagging and lemmatisation give information about the

¹In this case, English is substantially simpler than many other languages. For example, Dutch conjugates verbs in three singular and one plural form, and German has even more conjugations due to the overt case marking. Agglutinative languages such as Hungarian or Turkish are among the most conjugated languages, among others also expressing modality within the main verb.

words, parsing looks at the grammatical relation between words. In traditional parsing, words combine to create larger units, which are then combined into a single sentence unit. This creates a tree or graph structure called the *parse tree* or *syntax tree*. This creates a *parse tree* with the words as leaves and the abstract categories, such as noun phrases, as non-final nodes. A simpler alternative to represent the output of parsing is using a dependency tree. In a dependency tree, all nodes are words, so no abstract categories are used in the tree. The root (top) of the tree contains the head word of the sentence. Similarly, in each branch the root is the head of that branch. This reduces the size and complexity of the tree and makes it easy to see what the most important words and relations are.

The dependency tree for the example sentence *Obama broke forcefully with his former pastor* is displayed above. The verb *broke* is the head of the sentence, with *Obama* as its subject (su) and modified (mod) by *forcefully*. *With his former pastor* is the prepositional complement (pc) of *broke*, and *with* is the head of this complement. *His former pastor* is the object (obj) of *with*. *Pastor* is the head word of this object, with *his* as its determiner (det) and *former* as a modifier.

Dependency trees have the advantage of not creating any abstract nodes, thereby creating a smaller tree structure. The example dependency tree given above has 7 nodes, while the corresponding syntax tree with abstract categories would have at least twice as many. Since different linguistic traditions yield different parse trees for the same sentence, much like the different POS-tag sets discussed above, dependency trees are often used as a common base for using and comparing different parsers. For example, the Dutch Alpino parser that is used in this thesis is a Head-Driven Phrase Structure Grammar (HPSG) parser, but we use the dependency output rather than the full HPSG parse tree (Van Noord, 2006).

Parsing can be very useful if we are interested in investigating the relation between concepts or characterisations or evaluations of single concepts. The same basic relation, such as an actor being the subject of the verb *to fail*, can be expressed in many different ways by having structures such as relative clauses and adjectives in between. By searching for patterns in the dependency parse rather than the raw text, we can greatly increase the coverage of patterns. Like POS-tagging, we can also use this inductively, for example by compiling a list of all adjectives applied to politicians, making sure that no important adjectives are missing in our word lists.

Initially, parsers mainly consisted of hand-built rules that govern the joining of constituents to form larger units such as Verb Phrases. State-of-the-art parsers are generally trained on large corpora of manually

or semi-automatically parsed texts, such as the Penn Treebank (Marcus et al., 2004) or the Corpus Gesproken Nederlands (Corpus of Spoken Dutch; Van der Wouden et al., 2002).

3.3 Thesauri

A recurrent problem in searching in text is that of synonyms: there are often a number of words that share the same meaning. For example, if we are interested in whether texts talk about being rich or being poor, we want to look for all synonyms of ‘rich’ and ‘poor.’ Thesauri are lists of sets of synonyms, and a number of machine-readable thesauri are available.

For English, two well known thesauri are Roget’s Thesaurus (Kirkpatrick, 1998) and WordNet (Miller, 1990; Fellbaum, 1998). Roget’s thesaurus was created in 1805 (first published in 1852) by Peter Roget and has been continually updated and expanded since, currently containing over 200,000 entries in almost 60,000 synonym groups. WordNet is more recent, having been developed at Princeton University since 1985. Version 3.0 contains over 150,000 words in more than 100,000 synonym sets, and is used in many application in computer science. For Dutch, a good thesaurus is Brouwer’s thesaurus (Brouwers, 1989), originally published by Lodewijk Brouwers in 1931 and also continuously updated since. This thesaurus contains around 123,000 single-word entries, including around 20,000 verbs and 16,000 adjectives and adverbs, categorised into 1,000 fairly broad synonym sets. For each of the listed thesauri, a single word can appear in multiple synonym sets due to ambiguity or homonymy. For example, the word ‘spring’ can refer to a well, a season, a coil, or a jump, and it would be listed in the synonym sets corresponding to these meanings. Since the entries also contain the POS tag of the word, it is sometimes possible to disambiguate between word senses. For example, *safe* as a noun (*a money safe*) and as an adjective (*a safe house*) have different meanings.

As an indication of the sort of information contained in a thesaurus, a selection of synonyms given for the word ‘rich’ in the discussed thesauri are listed below:

Roget’s affluent, flush, moneyed, wealthy

WordNet affluent, flush, loaded, moneyed, wealthy, comfortable, prosperous, well-fixed, well-heeled, well-off, well-situated, well-to-do

Brouwers bemiddeld (*affluent*), rentenier (*person living off interest*), rijkheid (*richness*), weelderig (*wealthy*), goudvisje (*little goldfish*; figurative), gegoedheid (*well-to-do-ness*), luxe (*luxury*)

3.4 Evaluation metrics

In developing methods for automated (language) processing, we often want to know how well the automatic method performs. A common method for doing so is by creating a Gold Standard by a native speaker or domain expert, and comparing the output of the automatic method to that Gold Standard. In this comparison, we assume that we have a number of cases to which the method has to assign a class, and we compare the automatically assigned class to the Gold Standard class on a case-by-case basis using an *evaluation metric*.

A simple metric is *accuracy*: the percentage of cases that was processed correctly. However, this is not very informative when some classes have low frequencies. For example, if we want a method to classify whether a word refers to a politician, a system that assigns the negative class to each word probably gets an accuracy of over 90% because most words do not refer to politicians.

For these reasons, in Machine Learning and Statistical NLP, the metrics *precision*, *recall*, and *F1 score* are often used (Manning and Schütze, 2002, p.267–271). These metrics are all calculated separately for each class and then averaged if needed. Precision is an indicator of how often the automatic method was correct when it assigned that class. Recall is an indicator of coverage: how many true instances of that class did the method find? The F1 score is the *harmonic average* of the two metrics, generally being closer to the lower of the two.

To calculate these metrics, consider the diagram in figure 3.2. The thick grey circle indicates which cases belong to a target class according to the Gold Standard; the thin black circle contains the cases assigned to that class by the automatic method. The intersection of these circles are the *true positives*: all cases in that region belong to the target class

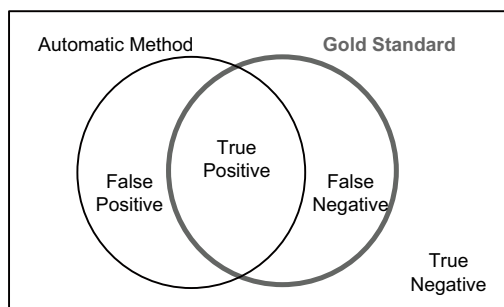


Figure 3.2: Schematic representation of classification according to an automatic method and a Gold Standard

according to both model and Gold Standard. The *false positives* are misclassified by the model as belonging to the class (errors of the first kind), and *False negatives* are misclassified by the model as not belonging to the target class (errors of the second kind). *True Negatives* are the cases outside both circles, i.e. the cases correctly classified as not belonging to the target class. *Precision* is then defined as how often the model was correct when it classified a case as belonging to the target class, and *Recall* is the percentage of cases actually belonging to the target class (according to the Gold Standard) that was found by the model. The F1 Score is defined as the harmonic average of those two measures, and can be reported either per target class or as an average over all classes.

If we count the number of cases in each of these regions, we can apply the following formulas to determine the metrics:

$$\begin{aligned} \textbf{Precision} \quad pr &= \frac{\textit{True Positive}}{\textit{True Positive} + \textit{False Positive}} \\ \textbf{Recall} \quad re &= \frac{\textit{True Positive}}{\textit{True Positive} + \textit{False Negative}} \\ \textbf{F1 Score} \quad F1 &= \frac{2 \cdot \textit{Precision} \cdot \textit{Recall}}{\textit{Precision} + \textit{Recall}} \end{aligned}$$

In the example, suppose there were 4 true positives, 2 false positives, and 6 false negatives. This gives a precision of $4/6 = .67$, and a recall of $4/10 = .4$. The F1 score is then $2 \cdot .67 \cdot .4 / (.4 + .67) = .50$. The true negatives are not used in the calculation.

3.5 Conclusion

This chapter described a number of Natural Language Processing tasks that can be used to (pre)process texts before conducting Content Analysis. In particular, it described tokenising, Part-Of-Speech (POS) tagging, lemmatising, and syntactic parsing. The chapter briefly discussed using thesauri to overcome synonymy problems. Finally, it discussed the evaluation metrics of precision, recall, and F1 score that are useful for evaluating automatic (language) processing methods by comparing the method output with a Gold Standard. This provides the background knowledge on Natural Language Processing assumed in parts II and III

CHAPTER 4

Knowledge Representation and the Semantic Web

'The words remain equal, but the meanings shift'

(De woorden blijven gelijk, maar de betekenis verschuift; *Trouw*, October 8, 1994)

Techniques from the Semantic Web, a field of Knowledge Representation, are useful for the formal representation of Semantic Networks. Such a formal representation can help in analysing, sharing, and reusing Semantic Network Data. This chapter provides some background regarding Knowledge Representation and the Semantic Web, focusing on the techniques used in part III.

4.1 Introduction

As described in chapter 2, one of the biggest challenges facing Content Analysis is the need to share, combine, and reuse data sets. This need is driven by the fact that modelling complex interactions between source, message, and sender, such as that between politics, media, and public, requires large data sets from different countries and/or time periods. Gathering Content Analysis data is expensive and time-consuming, which makes it expedient to use and reuse data to the fullest extent. One of the fundamental problems in (re)using Content Analysis data is the semantic gap between the symbols in the message and the theoretical concepts used in the research question. In particular, there is an *abstraction gap* and a *complexity gap*. The *abstraction gap* is the gap between the textual symbols that refer to concrete objects, such as Bush, and the theoretical concepts that are more abstract, such as President. The *complexity gap* refers to the fact that the variables used in a theory or model are non-structured, such as the *government responsibility frame* mentioned in section 2.2.6, but are expressed in text using complex structures, such as a concrete government actor being accused of causing a specific problem. To overcome these problems, chapter 2 argued for a Semantic Network Analysis where the text is coded as a network of concrete objects. These objects are placed in a hierarchy of concrete objects to abstract ones. This *representation* of the Semantic Network and object hierarchy can then be *queried* to answer the original research question in a systematic way.

A first approach to solving the abstraction gap might be to store the hierarchy as a simple list, for example in an Excel file, and define the patterns procedurally, for example using SPSS syntax. Such informal definitions work fine for a single study. To reuse this data in another study, however, the hierarchy will probably need to be adapted for the new research question, which often has a different context or view on the objects. Combining data sets is possible by making sure both hierarchies aggregate to the same concepts, so the aggregated networks can be combined without any problem. This approach has two main drawbacks. The first is that it is cumbersome and error-prone to align and change the hierarchy manually and to answer the research questions procedurally. The second and more fundamental objection is that it is not transparent: the meaning of the objects in the hierarchy is only specified in the documentation, and the operationalisations of variables are often difficult to deduce from a procedural description. If we want to create a system where researchers can easily reuse, share, and combine data, these problems will have to be addressed.

Knowledge Representation is a field of Artificial Intelligence that deals with the formalisation of knowledge. Knowledge Representation can

solve some of the problems described above by formalising the data and background knowledge in such a way that the computer can combine data sets and interpret them using (partially) shared background knowledge. Sowa (2000, p.xi) describes three elements of Knowledge Representation: *Ontology* to define the vocabulary of description, *Logic* to formalise the relations between items from the vocabulary, and *Computation* to put the representation to practical use.

Ontology Literally meaning the study of being, ontology is a field of philosophy that is concerned with describing the things that exist, i.e. the vocabulary used in logical statements. On the most abstract level, ontology studies how the ‘things that are’ can be divided into top-level categories such as physical versus abstract, independent versus mediating, and continuant versus occurrent. On a more concrete level, ontology also deals with more issues such as describing roles, collections, time, and space (Sowa, 2000, p.51–123). Computer scientists speak of *an ontology* (using a count noun) to refer to practical implementations of ontological systems. Such an implementation uses a logical formalism to represent the categorisations, working down from a universal Thing or Entity to concrete concepts such as persons. An example is the general-purpose ontology CYC, containing over 200,000 general terms, and over two million facts and rules about those terms expressed in predicate calculus¹ (Lenat and Guha, 1990). In Content Analysis, ontologies give us a way of formalising the background knowledge contained in the *ad hoc* hierarchy described above, allowing us to clearly define what a concept means and how it relates to other concepts. This will make it easier to combine and adapt Content Analysis data sets, and allows us to create one data set with an ontology with multiple different relations, allowing for different views on the same data as required by different research questions.

Logic Stemming from a study of valid modes of reasoning, logic refers to systems for formally describing relations between objects and valid inferences that can be made from these relations. One of the most well-known logics is propositional logic, which allows statements such as *If the sun shines, John will play tennis*, where the *sun shining* and *John playing tennis* are atomic propositions. First Order Logic (FOL) or Predicate Logic increases the expressibility by allowing predicates about objects, so *John playing tennis* can be expressed as the structured predicate *playTennis(john)* rather than as an atomic proposition. Moreover, Pred-

¹http://www.cyc.com/cyc/technology/whitepapers_dir/sksi%20_data_sheet.pdf

icate Logic allows universal and existential quantifiers over variables, allowing us to express the famous reasoning: *Every human is mortal; Socrates is a human. Thus, Socrates is mortal.* In that case, being a human is a predicate, which is said about Socrates, represented as *human(Socrates)*. Universal quantification is used to make general statements: the fact that every human is mortal is represented as “For every x , if *human(x)*, then *mortal(x)*.” Formal inference systems can describe what conclusions can be validly drawn from a set of assumptions, allowing the logic to be used for reasoning as well as representation. For Content Analysis, logic gives us the formal mechanism for representing both the background ontology and the Semantic Networks extracted from the message, and allows us to perform the aggregation, combination, and pattern identification steps in a formal, transparent manner using these inference rules.

Computation Logic prescribes which inferences are sanctioned given a set of premises, but it doesn’t prescribe *how* these inferences should be drawn algorithmically. Research in AI since the early 1960’s has yielded a powerful set of algorithms to ‘enact’ the inferences specified by a logic, and a set of techniques to implement such algorithms on modern computer hardware. This has ensured that logic is not only a theoretical tool for the analysis of reasoning, but also a practical tool to actually compute or infer the prescribed consequences. Since Content Analysis is in search of practical solutions for problems of data analysis and reuse, the computation- or implementation-related aspects of Knowledge Representation are at least as important as the (onto)logical aspects.

4.2 The Semantic Web

The Semantic Web is a relatively recent Knowledge Representation implementation (Antoniou and Van Harmelen, 2004), and is the implementation used in part III of this thesis. The intuition behind the Semantic Web comes from the limitation of the ‘conventional’ World Wide Web.

The World Wide Web is an enormous collection of linked web pages containing (true and false) information about almost everything that can be imagined. Most of this information is written in natural language and embedded in images, limiting the role of the computer to finding, retrieving, and displaying the web-site, leaving interpreting and understanding to the human user. If these web pages could somehow be understood by the computer, we would be able to conduct more intelligent searches, and the computer might be able to use the knowledge contained in the pages to perform tasks more efficiently and draw new conclusions. The *Semantic Web* refers to the vision of Tim Berners-Lee and

colleagues to create such a system (Berners-Lee et al., 2001). It assumes that web servers, next to serving human-readable HTML content, would serve information in a computer-understandable format. The client computer could then interpret this information and process it automatically to answer the user's question.

A Semantic Web of computer-readable data is not necessarily interesting to a Content Analyst. However, there are good reasons to assume that the technology developed to create this Web will be useful for Content Analysis as well. The basic task of both the Semantic Web and Content Analysis is to describe the meaning of messages in a structured manner. Both assume a large amount of data coming from different sources without a central authority, using different vocabularies representing different views on the world. Semantic Web technology needs to be able to represent and handle this large amount of heterogeneous data, and allow syntactic and semantic interoperability without resorting to centralisation and without breaking on inconsistent input, which are also the requirements for combining Content Analysis data. Because of these similarities, it is plausible that Semantic Web technology will help solve the Content Analysis problems outlined above.

The rest of this chapter will describe three core Semantic Web Technologies: the Resource Description Framework (RDF) for making statements about resources, RDF Schema (RDFS) for describing a vocabulary of resources and relations, and the Web Ontology Language (OWL) for a detailed description of formal ontologies.

4.2.1 *RDF: The Resource Description Framework*

The Resource Description Framework (RDF) is a standard representation specified by the World Wide Web Consortium (W3C) for describing documents and other resources on the Internet, creating an interconnected Semantic Web (Antoniou and Van Harmelen, 2004). Using a graph as its data model and using XML syntax to describe information, RDF allows data to be mixed, exported, and shared across different applications. Since Semantic Network Analysis has a graph-like data model, it makes sense to base a standard on an existing graph representation like RDF. This enables us to utilise existing tools and language bindings, making it easier to develop specialised tools since the elementary operations such as parsing the file format are performed for us. Using an existing standard also offers a potential to easily combine Content Analysis data with other Semantic Web data encoded in RDF, such as WordNet or the CIA factbook.

Figure 4.1 shows a very simple RDF network describing a single content statement, "Bush invites Elizabeth." The network contains only one

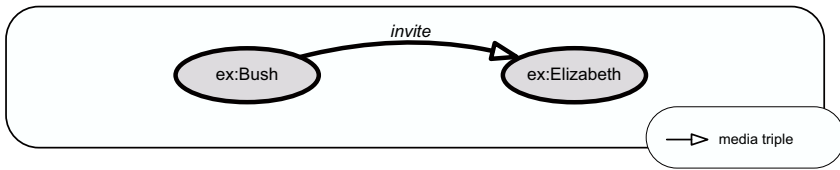


Figure 4.1: Example relation in RDF

relation, labelled *invite*, which links Bush to Elizabeth. In the terminology of graph theory, which is often used in the Semantic Web, a network is called a *graph*, the objects in the network are called *nodes*, and the relations or links between the nodes are called *edges*. In this thesis, the terms network and graph, and relation and edge, are used more or less interchangeably.

Often, we have more information than just the raw data: we also know things about the nodes and edges used in the statements. In RDF, such background knowledge is encoded within the same graph as the data itself, using different names for the relations to make the distinction between data and background statements. These extra data will generally be known beforehand and specified as a vocabulary definition or *ontology*, and combined into a single graph by the application. In our example, we could add two more relations, specifying that Bush is *presidentOf* of the United States and that Elizabeth is *queenOf* of the United Kingdom.

In order to understand the resulting graph, it is necessary to know what the nodes (such as Bush and United Kingdom) and edges (such as *presidentOf*) refer to. In particular, if we combine this graph with another graph that features a node labelled Bush, it is important to know whether these nodes refer to the same person or not. In RDF, this problem is solved by using Uniform Resource Identifiers (URIs) as the labels on nodes and edges. URI is an Internet standard for identifying resources, for example <http://content-analysis.org/voca#bush>². Since URIs contain a domain name, they can be ‘owned’ by owning the respective domain. Hence, if we use a URI in our own domain, it means ‘our’ definition of Bush, while if we use a URI in another domain we make a statement about ‘their’ resource.

Using URIs, we can link our nodes to existing vocabulary definitions, which can make our graph more interpretable to third parties. For example, we can state that Elizabeth is queen of the United King-

²URIs are similar to Uniform Resource Locators (URLs) used in browsers, but a URI does not necessarily point to an actual retrievable resource.

dom as defined by the *CIA Factbook*³ by pointing to <http://www.daml.org/2001/12/factbook/uk>, which we abbreviate to *cia:uk*. The same holds for relations. For example, we can use the relations used in the World Events Interaction Survey (WEIS; McClelland, 1976)⁴, using the WEIS *state invitation* relation to code the relation between Bush and Elizabeth. These additions result in the graph displayed in figure 4.2. Suppose this graph is read by an application that knows the WEIS and CIA Factbook, but is unaware of our vocabulary, it can still read that an actor connected to the U.S. issues an invitation to an actor connected to the U.K.

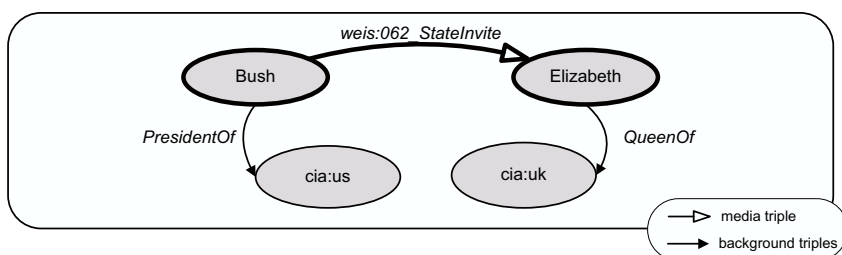


Figure 4.2: Using different namespaces in an RDF graph

4.2.2 RDFS: RDF Schema

Another way to enhance the interpretability of graphs is by using RDF Schema (RDFS). As stated above, the interpretation of RDF graphs depends critically on interpreting the vocabulary used in that graph; the only thing that is specified by the RDF standard is the graph itself. Since different data sets often use different vocabulary, a mechanism for mapping vocabularies to shared concepts is necessary to detect the greatest common factor between vocabularies, which is essential for interoperability. RDF Schema supplies this mechanism within RDF with a set of distinguished vocabulary items.

An important RDFS relation is *rdfs:subClassOf*, which indicates an ‘is-a’ relation indicating that the subject class is a specialisation of the object class. This is closely connected to the *rdf:type* relation, which specifies that a resource is an instance of a certain class. For example, we can specify that the U.K. is of type *Country*, and that *Country* is a subclass of

³<https://www.cia.gov/library/publications/the-world-factbook/index.html>. The RDF version of the factbook used in this example is located at http://simile.mit.edu/wiki/Dataset:_CIA_Factbook

⁴As far as we know, there is no RDF definition of relations used in WEIS, so the abbreviation *weis:* is by way of example only.

Actor. From this information, an *RDFS inferencer* can conclude that the U.K. is also of type Actor: it is a country, and every country is-an actor. A similar specialisation relation exists between relations: *rdfs:subPropertyOf* declares a relation to be a ‘subproperty’ of another relation. For example, if the CIA Factbook contains a definition for head-of-state, we could declare our *presidentOf* and *queenOf* to be *subPropertyOf* *cia:headOfState*. Finally, it is possible to specify the domain and range of a relation using *rdfs:domain* and *rdfs:range*. In our example, *cia:headOfState* can be defined to have domain Person and range Country, and the inferencer would automatically conclude the class memberships of U.K. and Elizabeth given the *queenOf* relation: Elizabeth is *queenOf* something, and therefore *headOfState* of something, and is therefore a Person, and hence an Actor.

Adding these background knowledge statements to the earlier graph results in the network displayed in figure 4.3, where the dotted relations are the automatic inferences made by an RDFS inferencer (to save space, the inferred relations regarding Elizabeth are not shown). From this graph, without knowing any of our vocabulary, a third party can understand that the head-of-state of the U.S. is inviting his British colleague. This shows how using RDFS to define new vocabulary in terms of existing definitions can help the interpretability of data. Moreover, we could combine this graph with other graphs about countries or people, for example about attendance at international conferences, even if that graph uses its own special vocabulary: as long as that vocabulary is also linked to (for example) the CIA factbook, we can understand the combined graph at the level of this shared vocabulary.

RDF is an important part of the Semantic Web community and has

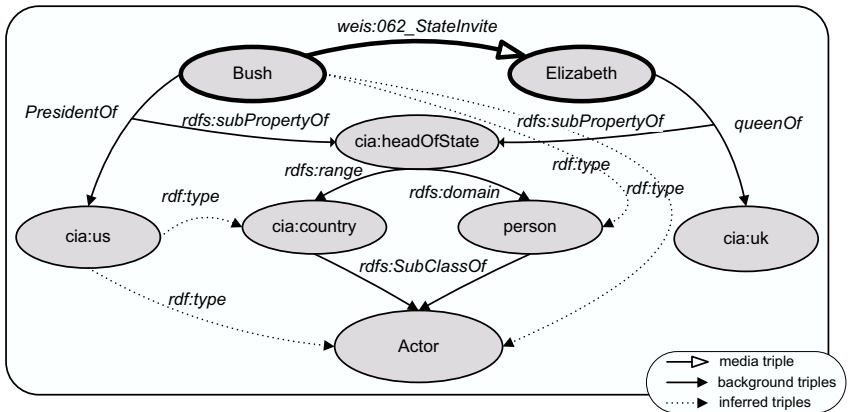


Figure 4.3: Adding background relations using RDFS

a large user base in both academy and industry. As a result, there are a number of tools such as RDF repositories, language bindings, and visualisation tools.⁵

4.2.3 OWL: *The Web Ontology Language*

As described above, RDF plus RDFS is a useful formalism for describing graphs and vocabulary using typing and is-a hierarchies. By design, RDFS has rather limited expressivity. For example, it is impossible to express negation or disjointness: you cannot state that countries cannot be persons, and RDFS will simply conclude that Bush is a Country if we accidentally make him the object (rather than the subject) of a *headOfState* relation with range Country. This lack of expressivity has strong advantages: computing inferences in RDFS graphs is relatively easy to do computationally, and RDFS graphs cannot be inconsistent. The latter point is important, since logics often collapse if there is a single inconsistency in a set of statements, which could easily arise if one combines statements from different sources.

Sometimes, it can be useful to have a more expressive language. For example, if we are creating our background ontology for the political domain, it would be good to be warned if we accidentally state that the US is the head of state of Bush, or if both Bush and Elizabeth are heads of state of the US. The Web Ontology Language OWL⁶ offers this additional expressivity as an extension of RDFS. The advantage of OWL being an extension of RDFS is that a pure RDFS inferencer will simply ignore the OWL-specific statements, so we have the best of both worlds: we can use OWL expressivity for designing and using our own knowledge base, and use the computational simplicity and robustness of RDFS when combining our graph with other data sets. In terms of expressivity, the Description Logic on which OWL is based is somewhere between RDFS and the Predicate Logic described above. In contrast to Predicate Logic, this Description Logic is computationally *decidable*, which means that a program can always compute the logical inferences from OWL statements in a finite amount of time.

A full description of OWL is beyond the scope of this chapter, and the techniques described in this thesis do not directly use OWL. However,

⁵For example, see <http://planetrdf.com/guide#sec-tools> for a list of 126 RDF tools

⁶“The natural acronym for Web Ontology Language would be WOL instead of OWL. Although the character Owl from Winnie the Pooh wrote his name WOL, the acronym OWL was proposed without reference to that character [...]. And, to quote Guus Schreiber, ‘Why not be inconsistent in at least one aspect of a language which is all about consistency?’, from http://en.wikipedia.org/wiki/Web_Ontology_Language#The_acronym

the extra expressivity of OWL can be useful for expanding the representation of background knowledge, as described in section 8.5. For this reason, we shall briefly describe some of the things expressible in OWL that cannot be expressed using RDF(S):

Disjointness In OWL, two classes can be defined *disjoint*, which means that an instance cannot be a member of both classes.

Cardinality The relation between two classes can contain *cardinality restrictions*. For example, a country has only one capital or head of state, and a politician is a member of exactly one political party.

Transitivity Relations can be defined as *transitive*: if *partOf* is transitive, and the U.K. is *partOf* North West Europe, and North West Europe is *partOf* Europe, then the U.K. is *partOf* Europe. Note that although the *rdfs:subClassOf* and *rdfs:subPropertyOf* relations are transitive, RDFS does not have a mechanism for defining other transitive relations.

4.3 The Semantic Web as a Knowledge Representation framework

As defined above, Knowledge Representation is characterised as the combination of ontology, logic, and computation. Describing RDF, RDFS, and OWL in these terms shows how these techniques offer an increasingly sophisticated Knowledge Representation Framework:

Ontology As defined above, ontology is the description of the things that are, i.e. the vocabulary. The semantic web does not assume or provide a certain definition of vocabulary. RDFS and OWL do provide mechanisms to define vocabulary, for example using the type, subclass, and subproperty descriptions mentioned above.

Logic The logical foundation of RDF is a very simple conjunctive-existential logic with only binary predicates. It is possible to link resources in statements and create anonymous resources (blank nodes), but it is impossible to define disjunction or general statements using universal quantifiers. RDFS adds to this a fixed set of axioms with universal quantification, for example for the transitivity of subclass definitions and the type inference for domain and range. OWL has a more expressive logical foundation in Description Logics, allowing for the expression of transitivity, disjointness, cardinality, etc. as discussed above.

Computation RDF does not have an inference mechanism, but RDF query languages such as SeRQL or SPARQL do allow for the definition of

patterns that can be matched to an RDF graph, essentially a form of model checking. RDFS specifies a fixed number of inference rules; the closure of these rules is often computed at the moment the data is loaded. Since RDFS contains neither negation nor disjunction, RDFS graphs and their closure grow monotonically as more data is added, and computed inferences always remain valid. OWL has more sophisticated reasoning capacities, and a number of OWL reasoners exist. Such reasoners can determine things such as whether an ontology expressed in OWL is consistent and whether any description, such as a conjunction of two classes, is satisfiable. OWL reasoning is not monotonic, so it is not possible to compute all inference in advance. Moreover, OWL reasoning is generally more computationally intensive than computing the RDFS closure and RDF(S) querying.

4.4 Conclusion

This chapter provided a high-level overview of Knowledge Representation, and described three related Semantic Web languages: RDF, RDFS, and OWL. RDF(S) is the Knowledge Representation framework used in part III of this thesis. Taken together, chapters 2 – 4 provide the background knowledge assumed in the remainder of this thesis.

Part II

Extracting Semantic Networks

CHAPTER 5

Extracting Associative Frames using Co-occurrence

'Muslims condemn terrorism'

(Moslims hekelen terrorisme; *Dagblad van het Noorden*, July 7, 2007)

'Muslims on TV, No Terror in Sight'

(*The New York Times*, November 11, 2007)

If two concepts constantly occur together, for example Muslims and terrorism, or immigrants and crime, they can be seen as associated. Even if the two are not related or are even explicitly dissociated, this tells us something about the worldview of the source of the messages containing both concepts, and it can cause the receiver of those messages to relate the two concepts. This chapter presents a Associative Framing, a method for determining and interpreting such associations.

This chapter is an edited and expanded version of:
Nel Ruigrok and Wouter van Atteveldt (2007), *Global Angling with a Local Angle: How U.S., British, and Dutch Newspapers Frame Global and Local Terrorist Attacks*, *Harvard International Journal of Press/Politics* 12:68–90

5.1 Introduction

In Semantic Network Analysis, we are interested in extracting relations between concepts. Concepts can be actors, issues, or abstract values, and can be expressed in a text using names, common nouns, prepositions, or other words. This chapter describes how the associations between concepts can be measured automatically and how these associations can be interpreted within communication theory. Chapter 6 expands on this by using syntactic patterns to differentiate between the source, subject, and object of relations, while chapter 7 uses Sentiment Analysis techniques to distinguish between positive and negative relations. Taken together, these three chapters describe the techniques needed to automate Semantic Network Analysis.

This chapter describes a method for automatically determining the association between concepts in texts. Determining the associations between concepts requires identifying references to these concepts in text. Automatically determining whether a word refers to one of the concepts we are interested in is not a trivial task. Actors, especially persons, are relatively easy to extract because, at least in journalistic or other professional texts, they are referred to by their full name the first time they are mentioned. Issues are more difficult to recognise automatically, as many issues do not have a single name, and are sometimes described in a phrase rather than a word, such as ‘the gap between citizens and politics’. Abstract values and emotions suffer from the same problem as issues: there are many synonyms for ‘angry’ or ‘good.’ To an extent, this problem can be tackled using word lists in what Krippendorff (2004, p.283–289) calls the “dictionary approach” to automatic content analysis. This approach has been used in a number of studies and has yielded good results (e.g. Stone et al., 1962; Pennebaker et al., 2001; Fan, 1988; Hart, 1985).

If we know in which parts of a text each concept occurs, we can also determine their *co-occurrence*. The fact that two concepts co-occur in a document is interesting in itself: suppose a source mentions terrorism in every sentence in which Muslims are also mentioned. Even if we do not know the exact relation that the source draws between the two concepts (if any), we know that the source apparently associates Muslims with terrorism, and this might have an effect on readers if this association is strongly present in many (media) messages. Methods based on automatic co-occurrence have long been used in Content Analysis. For example, Iker and Harway (1969) created a program to perform factor analysis on words co-occurring within fixed word windows; Woelfel (1993) created a Neural Network of weights between words that increase as words co-occur; Diesner and Carley (2004), discussed earlier, built a network of

ties between words or concepts based on co-occurrence in a fixed window. These approaches generally suffer from a number of problems. Many of them analyse the co-occurrence of words rather than concepts, rendering very low-level linguistic results that are difficult to relate to theoretical concepts. A related problem is that most of these approaches do not have a foundation in communication theory or processes, and work inductively to distill patterns from the text. Although such patterns can be very interesting, especially in qualitative or exploratory analysis, they are not suited for developing models or testing theories of communication structure or effects. Finally, the operationalisation of the co-occurrence is often in terms of weights, such as neuron weights or link strengths. These quantities are very difficult to interpret, as a weight of .1 in a neural network does not have an innate meaning in the context of the communication.

This chapter describes a method for co-occurrence analysis called Associative Framing. As the name suggests, this method is grounded in frame analysis, especially the cognitive frames as described by Tversky (1977), and the emphasis frames described by Druckman (2004). Associative Framing uses a dictionary approach to recognise concepts, and calculates the co-occurrence of these concepts. This method creates a network of concepts rather than of words. Associative Framing operationalises associations as the conditional reading (or ‘writing’) chance: the association of concept A with concept B is the probability that a random message is about concept B given that that message is about concept A. This gives a simple substantive interpretation of the association strength: if association of Muslims with immigration is .4, this means that if I draw a random message from the messages about Muslims, there is a 40% chance that that message will also be about immigration. In line with the observations on mental similarity made by Tversky (1977), such associations are asymmetric: one might strongly associate Muslims with immigration while not directly associating immigration with Muslims, or the other way around.

Research questions that require knowing the emphasis on or presence of a relation can be directly answered using Associative Framing. If a question supposes a specific relation, such as a causal or negative relation, Associative Framing can still be a useful tool for explorative analysis. It can show in which sources or time periods the relevant associations are strong or varying, indicating periods that would be interesting to investigate with more sophisticated methods.

In contrast to chapters 6 and 7, this chapter does not present an evaluation by comparing the extracted Semantic Networks to manual codings. Consequently, the main contribution of this chapter is not in presenting a novel way to measure associations, but lies rather in the interpreta-

tions of such associations in terms of recent literature in communication science. The literature contains a number of studies that perform such evaluations and report good results, giving us confidence that it is possible to reliably measure the occurrence of concepts, and hence their co-occurrence. For example, Rosenberg et al. (1990) report that the Harvard Sociopsychological Dictionary used in the General Inquirer (Stone et al., 1966) outperforms manual phrase-based content analysis on diagnosing authors. Pennebaker and Francis (1996) report correlations of around .7 on recognising emotions and cognitive strategies using the Linguistic Inquiry and Word Count (Pennebaker et al., 2001). Fan reports an accuracy of 75%–90% for English text (1988) and a correlation with manual coding of .7–.9 for German texts (2001) using lists of words and specific combinations of words.

In the next two sections, the theoretical foundation and operationalisation of associative framing are described in more detail. This is followed by a concrete use case: a study of associative frames in newspaper coverage of terrorist attacks. We will investigate how the associative framing of Muslims changes after a terrorist attack, and to what extent these attacks are globalised or localised in the press. This use case shows that associative framing is a useful technique that yields interesting results that can be readily interpreted theoretically and used for testing substantive hypotheses. Moreover, since the study is based on over 140,000 newspaper articles from three countries spanning 5 years, it showcases how automatic content analysis can quickly analyse very large data sets.

5.2 Frames as Associations

In the 1990s framing theory gained an important place in the field of communication research. Entman (1993, p.52) defined framing as selecting “aspects of a perceived reality and [making] them more salient in a communicating text, in such a way as to promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation for the item described.” The definition itself shows already the multi-faceted nature of framing research. It is about selection, salience, and recommendation, including not only the communicator but also the audience. As Entman (1993) points out, there are at least four locations of framing that can be studied: the communicator, the text, the receiver and the culture. The research on framing so far shows a division between studies examining media frames and research into audience frames (Cappella and Jamieson, 1997; Entman, 1993; Scheufele, 1999). The former branch of research focuses on how issues are presented in

the news (Norris, 1995; Patterson, 1993; Semetko and Valkenburg, 2000) while the latter focuses on how individuals perceive and interpret issues presented to them (Domke et al., 1998; Nelson et al., 1997; Price et al., 1997; Rhee, 1997; Valkenburg et al., 1999). A combination of these branches is found in a few studies examining both media frames and the effects of these frames on the public (e.g. Cappella and Jamieson, 1997; Iyengar, 1991; Neuman et al., 1992). Within framing research there are two critical questions: "What are frames?" and "How are frames transferred between media and audience?"

5.2.1 *News Frames, Equivalency Frames, and Emphasis Frames*

With respect to the question of what frames actually are, a distinction is made between equivalency frames and emphasis frames. 'Equivalency Frames' present an issue in different ways with "the use of different, but logically equivalent, words or phrases" (Druckman, 2001, p.228). In experiments, researchers found systematic changes in audience preference when the same problem was presented in different wordings, such as rescuing some versus sacrificing others (Quattrone and Tversky, 1988; Tversky and Kahneman, 1981). Emphasis frames, later called "issue framing" (Druckman, 2004), on the other hand, highlight a particular "subset of potentially relevant considerations" (Druckman, 2004, p.672). In line with Entman's definition, issue framing can be defined as a process of selecting and emphasising certain aspects of an issue on the basis of which the audience can evaluate the issue described or the protagonists associated with the issues.

5.2.2 *Linear versus Interactive Frame Setting*

The second question mentioned above — how are frames transferred from the media to the audience? — also leads to a number of different hypotheses. Some researchers consider the transfer of salience a linear process, straight from the sender into the audience (Eagley and Chaiken, 1998; Zaller, 1992, 1994). Research in this field is based on the mathematical model of a one-way, linear transmission of messages (Shannon, 1948). Other researchers suggest a more complex situation in which meanings are produced and exchanged between the sender, the receiver and the larger community in which they operate (Nelson et al., 1997). In other words, the framing process can be regarded as an interaction between message content and the interpreter's social knowledge. This interaction process leads to a construction of a mental model as a resulting state of interpretation (Rhee, 1997). Besides the creation of these mental models, the framing process can trigger a mental model or frame that already

exists within the receiver's perception. Graber (1988) describes the way people use schematic thinking to handle information. They extract only those limited amounts of information from news stories that they consider important for incorporation into their schemata. Snow and Benford (1988) state in this respect that media frames and audience frames interact through 'frame resonance,' where media frames that correspond to existing frames 'resonate' and are more effective than non-corresponding frames (see also Snow et al., 1986).

5.2.3 *Frames as Cognitive Networks*

The construction of mental models, schemata or frames is a central part of the cognitive approach to framing. Grounded in cognitive psychology, this approach uses the associative network model of human memory (Collins and Quillian, 1969), proposing that the concepts in semantic memory are represented as nodes in a complex hierarchical network. Each concept in the network is directly related to other concepts. Minsky (1975) linked this view to framing when he defined a frame as a structure containing various pieces of information. These discursive or mental structures are closely related to the description of a schema, which is "a cognitive structure that represents knowledge about a concept or type of stimulus, including its attributes and the relation among those attributes" (Fiske and Taylor, 1991, p.98). These cognitive structures are based on prior knowledge (Fiske and Linville, 1980).

As discussed above, the study of framing contains many perspectives and research lines. We perceive, however, a common denominator in that many studies base the idea of a frame on associations, either between concepts, between concepts and attributes, or on more complex networks of concepts. In this chapter, therefore, we will focus on what we call associative framing. Associative frames consist of patterns of relations between concepts; these relations have a weight and direction (i.e. they are asymmetric) but do not have a sign or specific meaning. In a wider context, these associative frames can be seen as a generalisation of the model used by second-level Agenda Setting (McCombs and Ghanem, 2001; McCombs and Estrada, 1997), which extends agenda setting theory by assuming the transfer of salience of object-attribute links as well as the original transfer of salience of issues (cf. Dearing and Rogers, 1996; McCombs and Shaw, 1972).¹ Contrary to (second-level) Agenda Setting, we do not hypothesise a linear, 'hypodermic' transfer of associations or frames. Rather, we are interested in examining different possible interactions and transfer hypotheses. These frames refer to the earlier described schemata of interpretation (Goffman, 1974), and

¹See section 2.2.1 on page 18

the main associations in a message can be seen as its “central organising idea” (Gamson and Modigliani, 1987). We conjecture that these media frames interact with associative networks in the receiver as described by Collins and Quillian (1969), which can thus be termed “associative audience frames.”

5.3 A Probabilistic Model of Associative Framing

In associative framing we assume that media messages can be reduced to *contexts* containing atomic *events* which have a certain probability of occurring, and within which co-occurrence is meaningful. More concretely, we assume that there is a document size, such as a sentence, paragraph, document, or newspaper, for which we can measure the occurrence of our target concepts and within which we want to know whether they co-occur. This section operationalises the visibility or reading chance of a concept as the marginal probability that a concept occurs: if we pick a random document, how high is the chance that that concept occurs in that document? The association between two concepts is operationalised as the conditional probability of one concept occurring given that the other concept occurred: if we pick a random document, and that document contains the primary concept, how high is the chance of that document also containing the associated concept?

The occurrence of concepts is measured using synonyms or keywords as indicators of the target concept, along with disambiguating conditions. An example of this is requiring the phrase “President Bush” to occur in a document as a condition of accepting the word ‘Bush’ as an indicator of the concept. This would prevent articles about George H.W. Bush or Governor Jeb Bush to be mistakenly counted while still allowing the use of just ‘Bush’ as an indicator in other sentences within the article. For example, in the two sentences from a constructed article, we will find the following keyword counts, assuming sensible keywords for the target concepts Bush, Immigration, and American Values:

| Sentence | Bush | Immigration | Values |
|--|------|-------------|--------|
| Bush Campaigns for Immigration Reform | 1 | 1 | 0 |
| New arrivals to this country must adopt American values and learn English, President Bush said Wednesday | 1 | 1 | 2 |

To transform these keyword counts to probabilities, we need a function from $[0, \infty >]$ to $[0, 1]$. We would wish this function to increase monotonically from zero towards one, and the probability of a concept with two synonyms should equal the probability of encountering either of the

two synonyms if they were separate keywords. A set of functions satisfying these constraints is given in equation 5.1, where c and m stand for the concept and message being investigated, and the parameter $\frac{1}{b}$ is the probability of a concept encountered only once.

$$p(c|m) = 1 - \left(1 - \frac{1}{b}\right)^{\text{count}(c,m)} \quad (5.1)$$

Using this to assign probabilities to the sentences above, taking $\frac{1}{b} = 50\%$ yields:

| Sentence | Bush | | Immigration | | Values | |
|--------------------|-------|----------|-------------|----------|--------|----------|
| | Freq. | $p(c m)$ | Freq. | $p(c m)$ | Freq. | $p(c m)$ |
| Bush Campaigns ... | 1 | .5 | 1 | .5 | 0 | .0 |
| New arrivals ... | 1 | .5 | 1 | .5 | 2 | .75 |

In the first document, the reading chance of both Bush and Immigration is given as .5 or 50%. Values does not occur in that document and has a reading chance of 0%. In the second document, the concept values is mentioned twice and hence has a reading chance of 75%. In essence, the formula above translates this into a term by document matrix containing probabilities as cell values. On this matrix we define the two measures comprising associative frames, visibility and association, as the marginal and conditional probabilities.

Visibility is the marginal probability of a concept. In other words, visibility is the chance that if a single message is received from the set of messages, that message will contain that concept. This probability is based on the chance of a concept occurring in a message and the chance of receiving that message. This latter probability could be based on message properties, such as position in a newspaper and reach of that newspaper for articles, but also characteristics of the potential receiver of the message in individual-level analyses. In a formula, where $p(m)$ is the chance of receiving a message (the normalised weight of a message):

$$\text{Visibility}(c) = p(c) = \sum_m p(m)p(c|m) \quad (5.2)$$

The association between two concepts, called the *base* and *target* concepts, is defined as the conditional probability of a message containing the *target* concept given that that message contains the *base* concept. This corresponds to the following formula, where c_t is the target concept and c_b is the base concept.

$$ass(c_b \rightarrow c_t) = \frac{\sum_m p(m) \cdot p(c_b|m) \cdot p(c_t|m)}{Visibility(c_b)} \quad (5.3)$$

In our two-sentence example, this leads to the associations below:

| Concept | Visibility | Association with | | |
|-------------|------------|------------------|-------------|--------|
| | | Bush | Immigration | Values |
| Bush | .50 | — | .50 | .19 |
| Immigration | .50 | .50 | — | .19 |
| Values | .38 | .38 | .38 | — |

The association between Bush and Values is 0.19, meaning that the probability of encountering values in a document containing Bush is 19%. The reverse association is .38, so the chance that a document in which values occurs also contains Bush is 38%. The same associations exist between values and immigration. Both the association of Bush with immigration and the reverse are .5, so the chance of one occurring in a document given that the other occurs is 50%.

5.3.1 Motivation and Relation to other Methods

In the preceding section, we proposed using marginal and conditional probability to describe association patterns. This section will give a number of reasons why we think this is a good representation, and also discuss how it relates with existing association measures.

The simplest alternative to probabilities is using the raw keyword co-occurrence counts, such as in Automap (Diesner and Carley, 2004). These numbers, however, are very difficult to compare between data sets and even concepts, and also suffer from strong autocorrelation between different edges from the same node (cf. Krackhardt, 1987). Moreover, outliers such as very long documents can strongly influence these counts, necessitating a normalisation using local and global weighting in studies such as Deerwester et al. (1990). The resulting normalised numbers are hard to interpret. Probabilities do not suffer from these problems, having a very clear substantive interpretation as (conditional) reading chance.

Associative framing using marginal and conditional probabilities is a direct extension of associations using simple dichotomous occurrence of concepts Tversky (cf. 1977). This is similar to the crisp versus partial set membership such as used by Fuzzy Logics. Since the method is a direct extension, it is valid to use deterministic concept occurrence, i.e.

a concept is either present or not, while still remaining within the associative framework. In that case, visibility reduces to the (weighted) proportion of documents mentioning a document, and associations are the proportion of documents containing the base concept that also contain the target.

Another advantage of using probabilities is that Statistical Natural Language Processing methods generally return a probability distribution over possible outcomes or at least a confidence estimation of the best outcome. Using a probabilistic graph representation allows the seamless integration of such qualified information.

Additionally, probability calculus is a well-established field of mathematics, and many other methods are built on its foundations. Hence, probability calculus gives us a natural way to extend the models presented here by making the (conditional) probability models more complicated, some examples of which will be given below. Another possibility is to use a generative model for the media producer, viewing media content as something that is produced based on an internal state of the media producer. This could be a natural way to estimate confidence of media data and might be a useful way to model theories on media production. Although all of this would require substantial theoretical and methodological work before yielding results, building the graph representation on a probabilistic foundation makes it easier to use such established methods, and might be a first step in fruitful interdisciplinary research.

Another important choice was to use an asymmetric association measure. The main argument for that is substantive: currently John Bolton only appears in the news in articles on the United Nations, making the association from Bolton to UN very strong, while the reverse association is fairly weak. Tversky (1977) also notes that the semantic distance of one concept to another is often different from its inverse. For example, they find that Hospital is more similar to Building than the other way around. This choice rules out many existing symmetric metrics such as correlation and the cosine distance often used in Information Retrieval systems.

As a final note we would like to state that our association metric is fairly similar to metrics like cosine distance, correlation, and regression. All these metrics are based on the dot product of the variable vectors with some normalisation. Cosine distance and correlation both normalise on the length (standard deviation) of both vectors, while correlation is also based on mean centred variables. Regression coefficients model only on the standard deviation of the target variable, making it equivalent to our metric except for the centring. A statistic often used in linguistic co-occurrence analysis, the χ^2 test, is a measure of the signifi-

cance rather than the strength of the association (Manning and Schütze, 2002), making it less relevant for describing media frames, although it could be used to test whether such found frames deviate from some prior expected distribution.

5.4 Use Case: Terrorism in the News

This section presents a concrete use case to show the applicability of associative framing to large-scale communication research. In particular, we will investigate the extent to which terrorist attacks are framed as a local event or a global event, and how the framing of Muslims changes after terrorist attacks.

Following the attacks on the World Trade Centre in New York, there has been an accelerating trend towards a global polarisation of society into 'Western' cultures on the one hand and Islam on the other, epitomised in the 'War on Terror' against the 'Axis of Evil'. Within this process we can see an increasingly important role for the media. The fall of the Berlin Wall has loosened the coherence and narrative power of the safe 'Cold War' frame that helped journalists clearly distinguish 'us' from 'them' for several decades. After 9/11, journalists enthusiastically embraced the new framework of the 'War on Terror' in order to interpret the 'friends' and 'enemies' of a state, easily expanding the notion of 'enemy' to include all Muslims, both in the Middle East and the West (see also Lippmann, 1922; Norris et al., 2003).

In the context of these events, a school of hyperglobalisers argue that the national public sphere has been replaced by McLuhan's (1960) 'global village,' in which citizens are informed of global issues and receive the same information around the world. According to this view, the media has transcended the nation and now functions on a global level (Hjarvard, 2001, p.20). Another school of thought argues that, although we learn a lot about the world around us, this is still from a domestic point of view. Media continue to preserve traditional culture despite internationalising forces (Hjarvard, 2001, p.22). Especially after 9/11, familiar local, domestic contexts are being used to integrate the global event in increasingly local discourses — in a process of 'regionalisation' (Volkmmer, 2002, p.239).

5.4.1 *Globalisation, Domestication, and Glocalisation*

A news item must be meaningful for the audience before it can become news. Proximity, as Galtung and Ruge (1965) labelled it in their seminal study into news values, makes an event more meaningful for a country

and its audiences. Researchers found that proximity affects both news selection as well as the coverage and framing of news items (Entman, 1991; Grundmann et al., 2000; Kaid et al., 1993). Gurevitch et al. (1991, p.207) concluded that in order to be judged newsworthy, an event has to be anchored "in a narrative framework that is already familiar to and recognisable by newsmen as well as by audiences." Within the discussion about the globalisation of the news Cook (1994) illustrates tellingly the notion of 'domestication' of news by considering a French and an American report of a diplomatic event that took place before the Gulf war. He states that in the French report "the world was first globally constructed, then ideologically constructed. By contrast, for the American broadcast, the world was first domestically constructed, then institutionally constructed" (Cook, 1994, p.105).

In contrast to this localisation, hyperglobalisers argue that we are entering one 'global village,' replacing the national public spheres (Clausen, 2004; Volkmer, 2002). Citizens are informed of global issues and receive the same information around the world. According to this view, the media have transcended nation states and exist on a global level (Hjarvard, 2001, p.20). The new term glocalisation was coined to indicate the synthesis of domestication and globalisation. Robertson (1995) argues that this term, which originated in the Japanese business sector to describe a global outlook adapted to local conditions, can serve as a more precise term than globalisation. Ritzer (2004, p.77) defines glocalisation as 'the interpenetration of the global and local resulting in unique outcomes in different geographic areas.' Lee et al. (2002) studied glocalisation in media coverage while looking at the handover of Hong Kong from Britain to China. The researchers demonstrated how a global event was glocalised by media in different countries, how the notion of glocalisation 'captures the global media production of the local and the local media productions of the global' Lee et al. (2002, p.53).

Studies into the news coverage of terrorism support the observation of a domestic culture filter (Simmons and Lowry, 1990; Van Belle, 2000; Winn, 1994). Winn (1994) examined the news coverage of terrorist events in *The New York Times* and the three major US networks from 1972 to 1980. The researchers found that the location of the event and the nationality of the victims were both significant, especially for television news (Winn, 1994, p.78).

A study into the news coverage of a Swedish newspaper about terrorist attacks in Kenya and Tanzania in 1998, and Madrid in 2004, showed a significant difference in the amount of attention paid to both events (Persson, 2004). Madrid received far more attention than the African countries. Moreover, the study reveals differences in the interpretation of the events. Kenya and Tanzania were framed as a tragedy and crime,

while Madrid was a moral outrage everyone should care about; terrorism was labelled as something 'new,' 'Islamic' and 'global', increasing the association between Islam and terrorism. The description of the causes of terrorism remained very limited in the news (Persson, 2004, p.36).

Schaefer (2003), in his research into the framing of the US embassy bombings and September 11 attacks in African and US newspapers, found differences between the American and African media. Schaefer (2003, p.110) concludes: "Because journalists searched for local angles and reflect the biases in their societies, American and African newspapers were ethnocentric in putting their own concerns and structural frames first and not challenging what they already thought about the other."

Clausen (2003, p.113) found the same domestication of the news, researching the coverage of the first commemoration of 9/11: "Stories were framed, angled, geared, and worded to suit the emotional and cognitive framework of audiences at home." There is no such thing as 'global' news congruent in theme, content, and meaning.

Following the Official Slant

An important aspect when looking at the domestication of the news is a focus on official sources. Numerous studies show the importance of official sources especially with regard to foreign policy and security issues (Bennett and Paletz, 1994; Hallin, 1986; O'Heffernan, 1991). Ryan (2004) came to the same conclusion studying the editorials of 10 US newspapers after 9/11. Bush's 'war on terror' frame was accepted without any counter arguments and even reinforced by a selective choice of sources. According to Norris et al. (2003, p.4), 9/11 forms a 'symbolising critical culture shift in the predominant news frame used by the American mass media for understanding issues of national security etc.' Research into news coverage on CNN right after the attack support these findings (Reynolds and Barnett, 2003). Relying heavily on official sources, CNN's coverage showed a clear, dominant frame consisting of three thematic clusters that involved war and military response, American unity, and justification. Keywords within the war and military response and justification cluster included statements referring to the United States more frequently as "America" instead of "the United States"; using the words "war" and "an act of war" to describe the attack, labelling the event as "horrific" and "unbelievable." In the coverage, words such as "cowards" and "madmen" were used to describe the terrorists. Moreover, journalists made atypical references to "God" and the need to "pray" or for "prayer"; and used words like "freedom," "justice," and "liberty" as simple descriptors of America and its ideals. Finally, symbolic compar-

isions to Pearl Harbor were made. Political leaders, both Democratic and Republican, were unified in their support for the president, their usage of catchphrases such as "God Bless America," and their emphasis on the American way of life, giving the impression that the entire country was unified. Not once did anyone, source or journalist, suggest that an option other than supporting the president would exist.

Framing Islam before and after 9/11

As long ago as the early 80s, Said in his writing about Islam in America states that the media have 'portrayed it [Islam], characterised it, analysed it, given courses on it...based on far from objective material' (Said, 1981, p.x-xi). Other research shows that this bias is not confined to the US. The Australian media's reporting on Muslims and Islam during the first Gulf War focused on themes such as terrorism and disloyalty. Richardson (2001, p.239) shows the same tendencies in the British broadsheet press. British Muslims are largely excluded from British broadsheet coverage and if included, Muslims are presented within a negative context, such as violence, terrorism, fundamentalism, or villainy.

After 9/11, several new studies investigated the portrayal of Muslims in the media and produced different results. Nacos and Torres-Reyna (2003, p.151) found a shift from "limited and stereotypical coverage in the pre-9/11 period to a more comprehensive, inclusive, and less stereotypical news presentation." The researchers not only found that more access was granted to Muslims, but also that there was a difference in the content of the news. Whereas the media associated Muslims with negative and stereotypical topics before 9/11, they focused on a bigger range of topics post-9/11. Moreover, the researchers found a shift from the episodic framing towards more thematic framing patterns. As one expert in the field pointed out as early as 1981, the cultures and peoples of the Middle East "are not easily explained in quick two-minute network news stories" (Shaheen, 1981).

Other researchers, however, argue that after the initial period of disorientation, news coverage recaptured the old frames with which they shaped the news about Arabs, associating them with violence, terror, and Islam (Karim 2002, p. 12; see also Persson 2004). These findings are in line with the research of Brosius and Eps (1995), who studied the impact of four key events on news selection in the case of violence against aliens and asylum seekers in Germany. They found that the amount, as well as the shape, of coverage increased significantly after these key events. According to Brosius and Eps (1995, p.407), key events have a prototyping quality that "point to an interactionist view of news selection. Both reality, represented by key events, and journalists' schemata

and routines work together in creating the media's picture of the world."

Hypotheses

The literature has shown that localisation of the event affects both news selection and presentation. In other words, all the news is framed from a local perspective. On the other hand, we expect journalists in the globalised world to present the local events within a wider, global framework:

H1 All news is local: global news is linked to the audience's world.

H2 Local news is globalised.

With respect to newspaper journalists, we expect that their focus will be heavily patriotic: they will support government statements and have a "rally around the flag function":

H3 The local media will perform a "rally around the flag" role.

When looking at the portrayal of Muslims before and after 9/11, we found contesting opinions. On the one hand, researchers argue there is more positive coverage of Muslims, while other researchers argue that the key event just emphasises already existing stereotypes:

H4a Muslims are portrayed more negatively after a local event.

H4b Muslims are connected with terrorism more strongly after a local event.

5.4.2 Use Case: Data and Concepts

This use case focuses on the coverage of four events in three countries. The four events are (1) the attacks on the World Trade Centre on September 11, 2001; (2) the bombing on the Madrid subway on March 11, 2003; (3) the murder of the film producer Theo van Gogh in the Netherlands on November 2, 2004; (4) and the bombings on the London subway on July 7, 2005. The three countries are the United States, Great Britain, and the Netherlands.

In each of these countries, we analysed two newspapers, one quality broadsheet newspaper and one popular newspaper. For the US, these will be *The Washington Post* and the *USA Today*; for the UK, *The Guardian* and *The Sun*; and for the Netherlands, *de Volkskrant* and *De Telegraaf*. From these newspapers, we analysed all articles between January 1st, 2001, and January 1st, 2006 that contain one or more of the keywords indicating one of the four key events, terror, or Islam — resulting in a total of 140,456 articles.

In addition to the four key events, terror, and Islam, we define a number of concepts and related keywords to measure the associations with: actors in the four countries, here confined to members of the executive and legislative branches of the government; reference events in the four countries and the broader world, here Pearl Harbor, the African Embassy bombings; the 1993 World Trade Centre bombing in the US; the IRA bombings for the UK; the ETA for Spain; the Moluccan train hijackings for the Netherlands; and the Palestinian Intifada and the Kurdish independence movement in Turkey as global related events. Finally, we define three categories of symbolic or emotional associations: positive (innocent, peace, freedom), negative (violence, madmen, fundamentalist), and patriotic (unity and support but also government jargon such as 'war on terror').

We decided to use the article as the unit of co-occurrence because we are not just interested in direct connections between objects, in which case paragraphs or sentences would be a more meaningful unit. Rather, we are interested in seeing associations in a whole line of argument, which necessitates the larger associative context. In other words, we are interested in the general organizing themes of an article rather than the manifest connections made in sentences (Goffman, 1974, cf).

5.4.3 Results

For the first four hypotheses we are interested in the framing of an event and only consider the two months after each event. For the last hypotheses we are interested in the long-term effect of events on the framing of a group, so we consider the whole data set. Actors and media are considered local to an event if they are officials, institutions or residents of the country in which the event took place. Analogously, an event is local to a medium if it took place in the country that that medium is produced in, and global if not.

Attention paid to global and local events

Before moving on to the hypotheses described earlier, figure 5.1 gives an overview of the number of articles per newspaper over time. A first striking aspect in the graph is the similarity between the different newspapers. They all show the same pattern in their news coverage of terrorism. Differences, however, are found with respect to the quantity of the news. By far the most news coverage is found in *The Washington Post*, with 50,000 out of the total number of 140,000 articles. The other quality newspapers investigated, *The Guardian* and *de Volkskrant*, follow at a distance with about 25,000 articles each. The popular press gives the in-

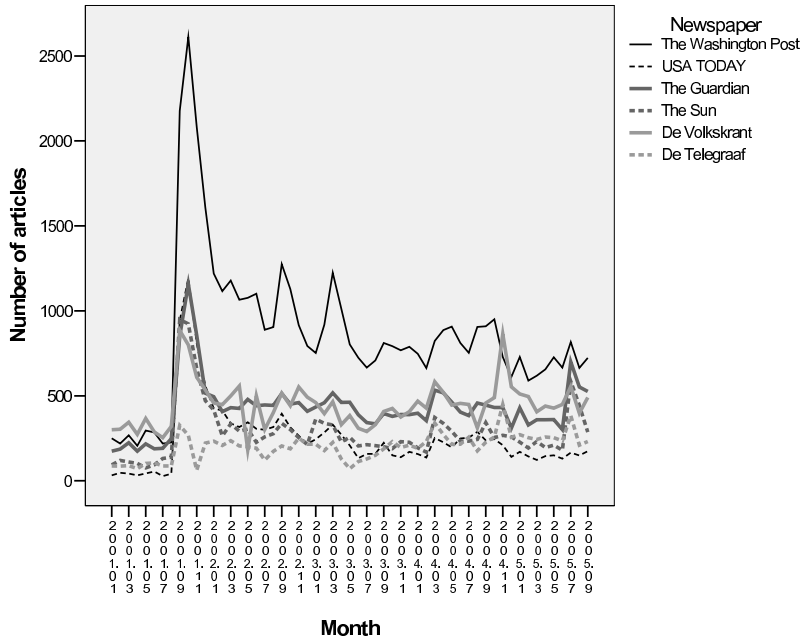


Figure 5.1: Number of Articles in the Sample per Newspaper per Month
(N = 140,456)

investigated events even less coverage, with 11,000 to 16,000 articles each. We can clearly see that most news is found around the WTC attacks in September 2001. One year later, the first commemoration of the events caused another peak in the attention over time, followed by a summit of attention in the building-up phase to the war in Iraq, in April 2003. After that we see a decline of attention in *The Washington Post* and *USA Today*, with a small revival of attention when the bombs in Madrid exploded. We see the same pattern for both the Dutch and British newspapers, although with smaller differences in the amount of attention. The British newspapers follow the same line as the other newspapers, with a revival in July 2005 with the bombings in London. In both *The Sun* and *The Guardian* we see an increase of attention in this month.

Globalization versus localization

After the news value of geographical proximity, we now focus on the extent to which the events were either localized or globalized, as stated in hypotheses 1 and 2. The absolute localization can be seen as the as-

sociation between an event and local actors or reference events. This is compared to the international association with these actors, to make sure that it is a local association rather than a local actor playing a global role. It is also compared to the association with global actors and reference events to obtain a relative measure.

Analogous to the domestication of global events, the globalization of local events is considered to be the association of that event with global actors and other global events of the past by the local newspaper. Figure

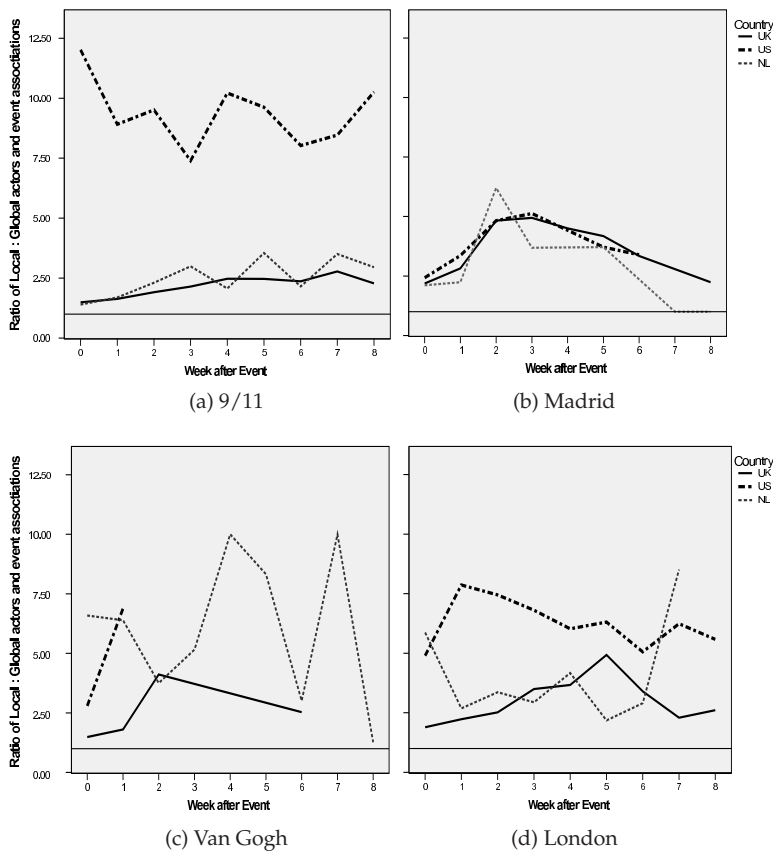


Figure 5.2: Globalization and Localization of the Four Events
(a. N = 6,022, U.S. = 4,494, U.K. = 1,140, Netherlands = 338;
b. N = 580, U.S. = 172, U.K. = 223, Netherlands = 185;
c. N = 656, U.S. = 17, U.K. = 17, Netherlands = 622;
d. N = 1,289, U.S. = 322, U.K. = 759, Netherlands = 208)

5.2 shows the results of both hypotheses during the four events. The lines represent the ratio between the localization and the globalization of the respective event in the different countries' media.

Looking at 9/11, we see that the US press localized the event to a great extent, with an average of 10, meaning that the local actors were associated with 9/11 ten times more often than global actors. In the other newspapers we see that this global event was domesticated, especially after the first week, although international actors stay important as well, receiving about a third of the attention in the first weeks, which drops to around a quarter in the later weeks. Apparently, after the first week in which the global framework was determined in the news, with a focus on the protagonists' reactions, the media turned to the local actors in giving the events a domestic meaning. Chi-square tests show that local actors were used in the local newspapers significantly more often than in the other news for all countries and weeks ($p < 0.01$).

The attacks in Madrid show a similar picture. The newspapers in all three countries follow the same pattern: after an initial focus on Spanish actors, the news media turned especially to local sources for their news, strongly domesticating the event in the following weeks. After a few weeks, the attention paid to Madrid decreased rapidly, especially in the US papers. Significance tests show that the localization is significant overall and for the first three weeks; after that period the frequencies are too low for the use of local actors to be significantly different from the global use of these actors

Figure 5.2c (the Van Gogh murder) is less informative because hardly any attention was paid to the event by the international newspapers. The event was globalized by the Dutch press slightly more than 9/11 was in the American press, but still only one non-local event or actor was mentioned for every six local mentions. The overall localization by the British press was significant at the $p < 0.05$ level, but due to the limited number of articles, none of the weekly subsamples showed significant localization and the US coverage did not show any significant localization at all.

The London attacks, presented in graph 3d, show an interesting result: the event was more localized in the US than in the UK itself. For the US media, the attacks fitted within the earlier formed framework of 9/11 and focused more on local actors than the actors directly affected by the attack. In the British press, the globalization was almost 25%, with one international actor mentioned for three national actors or reference events. It should be noted that, due to our operationalization, this strong 'Americanization' of the London bombings is dependent on the qualification of 9/11 as an American event.

This tendency is also seen when we look in detail at the associations of the events with previous events (not presented here as a graph). When

the bombs exploded in Madrid, both the British and the Dutch newspapers refer mostly to the ETA, while the US press mentions 9/11 most often in their coverage. When we look at the London bombings, we see that both British and US press refer mainly to 9/11 followed by references to Madrid. In the Dutch press we see an inverse picture, with newspapers comparing the London bombings more often with Madrid than with 9/11. It seems that all media place these events in the global 'war on terrorism' frame, but with each retaining their own local focus through the events they use as a frame of reference.

We can conclude that the localization of all foreign events show the same pattern, with an increase of the localization after the first week. 9/11 was a local American event in US newspapers and was least domesticated by the international press. In contrast, the London bombings were strongly localized by the American newspapers while it was the event that was most globalized by the local British press. Van Gogh was an almost exclusively Dutch event, with low globalization by the Dutch press and almost no coverage by the other newspapers.

Rally around the Flag

'Rallying around the flag' is measured as associating the event with words indicating patriotism. Since this hypothesis predicts that this will happen in the local media more often than in other media, we can compute the ratio between the associations in these newspapers as a descriptive. Analogous to Hypothesis 1, we can test significance by testing the independence of an article about the event mentioning a patriotic word from the locality of the medium.

Figure 5.3 shows that for 9/11 the highest local patriotism is found in the US newspapers. Especially in the first week after the event, we see an increase of patriotism in US papers compared to the other countries' press. Chi-squared tests confirm that the overall US local patriotism is significantly local for 9/11 ($p < 0.01$), while weekly tests show significance only for the first two weeks and week 4 ($p < 0.05$).

In this respect, the figure concerning patriotism around the attacks in London is interesting. Here we see that the ratio in British newspapers is below 1 in the first seven weeks after the event. The US newspapers associate the London attacks with American patriotism more often than the British newspapers do. Apparently, for the US press the London attack fitted well in the framework of the war on terror. These findings suggest that patriotism is more related to the country of the newspaper than the happening of a local event, contrary to our third hypothesis.

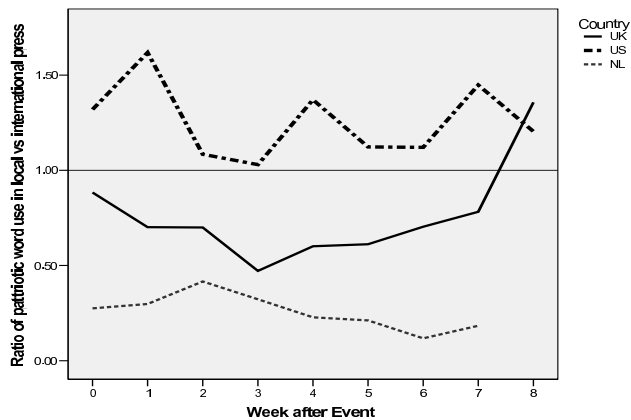


Figure 5.3: Relative Association with Patriotic Terms of the Local Events (N = 7,967; U.S. = 6,022; U.K. = 1,289; Netherlands = 656)

Muslims in the News Coverage

In the theoretical discussion, we saw contradictory findings with respect to whether Muslims and Islam were covered differently before and after 9/11 and after the events that were local as far as the British and Dutch newspapers were concerned. In this section we look at associations with Muslims and Islam in the news.

Figure 5.4 shows the associaton patterns of Islam and Terror with the government and with Positive, Negative, and Patriotic values. The left and middle columns show all associations greater than 20% in the year before and after 9/11, respectively. The right-hand column shows the difference between the association before and after 9/11 in percentage points.

In all cases, the most striking change is a strong increase in the association with terrorism. In the US, the reverse association of terrorism with Islam decreased. Interestingly, the association with the positive and negative values also decreases, and the other associations are stable, so it seems that the Terrorism discourse is expanded into other directions than those measured in this chapter. In the UK and the Netherlands, the association of Terror with Islam and with Government increases slightly (+0.1). In the UK, associations with positive and negative values remains constant, while in the Netherlands Islam is associated more with negativity, while Terror is less positive. In both the US and UK, the association between Islam and patriotism increases, and it would be very interesting to know what this increased relation is.

To test our fourth hypothesis, we look more closely at two aspects of

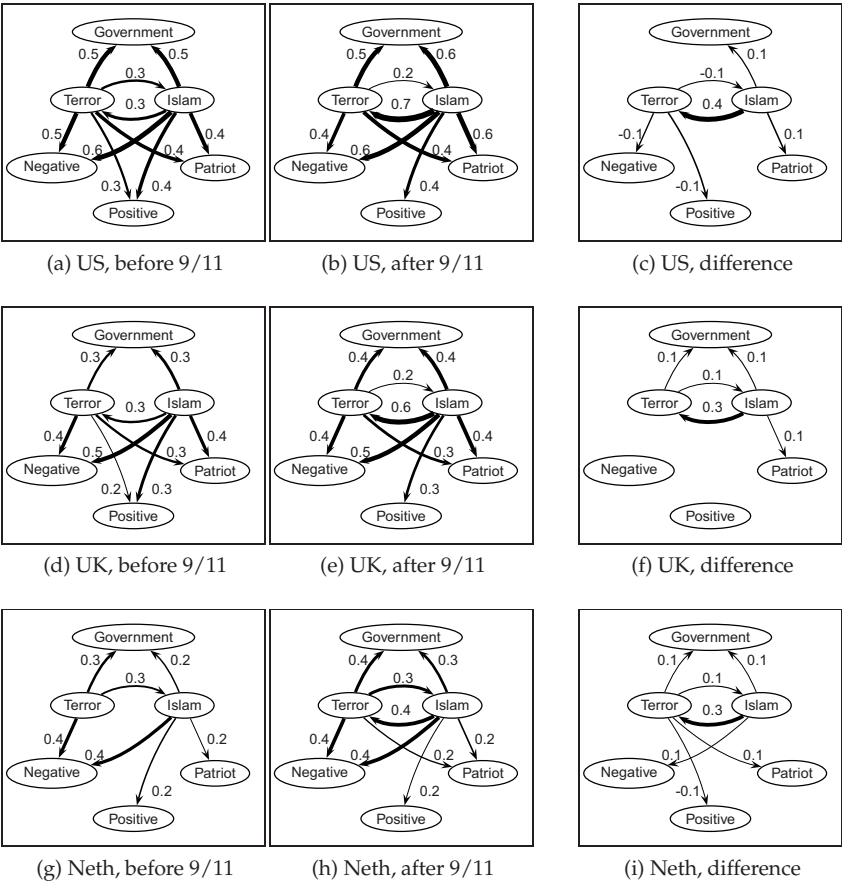


Figure 5.4: Associations of Terror and Islam

Reading example: In the US, before 9/11 the association of Islam with government was .5: half of all documents containing Islam also contained one of the governmental actors. This association increased to .6 after 9/11, so there was an increment of 0.1 as shown in subfigure (c). In the UK, before 9/11 the association of terror with Islam was below the threshold (0.2). After 9/11 this association was 0.2, an increase of 0.1 according to subfigure (f).

these networks: evaluations (associations with positive/negative) and associations of Muslim with terrorism. The evaluative portrayal of Muslims is measured as the ratio of associations with negative terms compared to positive terms. The connection with terrorism is measured as their association with terrorist terms. We can again use a chi-squared test to test whether using a negative word is independent of the article being before or after the event for all articles mentioning Muslims. Interestingly, although Muslims were portrayed much more negatively immediately after both 9/11 and the local key events, after the first months the negativity dropped to a level slightly, but not significantly, higher than it was before the event. Only the British press is overall significantly more negative after the event, but this might simply be because ‘their’ event is still relatively recent.

The association of Muslims with terrorism is more interesting, as shown in Figure 5.5. In all countries, Muslims were strongly associated with terrorism immediately after 9/11, and although this level decreased over time it stayed significantly higher than before 9/11. In the UK, the association with terror also increased significantly after the London bombings, although the long-term effects are not clear since the bombings are fairly recent. In the Netherlands, the association with terror actually decreased significantly after the murder of Van Gogh, even

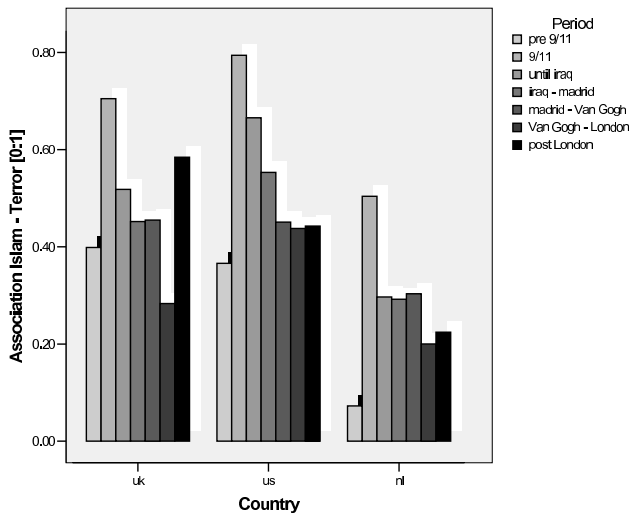


Figure 5.5: Associations of Muslims with Terrorism per Country per Period

(N = 43,868; U.S. = 16,929; U.K. = 12,258; Netherlands = 14,681)

though the negativity increased slightly (not shown here). Apparently, even though the murder was often called a terrorist attack by the press, this did not cause a framing of Muslims as terrorists afterwards (all significances $p < 0.01$).

It seems that the associative frame between Muslims and terrorism was created not by local events, but rather by 9/11 as a global event. This association was reinforced in the UK after the London bombings but decreased after the murder of Van Gogh in the Dutch press.

5.5 Conclusion

This chapter presented Associative Frames, a method for automatically extracting Semantic Networks based on the co-occurrence of concepts as measured using keyword lists. Agenda Setting, second-level Agenda Setting, and Emphasis or Issue Framing can all be seen as theories about the transfer of association patterns or networks from the media to the receiver. By measuring the individual concepts rather than the whole frame, we are able to use computer techniques to automate this measurement. Moreover, the data obtained using this method is less dependent on the specific definition of the measured frames, increasing the reusability of the data and making it easier to test different theories on the same data. In this way, we have given a clear interpretation of linguistic co-occurrence in terms of current communication theory, which helps to bridge the gap between theoretical sophistication and measurement techniques.

We also presented a methodology for calculating association scores as a conditional reading probability. This measure is asymmetric, conforming to substantive intuitions about the nature of association. Moreover, probability calculus is a well-understood field of mathematics, making the model easy to extend and compare with other methods. Finally, we gave an example content analysis of the associations between Islam and Terror in the Dutch, British, and American press between 2000 and 2006, showcasing the power of this method to analyse large amounts of text.

This method has some limitations, however. First, we accept that not all frames can be expressed as simple association patterns. For some frames, it will be necessary to extend the current model, for example by distinguishing between types of relations or by measuring whether the relation is positive or negative. Although these measurements are difficult to carry out linguistically, a lot of work is being done on such problems in the Computational Linguistics community, and the model can easily be extended as soon as acceptable accuracy is reached on the linguistic extraction. Also, certain framing theories, such as Equivalency

Framing, are even more difficult to fit into this model, since they are not directly based on association networks.

Another limitation is the difficulty of interpreting co-occurrence measures on keywords. An overly broad or narrow set of keywords for a concept can easily skew the results or measure something completely different from the intended concept. For this reason, it is very important to go 'back to the text' and qualitatively assess whether the found contexts are actually expressing the relation that one is interested in. Moreover, the validity of the extracted occurrence and co-occurrence of concepts should be tested by comparing extracted concepts and networks with manual annotations on a sample of documents, analogous to the reliability measurements common in manual content analysis. The creation and evaluation of keyword lists can be done systematically using Keyword-in-context programs or manual coding of a subset of documents, but it is ultimately the interaction between the quantitative measurement and the qualitative control that ensures a correct interpretation of the results. This makes automatic content analysis more labour-intensive than an 'automatic approach' might suggest, but once the lists are in place and they have been well tested, there are little extra costs in analysing more documents, making this approach particularly well suited for research topics that are ongoing or that cover a large number of texts.

These limitations notwithstanding, this chapter provides a clear communication theoretical interpretation and probabilistic operationalisation of co-occurrence. This yields a flexible method for the automatic analysis of text, either for testing hypotheses or for exploratory research, which is a contribution to the measurement techniques currently available to the communication scientist.

CHAPTER 6

Using Syntax to find Semantic Source, Subject, and Object

'Lubbers: ignore Wilders and Verdonk'

(Lubbers: negeer Wilders en Verdonk; *Het Parool*, October 24, 2007)

'Nieuwspoor reluctantly gives platform to 'free speech'

(Nieuwspoor geeft met tegenzin podium aan 'het vrije woord'; *Trouw*, March 7, 2008)

Populist politicians such as Geert Wilders and Rita Verdonk are highly visible in the news media, but are they given a platform to state their opinions and explain their positions or are they merely talked about as an issue? Do journalists pick sources who agree with their position? Are incumbent politicians presented more authoritatively in the media? This chapter presents methods to extract quotes sources and differentiate between semantic subject and object based on syntactic information, allowing the above questions to be answered automatically.

This chapter is an edited version of:

Wouter van Atteveldt, Jan Kleinnijenhuis, and Nel Ruigrok (in press), *Parsing, Semantic Networks, and Political Authority: Using syntactic analysis to extract semantic relations from Dutch newspaper articles*, accepted for publication in Philip Schrodtt and Burt Munroe (eds.), Special Issue on Automated Natural Language Analysis of Political Analysis.

6.1 Introduction

The previous chapter described how the simple fact that two concepts co-occur in a text can be used to answer a number of interesting research questions. However, a network of undirected and unsigned links is insufficient for answering many research questions. For some questions, we also need to know the direction (who is acting and who is acted upon) and sign (or valence, polarity: is the relation positive or negative?) or relations. Moreover, texts often contain a number of (quoted or paraphrased) sources, and it is often necessary to analyse the networks belonging to these sources separate from the main network. This chapter describes a method for using the syntactic structure of text to enrich the Semantic Network in two respects: it differentiates between the semantic subject (agent) and object (patient), that is, it creates directed links between the concepts; and it recognises quoted text and identifies the source and the quote, making it possible to create a separate network for the quoted text that is attributed to the quoted source.

The general approach taken in this chapter is rule-based: All sentences are syntactically parsed using the Dutch Alpino parser (Van Noord, 2006), and a number of patterns are used to extract source–quote and subject–object relations. Subsequently, the political actors are identified using label matching and simple anaphora resolution. This yields a directed graph with actors as nodes and separate contained subgraphs attributed to quoted sources.

To assess the quality of the automation by this method, the manual semantic network content analysis of the 2006 Dutch election campaign will be used as a Gold Standard (Kleinnijenhuis et al., 2007a). The availability of human codings allows us to assess the validity and the automatic method. In order to test this validity, we present a use case in which the authority of political actors in the media is operationalised based on their occurrence as source, subject, and object. Construct validity is then defined as the extent to which the method measures the theoretical constructs correctly, and is defined as the agreement with human coding on the level of analysis, in this case authority scores per week of news on an issue. Finally, we test whether the predictive validity of automatic coding is satisfactory: it should make no difference whether substantive hypotheses about the effects of political communication are tested with data derived from an automatic content analysis or a content analysis by human coders. This is determined by testing a number of hypotheses that explain which politicians should be authoritative based on the newspaper and news topic.

The remainder of this chapter is structured as follows. The next section will describe the method used for extracting the source–quote and

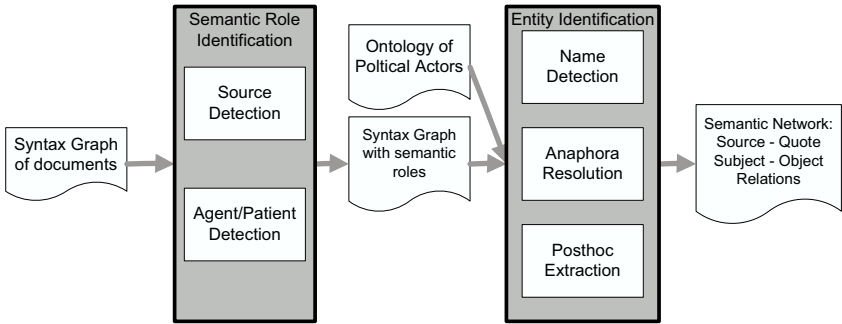


Figure 6.1: Overview of the method for recognising source, agent, patient

subject–object relations. The third section describes the methodology for testing the validity of the method, including the case study on measuring media authority. The fourth section presents the results of these tests, and the fifth section will investigate which parts of the method perform well and what the causes of failure are.

6.2 Determining Semantic Roles using Syntax Patterns

Figure 6.1 gives an overview of the method presented in this chapter. The method assumes that the documents to be analysed have been fully parsed in the pre-processing¹. The input of the method consists of the syntax (dependency) graphs of these documents. The semantic roles are identified by pattern matching on these syntax graphs by the Source Detection and Subject–Object detection modules. This yields a syntax graph enriched with the identified semantic roles. This graph is combined with an ontology of political actors² and fed into the Entity Identification module. This module detects the occurrence of the entities in the ontology in three steps: direct name matching, anaphora resolution, and post-hoc extraction of remaining names that were not identified as playing a semantic role. This results in a Semantic Network of source–quote and subject–object relations between objects from the ontology. The following sections will describe the modules comprising the method.

¹See section 3.2.4 on page 46

²See chapter 4 and section 8.4 on page 156

6.2.1 Source–Quote recognition

Journalists use a limited number of relatively fixed patterns when including quotes and paraphrases in newspaper articles, and a limited number of verbs are used as indicators of these patterns. As a consequence, it is feasible to recognise these patterns from the text and dependency graph using a limited number of patterns. In principle, we are only interested in extracting quotes as assertive speech acts (Searle, 1969), not as directives (orders) or commissives (promises).

Figure 6.2 shows the dependency trees of two fictive newspaper sentences, corresponding to the main patterns used in this module. Figure 6.2a shows the parse of the sentence ‘Bush denkt dat Blair een leugenaar is’ (*Bush thinks that Blair is a liar*). This exemplifies the first pattern: [source– v_{says} –quote]. The key to this pattern is the occurrence of one of 27 ‘says’ verbs, such as *zeggen* (*to say*) or *denken* (*to think*). The subject of this verb is the source of the quote, and the word *dat* (*that*) or (if *dat* is omitted) the verbal complement of the main verb is the quote. In the figure, the identified key word is indicated with a double line, the source has a thin black line, and the quote has a thick grey line.

Figure 6.2b is an example of the second main pattern: [$p_{volgens}$ source, quote]. In this sentence, ‘Volgens Bush is Blair een leugenaar’ (*According to Bush, Blair is a liar*), the key word is the preposition *volgens* (*according to*). The direct object of this preposition is the source of the quote, while the parent of the preposition and its children, except for the preposition itself, forms the quote. This pattern is often used when citing reports

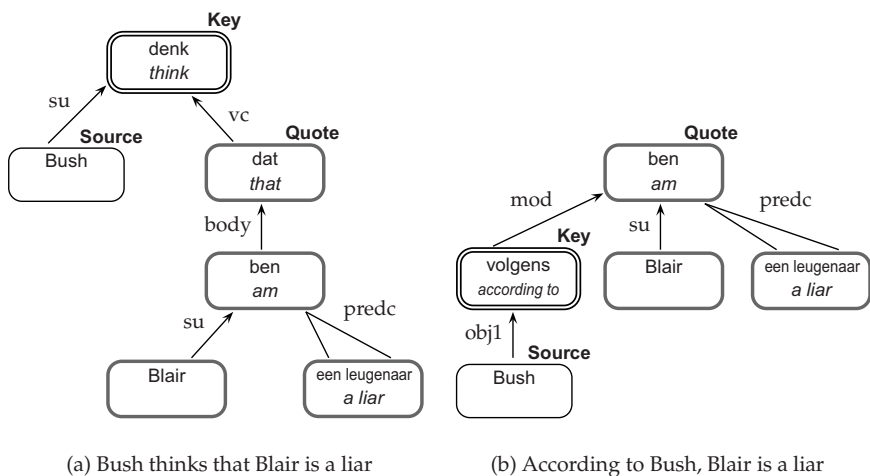


Figure 6.2: Example syntax patterns for source recognition

(e.g. *According to a report from ...*); other key words are ‘aldus’ (*as stated by*) and ‘blijkens’ (*as appears from*).

Apart from these main patterns, we created a set of minor patterns for special constructions such as *Bush laat weten dat ...* (*Bush makes it known that ...*) and *Dat blijkt uit ...* (*This is evident from ...*). Moreover, surface (string) matching was used to extract explicit quotes such as *Bush: “Blair is een leugenaar”* (*Bush: “Blair is a liar”*).

6.2.2 Subject–Object recognition

This module aims to overcome the discrepancy between the syntactic subject-object and the semantic subject-object (or agent–patient). For our purposes, the semantic subject (or agent) should be the actor or object primarily doing or causing something, while the semantic object (or patient) is the actor or entity that it is done to (Dixon, 1991). Below are some simple example sentences, with the syntactic subject and object indicated above the text and the semantic subject and object below it:

1.

| | | |
|----------------|---------------|----------------|
| <u>Subject</u> | | <u>Object</u> |
| Bush | vertrouwt | Blair |
| <i>Bush</i> | <i>trusts</i> | <i>Blair</i> |
| <u>Agent</u> | | <u>Patient</u> |
2.

| | | | |
|----------------|----------------------|--|--------------------|
| <u>Subject</u> | | | <u>Prep.Object</u> |
| Blair | wordt vertrouwd door | | Bush |
| <i>Blair</i> | <i>is liked by</i> | | <i>Bush</i> |
| <u>Patient</u> | | | <u>Agent</u> |
3.

| | | |
|----------------|---------------------|--------------------|
| <u>Subject</u> | | <u>Prep.Object</u> |
| Bush | heeft vertrouwen in | Blair |
| <i>Bush</i> | <i>has trust in</i> | <i>Blair</i> |
| <u>Agent</u> | | <u>Patient</u> |
4.

| | | | | |
|----------------|-------------|------------------|-----------|--------------------|
| <u>Subject</u> | | <u>Object</u> | | <u>Prep.Object</u> |
| Bush | stelt | zijn vertrouwen | in | Blair |
| <i>Bush</i> | <i>puts</i> | <i>his trust</i> | <i>in</i> | <i>Blair</i> |
| <u>Agent</u> | | | | <u>Patient</u> |

As can be seen from these examples, sometimes the semantic object is the direct object, sometimes the indirect object, and sometimes the object of a prepositional phrase. Moreover, in passive sentences the subject is the semantic object, and the (prepositional) object the semantic subject. Extracting such relations from syntactic structures has been done by Katz and Lin (2003), who extracted specific relations from a dependency parse of English texts. Bouma et al. (2003) describe a system for extracting relations from Dutch texts using the Alpino parser in the context of

a Question Answering system. Similar to these systems, our module works as outlined below:

1. A predicate is formed from the ‘verb chain’ in the tree. The verb chain consists of the finite verbs and all verbal complements. Also, infinitive *te*-constructions (*to*-constructions, e.g. *agreeing to disagree*) are considered part of the verb chain.
2. All predicative complements of the verbs in the chain are included in the predicate. An example of a predicative complement is the preposition attached to a verb (e.g. *search for information*).
3. If any of the nodes in the predicate has an indirect object, all direct objects are included in the predicate.
4. The predicate is inspected to determine whether the sentence is active or passive. If the sentence is active, the subject node of the top verb is identified as the semantic subject, and all (indirect) objects of the nodes in the predicate are identified as the semantic object. For passive sentences this is reversed. Passive verbs (e.g. Dixon, 1991) were not included in the version of the module tested here.

An alternative approach is using machine learning techniques to extract relations, such as done by Zhang et al. (2008) and Jijkoun (2007). This is a promising technique for improving this module given that we have a training corpus of human codings, but is beyond the scope of this chapter.

As an example of the source and subject–object detection, consider the annotated syntax tree in figure 6.3 of the Dutch sentence reading ‘Kerry zegt dat Bush hem een draaikont heeft genoemd’ (*Kerry says that Bush called him a flip-flop*). Figure 6.3a is the raw dependency graph.³ The top relation is Kerry as the subject of *says*, which has a verbal complement *that*. This complement contains the clause with Bush as the subject of *has called*, which has an indirect object (*obj 2*) *him*, and a direct object (*obj 1*) *flip-flop*. Finally, *flip-flop* has a determiner *a*. Figure 6.3b shows the dependency graph enriched with semantic roles. The circles indicate a source construction: *says* is the key to the construction, indicated by a dashed line, while the subject *Kerry* is indicated by a solid black circle and the quote *that* (and all underlying nodes) by a thick grey circle. The rectangles indicate a subject–predicate–object construction. The predicate is *has called flip-flop*, which is displayed using a double rectangle. The subject, *Bush*, uses a solid line while the line for the object, *him*, is thick grey. Note that *flip-flop*, although it is the grammatical object, is correctly classified as part of the predicate.

³See section 3.2.4 on page 46

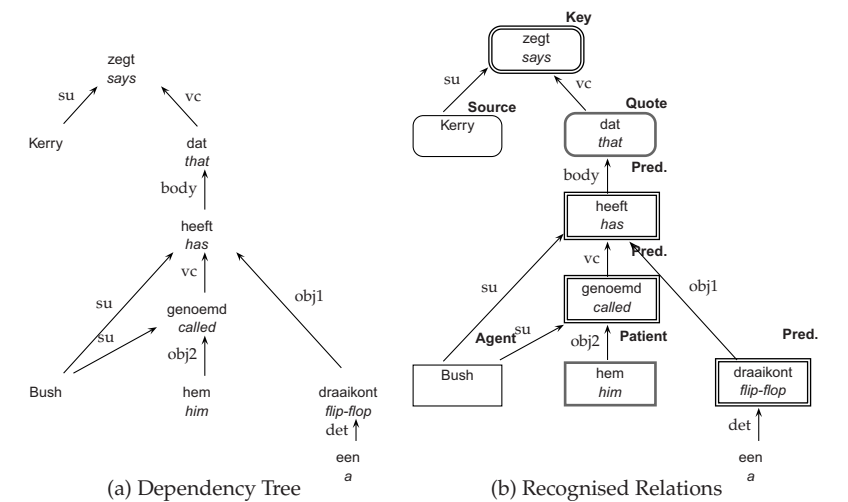


Figure 6.3: Example of a parsed sentence with semantic relations
Kerry zegt dat Bush hem een draaikont heeft genoemd
Kerry says that Bush called him a flip-flop

6.2.3 Entity Identification

The task of the next module of the method is recognising which politicians or political parties are referred to in the text. Since we know the names of the relevant political actors in the campaign beforehand, and politicians are generally referred to by their name, this is a fairly easy task. The recogniser identifies a politician if either his last name is mentioned, or his first name is mentioned and his last name was mentioned earlier in the same text. Political parties are identified if the party name appears but is not part of a noun phrase also containing the name of a party member (since the party name is often mentioned between parentheses when a politician is first mentioned). Finally, a number of custom rules are used to recognise references to politicians by references to the office held by them. Examples are ‘the minister of health’, the ‘leader of the Labour Party’, and ‘the Prime Minister’.

Anaphora Resolution

The main difficulty in recognising entities is that, after the first mention of an politician, he or she is usually referred to using a pronoun (‘he’, ‘she’) or a definite noun phrase (‘the representative’, ‘the minister’). Such references are called anaphoric references. Anaphoric reference resolution is an active field of (computational) linguistics. Although

completely solving this problem is very difficult, a number of systems attain good performance using fairly simple heuristics, as surveyed by Mitkov (2002). The general approach in such systems is to identify all candidates, filter out impossible candidates, and rank the remaining candidates using a heuristic function.

From the constraints and preferences surveyed by Mitkov (2002) and Lappin and Leass (1994), four seem immediately useful for our application: The first constraint is gender and number agreement: in English (and in Dutch), pronominal references have to match the antecedent in number and gender. This means that ‘he’ has to refer back to a singular male antecedent, while ‘they’ refers to a plural antecedent. To this we can add ‘party agreement’, in the case of a nominal reference to a member of a certain party such as ‘the socialist.’ A second group of constraints are the syntactic constraints. In simplified terms, these state that a normal pronoun cannot reference a noun in the same part of a sentence, while a reflexive pronoun (*zich*, *himself*) has to refer to such a noun: In ‘Bush likes him’, him cannot refer to Bush, while in ‘Bush likes himself’ the pronoun has to refer back to Bush. The third factor is syntactic parallelism: a noun in the same syntactic position as the anaphora is preferred over other nouns. The fourth factor is a salience heuristic described by Lappin and Leass (1994), that essentially prefers subjects to objects and recent sentences to earlier sentences.

From this survey, we designed the following fairly simple implementation for anaphoric references:

1. **Identification.** Any ‘animate’ pronoun and definite noun phrase that describes a political function is identified as an anaphora
2. **Candidate selection.** A list of possible noun phrases is made from the text preceding the anaphora, using the following ranking:
 - (a) Antecedents are ranked from the current sentence to the first sentence in the paragraph
 - (b) Within a sentence, antecedents are ranked by their grammatical function: [function-of-pronoun; Subject; Direct Object; Indirect Object]
3. **Candidate Filtering** based on gender and number and (within the same sentence) syntactic constraints. Since we are only interested in the best match, the candidates are generated according to their rank, and the first matching candidate is considered to be the antecedent of the pronoun. For example, in the sentence in Figure 2 above, for *him* there are two candidates in the same sentence: *Bush* and *Kerry*. However, *Bush* is a sibling of him and is excluded, making the module conclude that *Kerry* is probably the antecedent.

As another example of how these rules work, consider the following two sentences:

1. Hirsi Ali verliet de politiek omdat Minister Verdonk haar het Nederlander-schap leek te ontnemen
Hirsi Ali left politics because Minister Verdonk withdrew her citizenship
2. De minister heeft dit later heroverwogen
Later, the minister reconsidered this

In the first sentence, *her* refers to Hirsi Ali because Verdonk is rejected as antecedent as a sibling of the anaphora. If the pronoun had been ‘his’ or ‘their’, the module would have also rejected Hirsi Ali since it knows that Hirsi Ali is singular and female. In the second sentence, the minister is also considered anaphoric since the specific department is not mentioned (‘the minister of education’ uniquely identifies a person and is not considered anaphoric). Since no suitable referents exist in that sentence, the module looks back to the first sentence and identifies Hirsi Ali and Verdonk as candidate antecedents. Since it knows that the antecedent of the minister has to be a minister, it can rule out Hirsi Ali and selects Verdonk, of whom the module knows she was minister of immigration at that time. This shows how background knowledge of the domain is required for filtering out possible antecedents (chapter 8 gives a more formal description of this background knowledge). Note that we do not resolve the pronoun ‘this’, since we are only interested in actors in this paper and ‘this’ can only refer to an inanimate antecedent.

Post-hoc Extraction

Finally, it was observed during development that many sentences do not contain a full subject–predicate–object construct, but still contain relevant political actors. To identify these actors, all nodes that were not considered part of any semantic role are inspected. If any such node equals the last name of a politician or the name of a party, a rudimentary triple containing this actor is created, placing the triple at the object place if it is a modifier, object, or complement of another node, and — if none of these — subject.

6.3 Determining Validity

The purpose of this study is to investigate the quality of a method for extracting the semantic relations from text using grammatical analysis.

To evaluate this quality, we ask three specific questions, roughly matching the concepts of concurrent validity, construct validity, and predictive validity.⁴

Concurrent validity How well does it extract the relation from each unit of measurement?

Construct validity How well does it calculate variables at each unit of analysis?

Predictive validity How well can the results be used to model extra-media data?

For each of the three questions, we perform a comparison between the output of the method and manual coding of the same material. For this we use a corpus consisting of a selection of sentences from a manual analysis of the 2006 election campaign using the NET method (Kleinnijenhuis et al., 2007a). Sections 2.4 and 2.5 give more detail on the NET method and the 2006 election study and resulting corpus, respectively. From the 2006 election corpus, we used all relations containing a political analysis, including 15,914 out of 27,137 relations.

6.3.1 *Concurrent Validity*

The most straightforward method for establishing the performance of the method is by comparing the sources, subjects, and objects found by the method with those found by the human coders. Since we consider the codes found by human coders to be valid, if the computer codes match those codes they are concurrently valid. Since it is sometimes unclear to which sentence a statement belongs, especially in the context of anaphora and sentence-spanning quotes, this analysis is performed at the article level. For each article, we compute the multiset of items that the method and the human coder found, where an item is a combination of a place (source, subject, or object) and a political actor. From these multisets we compute recall, precision, and F1 Score as defined in section 3.4 on page 49.

6.3.2 *Face Validity*

The reliability measure presented above depends on the human coding being the only correct answer. In the context of semantic networks, some

⁴The concepts used here are a slight deviation from their actual meaning, since we use the human codings as a benchmark to judge the quality of the new automatic codings for each of the three types mentioned above, whereas in the methodological literature the comparison with existing measures is usually exclusively associated with ‘concurrent’ validity. Nevertheless, these terms are a close match and serve as convenient descriptions.

sentences are semantically ambiguous. Depending on the interpretation of the coder, one of the possible codings is chosen. This means that part of the disagreement between the computer and the human coding can be due to semantic ambiguity of the texts rather than errors made by the method. To investigate this, we manually inspect a set of sentences, and classify the output of the method as justifiable, false positive, or false negative. From this, we can also compute precision and recall scores as above. This checks whether the relations identified by the method are acceptable to experts, rather than identical to the relations they identified. Consequently, the face validity can be seen as an optimistic indicator of reliability. As this is not a rigid, formal evaluation, the results should be interpreted as an indication of the face validity of the output of the method rather than a definitive measurement of its reliability.

6.3.3 *Construct Validity*

Since we are interested in using the method to answer questions that are relevant to political analysis, it is important to determine the construct validity of the method: can the method measure the theoretical constructs needed for political analysis? This validity can be operationalised as the ability of the method to measure variables used for modelling political interaction at the unit of analysis. These model variables depend on the substantive research question, so we need to investigate a specific use case to determine the construct validity.

The case study that we present here analyses what Herbst (2003, p.489) calls media-derived authority: a 'legitimation one receives through mediated channels'. Media authority is an important aspect of media performance, and is a direct result of journalistic choices on how to portray a political actor. As such, it is an interesting variable both for studying media effects and the internal media logic. Media authority does not equate to visibility: if an actor is portrayed as untrustworthy or incompetent, he or she is visible but not very authoritative. A first definition of media authority is that the more an actor is given the possibility to talk directly to the voters, the more authoritative he or she comes across with the public (cf. Patterson, 1993; Donsbach et al., 1993; Donsbach and Jandura, 2003). This can be operationalised as the frequency with which an actor is used as the source of statements.

Another definition of media authority is being portrayed as acting, rather than being acted upon. This portrays a politician as being capable of and active in the solving of issues, rather than being the object of external circumstances and the topic of discussion and scandal (Klein-nijenhuis et al., 1998). This definition can be operationalised as the relative frequency with which an actors appears as the semantic agent rather

than as patient.

Given these definitions, we can compute the construct validity of the method by calculating the variable based on the manual coding and on the computer output, and determining the correlation between these calculations.

6.3.4 *Predictive Validity*

The predictive validity of this method is its ability to make correct inferences about real-world phenomena. To investigate this, we use the method to test three hypothesis pairs about media authority. For each hypothesis, we present the results based on the output of the method as well as those based on the manual coding, allowing a qualitative estimation of the external validity of the method. These hypothesis pairs are as follows:

H1: Indexing theory

Bennett (1990, p.106) formulated the theory of indexing: 'Mass media news professionals, from the boardroom to the beat, tend to 'index' the range of voices and viewpoints in both news and editorials according to the range of views expressed in mainstream government debate about a given topic'. This is in line with what Galtung and Ruge (1965), in their classic essay on news values, define as the 'relevance' of an individual as a product of his or her power (see also Weaver and Wilhoit, 1980; Olien et al., 1983; Stempel and Culbertson, 1984). Given the two definitions of media authority, this leads to two hypotheses:

H1a Powerful politicians are quoted relatively often

H1b Powerful politicians are portrayed as acting relatively often

The power of a politician is operationalised as the number of seats in parliament and as a dummy variable indicating incumbency. These two variables are put in a regression model with either of two definitions of authority as operationalised above. Table 6.1 lists the parties with seats in the Dutch parliament before the 2006 elections.

H2: Opportune Witnesses

The specific selection of sources is a central part of the theory of 'opportunistic witnesses' (Hagen, 1993). According to this theory, journalists prefer to quote sources that support the journalist's own political stance or the editorial line. By selecting sources that put forward a certain opinion, journalists can convey the impression that experts share their personal

Table 6.1: Parties in the Dutch Parliament (2003-2006)

| Party | | Incumbent | Seats |
|--------------|---------------------------------|-----------|-------|
| CDA | Christian Democrats | 1 | 44 |
| PvdA | Labour, Social Democrats | 0 | 42 |
| VVD | Conservative Liberals | 1 | 28 |
| SP | Socialist Party | 0 | 9 |
| LPF | Fortuynists (extreme-right) | 0 | 8 |
| GroenLinks | Green / Left | 0 | 8 |
| D66 | Progressive Liberals | 0 | 8 |
| ChristenUnie | Orthodox Reformed (progressive) | 0 | 3 |
| SGP | Orthodox Reformed | 0 | 2 |

views. Ruigrok (2005) showed that during the Bosnian war, Dutch journalists made extensive use of opportune witnesses in order to strive for a military intervention. Traditionally, the Dutch media were segmented or pillarised in different ideological camps (Lijphart, 1975). Although this is mainly a feature of history, newspapers still differ in the ideology of their readership and there are observable differences in the way they treat the different parties (Kleinnijenhuis et al., 2007a). This leads us to the following hypotheses:

- H2a** Newspapers cite politicians belonging to their traditional 'pillar' relatively often
- H2b** Newspapers portray politicians belonging to their 'pillar' as acting relatively often

For this analysis, the two definitions of authority are correlated with a dichotomous variable indicating whether the party is considered aligned

Table 6.2: Dutch national newspapers and their traditionally aligned parties

| Newspaper | Aligned parties |
|------------------|---------------------------|
| De Volkskrant | PvdA, SP, D66, GroenLinks |
| NRC Handelsblad | CDA, VVD |
| Telegraaf | CDA, VVD |
| Trouw | CDA, CU, SGP |
| Algemeen Dagblad | CDA, VVD |

with the newspaper. Table 6.2 lists the five large national newspapers in the Netherlands, along with our operationalisation of newspaper alignment. It should be noted that this operationalisation is rather arbitrary. Newspapers are no longer rigidly embedded in an ideological 'pillar', and do not proclaim an open allegiance to a certain party. Moreover, the three relatively conservative newspapers are all considered to be aligned with the same parties, even though there are substantial differences between these newspapers. However, the purpose of this paper is not to give the definitive answer on this hypothesis, so even if this operationalisation is debatable, it is interesting to investigate whether the method presented in this chapter can be used to test it.

H3: Issue ownership and Economic Voting

A well-established theory is that of issue ownership: due to historic and current ideological cleavages, certain parties are seen as the 'owner' of certain issues, meaning that they are generally regarded as the party with the best reputation for solving the problems of those issues (Budge and Farlie, 1983; Petrocik, 1996; Petrocik et al., 2003; Benoit, 2007). For example, the welfare method is traditionally owned by left-wing parties, while fiscal and economic issues are owned by right-wing parties.

Additionally, it is often found that (news about) positive economic development is beneficial for the incumbent party or parties, while negative economic development is harmful (Lewis-Beck, 2006). This body of literature on 'economic voting' or 'retrospective voting' is easily subsumed in the issue ownership hypothesis by assuming that the incumbent party owns valence issues such as economic growth and employment. The issue ownership theory has two parts. It assumes that parties who succeed in getting owned issues in the media will win at elections, which is backed up by empirical research (Petrocik, 1996; Kleinnijenhuis et al., 2007b). Moreover, it assumes that parties prefer to address their owned issues in the media. From the latter part, we hypothesise a relation between issue ownership and media authority:

H3a In reporting on an issue, politicians from parties 'owning' that issue are cited relatively often

H3b In reporting on an issue, newspapers portray politicians from parties 'owning' that issue as active relatively often

Recent research from the US suggests, however, that a neck-and-neck race in election campaigns forces all parties to address the same set of key issues (Sigelman and Buell, 2004; Kaplan et al., 2006), which makes it worthwhile to test these hypotheses outside the US.

Table 6.3: Issues and Issue owners in the 2006 Dutch elections

| Issue | | Issue Owner |
|-------------|---|--------------------|
| Valence | Economic growth, employment | CDA, VVD (incumb.) |
| Leftist | Social welfare | PvdA, SP |
| Rightist | Fiscal policy | VVD |
| Immigration | (Illegal) immigration, integration | VVD, extreme-right |
| Values | Christian / Traditional values | CDA, CU |
| Environment | CO ₂ reduction, national parks | GroenLinks |
| Reforms | Democratic Reforms | D66 |

Similar to H2, these hypotheses are tested by regressing authority on issue ownership. As shown in table 6.3, issue ownership is operationalised as a dummy variable, based on a survey of Dutch citizens in September 2006 (Kleinnijenhuis et al., 2007a, p.113).

Combined model

All hypotheses can be measures using one data set if we take all statements about the same party in all articles from one newspaper on one topic as the unit of analysis. Moreover, all hypotheses have media authority as the dependent variable. This means we can construct a combined regression model with all independent variables mentioned above, which allows us to compare the relative explanatory power of the different theories on media authority.

6.4 Results

The main research question of this study is whether it is possible to devise a method that uses the grammatical analysis of text to draw conclusions that are relevant for political analysis. This section will answer that question by presenting the results of the validity tests, answering the question 'How good/useful is this method?'

6.4.1 Concurrent Validity

The performance of the method at the article level is given below in table 6.4. The columns list the precision, recall, and F1 Score of the method. The first three rows contain the performance on identifying sources, subjects, and objects. The fourth row contains the combined performance,

Table 6.4: Performance of the method at the article level

| Measure | Precision | Recall | F-score |
|--------------------|-----------|--------|---------|
| Source | 0.53 | 0.31 | 0.39 |
| Subject | 0.57 | 0.51 | 0.53 |
| Object | 0.39 | 0.36 | 0.37 |
| Combined | 0.51 | 0.43 | 0.47 |
| Ignoring frequency | 0.56 | 0.58 | 0.57 |
| Aggregate | 0.65 | 0.66 | 0.65 |

N=16,454 found items (1944 sources, 9614 subject, 4896)

and the fifth row contains the combined performance ignoring the frequency of identical items. The last row lists the performance on the level of parties rather than individual politicians, ignoring item frequency.

It is difficult to pass a qualitative judgement on these numbers. If we treat the aggregate F1 Score as an agreement measure, it would be rated ‘Acceptable’ in the classic article by Landis and Koch (1977). Additionally, as noted above, we cannot assume that the human coders always agree, hence some of the error is due to the method finding the objectively correct answer while the human found another answer. In the light of these considerations, we find this performance acceptable, but there is certainly room for improvement.

6.4.2 Face Validity

As noted before, human coders also have difficulty agreeing on the semantic relations in a sentence. In semantically ambiguous sentences, multiple interpretations are possible, and there is no single perfect answer. Hence, it is possible that the disagreement between human and computer analysis is partly due to different choices made by the human coder and the method that are both valid. For this reason, we also investigated the face validity or acceptability of the answers of the method by manually investigating 200 sentences and judging whether the answer is justifiable, or a false positive or negative.

The result of this outcome is encouraging: the average precision calculated using these judgements was 0.83, recall was 0.78, yielding an F1 Score of 0.80. Although this is not a formal evaluation, it can serve as an indication of (the upper bound of) the performance of the method at the sentence level. As such, it suggests that a fair part of the disagreement with human coders is because an automated analysis will often not reveal the judgements which are preferred by human coders, but still

judgements which are justifiable. It strengthens our conclusion that the performance of the method is acceptable.

6.4.3 Construct Validity

Where the previous measure was calculated at the sentence or article level, the construct validity of the method depends on measuring the theoretical constructs at the level of analysis. Table 6.5 shows the correlation coefficients of the two dependent variables between the human coding and the output of the method.

The correlations on the frequency of sources is very good (0.83), and the correlation of the proportional variable is acceptable (0.61). This performance is expected to increase with higher N due to the aggregation cancelling out noise, which is confirmed by correlation coefficients of .97 and .99 for the relative frequency when increasing the unit of analysis to all articles per topic or per newspaper respectively. Since the current unit of analysis contains on average 124 articles per newspaper/topic cell, this can be considered an indication of the number of articles required per cell for the method to make a good approximation.

Table 6.5: Correlation between method output and human codings at the level of analysis

| Hypotheses | Variable | Correlation | Signif. |
|---------------|---|-------------|------------|
| H1a, H2a, H3a | Relative frequency of use as source | .83 | $p < .001$ |
| H1b, H2b, H3b | Relative proportion of occurrence as source or subject rather than object | .61 | $p < .001$ |

N=450 combinations of issue x newspaper x party

6.4.4 Predictive Validity and Substantive Results

To give an indication of the external validity of the method, we shall test the hypotheses listed in the previous section based on the automatic coding and the human coding. If both methods reach the same conclusion, then this is a strong indicator of predictive validity.

H1: Indexing Theory

- H1a** Powerful politicians will be quoted relatively often
- H1b** Powerful politicians will be portrayed as acting relatively often

Table 6.6: Regression analysis of Indexing: political power and media authority Calculated using Automatic and Human coding

| Hypothesis | Variable | Automatic | Human |
|-----------------------|-----------------------------|-------------|-------------|
| H1a (party is source) | Number of seats (β) | 0.31*** | 0.29*** |
| | Incumbent (β) | 0.34*** | 0.35*** |
| | Adjusted R ² | 0.35 | 0.34 |
| H1b (party is acting) | Number of seats (β) | <i>n.s.</i> | <i>n.s.</i> |
| | Incumbent (β) | 0.14* | 0.15* |
| | Adjusted R ² | 0.02 | 0.02 |

N=450 combinations of issue x newspaper x party; * $p < 0.05$, *** $p < 0.001$

Table 6.6 summarises the regression analyses of the number of seats in parliament and incumbency⁵), calculated using the output of the method and the human coding. For H1a, the dependent variable is the frequency of each party as a source compared to the other parties. For H1b, it is the relative proportion of occurrences as subject or source rather than as object.

These findings support H1a: more powerful parties are quoted more often than less powerful parties, where seats in parliament and incumbency are both important. H1b is mostly rejected: although there is a significant effect of seats on authority, the overall explained variance is very low. Most importantly, the findings based on the output of the method and on the human codings are identical, lending credence to the use of the method for practical research.

H2: Opportune Witnesses

- H2a** Newspapers cite politicians belonging to their traditional ‘pillar’ relatively often
- H2b** Newspapers will portray politicians belonging to their ‘pillar’ as acting relatively often

These hypotheses were tested using correlation coefficients of the traditional alignment of a newspaper ⁶) with the dependent variable. Table 6.7 shows the results of this analysis, which displays the same pattern as the results for the first hypothesis: H2a is confirmed by the strong correlation of alignment with source use, while the evidence for H2b is much weaker, with a trend according to the output of the method and a weakly

⁵see table 6.1 on page 103
⁶see table 6.2 on page 103

Table 6.7: Correlation analysis of Opportune Witnesses: aligned parties and media authority Calculated using Automatic and Human coding

| Hypothesis (dependent variable) | Automatic | Human |
|---------------------------------|-------------------|---------|
| H2a (party is source) | 0.32*** | 0.34*** |
| H2b (party is acting) | 0.10 [†] | 0.13* |

N=379 combinations of issue x newspaper x party; [†]*p* < 0.1, **p* < 0.05, ****p* < 0.001. Note that the N is lower than 450 because one (new) newspaper does not have a traditionally aligned party.

significant correlation in the human coding. Although the results on H2b differ for the output of the method and the human coding, the direction is the same and this is not an immediate cause for concern.

H3: Issue Ownership and Economic Voting

- H3a** In reporting on an issue, politicians from parties ‘owning’ that issue are cited relatively often
- H3b** In reporting on an issue, newspapers portray politicians from parties ‘owning’ that issue as active relatively often

Similar to H2, table 6.8 lists the correlation coefficients between issue ownership⁷) and the dependent variable. From these results we can reject hypotheses H3a and H3b: owning an issue does not lead to being cited more often or to being portrayed as more active. This is in line with the findings that in campaign time, the competition between the actors is so fierce that no party will be left alone on any issue (Sigelman and Buell, 2004; Kaplan et al., 2006), especially not if that party is seen as having an advantage in talking about that issue. It would be interesting to see whether these hypotheses hold outside election campaigns.

⁷see table 6.3 on page 105

Table 6.8: Correlation analysis of Issue Ownership and media authority Calculated using Automatic and Human coding

| Hypothesis (dependent variable) | Automatic | Human |
|---------------------------------|-----------|-------|
| H3a (party is source) | n.s. | n.s. |
| H3b (party is acting) | n.s. | n.s. |

N=450 combinations of issue x newspaper x party

Combined model

Since the unit of analysis and dependent variables are the same for each group of hypotheses, we can create a combined regression model to determine the relative magnitude of each effect.

Table 6.9 lists the model parameters and explained variance for the models derived from the automatic and human coding. This model shows two interesting effects. First, the opportune witnesses effect disappears. This implies that the effect might have been spurious, since alignment is correlated with power since the three right-wing newspapers are aligned with the relatively large VVD and CDA. Second, issue ownership becomes a negative indicator of source use. This suggests an interaction effect with party power, as the positive effect of the latter causes issue ownership to have a significant negative effect, while the two are positively correlated (.17 and .25, respectively, for number of seats and incumbent; $p < .001$). This is confirmed by running a new model with an interaction effect between incumbency and owned issue. The interaction effect has a beta of $-.16$ ($p < 0.001$), and its inclusion reduces the owned issue to $\text{beta} = -0.1$ ($p < 0.05$). The explanation for this is that especially powerful parties are attacked as soon ‘their’ issues are placed on the agenda, while the less powerful parties are relatively free to talk about their issues. This is confirmed in the case of the Socialist Party and the extreme right-wing parties, which were given an uncontested floor to talk about social welfare and immigration, respectively. Their direct competitors seemed to think that keeping them out of the spotlight was

Table 6.9: A multivariate test of Indexing, Opportune Witnesses and Issue Ownership Calculated using Automatic and Human coding

| Dependent variable | Variable | Automatic | Human |
|--------------------|---------------------------|-------------|-------------|
| Party as source | No. of seats (β) | 0.31*** | 0.28*** |
| | Incumbent (β) | 0.35*** | 0.34*** |
| | Own Issue (β) | −0.15*** | −0.14*** |
| | Aligned Party (β) | <i>n.s.</i> | 0.11* |
| | Adjusted R^2 | 0.40*** | 0.36*** |
| Party is acting | No. of seats (β) | 0.13* | 0.14* |
| | Incumbent (β) | <i>n.s.</i> | <i>n.s.</i> |
| | Own Issue (β) | <i>n.s.</i> | <i>n.s.</i> |
| | Aligned Party (β) | <i>n.s.</i> | <i>n.s.</i> |
| | Adjusted R^2 | 0.02*** | 0.02*** |

N=450 combinations of issue x newspaper x party; * $p < 0.05$, *** $p < 0.001$

more important than attacking them on their points, and concentrated on the battle for the centre. This was punished by the voters, who massively left the large centrist parties except for the CDA for the smaller fringe parties.

6.5 Error Components Analysis

Even if the performance of the method is acceptable, there is certainly room for improvement. This makes it relevant to investigate the strengths and weaknesses of the method. This section will present a quantitative and qualitative analysis of the errors made by the method, answering the question: “Which aspects of the system can be improved?”

Since the method consists of a number of modules, the first question is: Which of the modules is causing the problems? This is a difficult question to answer exactly, as the true values for the intermediate results are unknown. For the sequential components in the ‘pipeline’, it is not possible to use the method with one of the components turned off as the later components depend on the output of the earlier components. However, for the three parallel modules that detect entities (the Entity Recogniser, the Anaphora resolution, and the Post-Hoc extraction), it is possible to run the method with or without one or more of those modules. This gives an indication of the performance of the individual modules as well as the contribution of this module to the total performance, by comparing the performance using all components with the performance of the method without that component. The results of this are listed in table 6.10. All scores are computed at the raw level; the performance of the method with all modules activated from table 6.4 is copied for reference.

Table 6.10: Performance and contribution of the different entity recognition modules

| Module | Items found | Performance using only this module | | | Performance degradation without module | | |
|---------------------|-------------|------------------------------------|------|------|--|------|------|
| | | Pr | Re | F | Pr | Re | F |
| All modules | 17573 | 0.51 | 0.43 | 0.47 | - | - | - |
| Entity Recogniser | 11412 | 0.58 | 0.31 | 0.40 | 0.03 | 0.28 | 0.24 |
| Pronominal Anaphora | 1469 | 0.55 | 0.02 | 0.04 | -0.01 | 0.03 | 0.02 |
| Nominal Anaphora | 715 | 0.65 | 0.05 | 0.09 | 0.00 | 0.01 | 0.01 |
| Posthoc Extraction | 3977 | 0.44 | 0.09 | 0.15 | -0.05 | 0.07 | 0.03 |

As can be seen from this table, the contribution of all modules is positive: the full configuration outperforms all other tested configurations. However, the contribution of the individual modules, except for the entity recogniser, is quite low. The total performance difference between only the entity recogniser and all components is .07 F1 Score. The largest contribution is from the post-hoc extraction. Although this module has low precision and actually decreases the overall precision of the method, it increases the recall even more and the contribution to the F1 Score is positive. The precision of both anaphora modules is quite high, especially that of the Nominal anaphora resolution, but only a limited amount of items were found by these modules, yielding only a small overall increase. It is probably worthwhile to investigate existing anaphora resolution systems, such as those described by Bouma (2003) and Mur and van der Plas (2006), to see whether recall can be improved.

As noted above, this technique cannot be used to judge the contribution of the serial modules. However, it is important to have an indication of where in the 'pipeline' the problems occur, as an error in an early module generally causes the later modules to fail too: if the parse of a sentence is incorrect, it is very difficult to extract the correct agent/patient relations. Therefore, while judging the acceptability of the method on 200 sentences as described above, we also indicated our judgement of the root cause of the problematic sentences. This judgement was based on an inspection of the parse tree, the extracted source constructions and subject-object relations, and the actors detected by the Entity Identification component. The result of this (qualitative) inspection is listed in table 6.11. For each module, the percentage of errors attributed to that module are given, and a summary of the causes of failure.

Although the error frequencies are based on a small sample, there is no single component that is identified as the main cause of errors. As such, there is no single bottleneck that can be tackled. Rather, a number of different possibilities for improving the method were found, giving us confidence that it is possible to increase the performance of the method within the current framework.

Another analysis is correlating the errors with the characteristics of the sentence. For this analysis, we computed the number of errors and percentage of incorrect items for each sentence, and correlated this with the length of the sentence and a dummy variable indicating whether the sentence was the headline. The results of this are interesting: Although longer sentences contain more errors, they actually contain even more correctly found items. Since the parse tree of long sentences is generally complicated and error-prone, this is a surprising finding. Additionally, the headlines contained fewer errors than other sentences, both absolutely and as a percentage of the total number of items. Given that head-

Table 6.11: Frequency of components causing errors and summary of the causes of failure

| Cause (frequency) | Explanation |
|----------------------------|---|
| Tokeniser (11%) | Newspaper sentences often contain hyphens that have to be interpreted either as a sentence boundary marker or as a connector in compound words. Additionally, the quote character can be used for citations or for apostrophes in words. The parser expects this decision to be made by the tokeniser, and generally fails spectacularly if this decision is made incorrectly. |
| Parser (5%) | Especially on headlines, where auxiliary verbs and function words are frequently omitted. Long sentences are often not parsed completely, but generally usable. |
| Source constructions (15%) | The most frequent error seemed to be multi-sentence quotes, where the source of the quote is not mentioned in the present sentence. |
| Subject-Object (20%) | The most frequent error made by the subject-object pattern matcher was that in complex sentences the agent and patient are often spread over the parse graph, rather than being the arguments of a single predicate. Also, a grammatically identical structure can have a different agent-patient structure depending on the words, i.e. the sentence is ambiguous unless considering the lexical level, which current module does not. |
| Entity (24%) | Errors here are mainly unusual labels for politicians and parties (such as 'JP' for prime minister Jan Peter Balkenende), and descriptions of the political offices held by politicians (such as 'onderwijsminister' (<i>education minister</i>)). |
| Anaphora (25%) | In quotes, the first person pronoun should be resolved to the source of the quote. Also, NP-anaphora can often be difficult to spot, for example, 'die drie' (<i>those three</i>) or 'die' (<i>this</i>) referring to a person. |

N=75 errors found in 51 incorrectly analysed sentences

lines and complex sentences are generally difficult to parse, this seems surprising. The reason is probably that headlines and long sentences are generally self-contained, whereas short sentences in news texts (e.g. “He was not the only one”; “Their success depends on this effort”) often need information from other sentences to be interpreted. Human coders do this naturally, but the module for anaphora resolution probably falls short of resolving political entities in such short sentences.

Finally, it is interesting to investigate whether the performance of the method is related to the type of relations that are extracted. In the semantic content analysis method NET, a number of different relation types are distinguished. Table 6.12 lists the relative frequency of sentences for each type, giving an example sentence to explain the statement types, and the proportion of errors in statements of that type.

These percentages are fairly close for all types. However, it is notable that sentences describing Success and Failure contain a disproportionate amount of errors. This is probably because the coding instruction for human coders prescribes coding the actor whom it concerns in the object position of a Success and Failure statement. In the text, however, the actors are just as often the grammatical subject of the sentence. For example, the sentence “The democrats gained 3 seats” would be coded by human coders as [reality / gain of 3 seats for / Democrats]. This suggests that a special-purpose module for detecting such sentences might aid overall performance.

Table 6.12: Relation types and performance

| Statement Type | Example sentence | % of sentences | % of errors |
|------------------|-----------------------------|----------------|-------------|
| Success, Failure | Victory for Balkenende | 15 | 19 |
| Evaluation | Wilders is evil | 17 | 15 |
| Action | PvdA voted against the bill | 16 | 13 |
| Affection | VVD disagrees with PvdA | 44 | 45 |
| Other | - | 3 | 4 |

N=10,367 sentences coded with a single relation type

6.6 Discussion / Conclusion

In this chapter we presented a method for extracting semantic relations between politicians from Dutch newspaper articles. This method uses the syntactic structure output by a freely available syntax parser (Van Noord, 2006) to identify source constructions and semantic agent-predicate-patient relations. Subsequently, the method identifies the political

actors in those relations using a list of the names of all relevant politicians. Moreover, the method contains an anaphora resolution module that resolves pronominal and nominal anaphora.

In order to assess the performance of automated Semantic Network Analysis, its output was compared with human codings of the same articles. Four types of performance were considered. The concurrent validity of the method was measured by comparing the agreement between the output of the method and human codings at the level of articles (the unit of measurement). This resulted in a moderate F1 Score of 0.65. Since this moderate validity can be due to semantic ambiguity as well as error, we measured the face validity by manually judging the acceptability of the output at the sentence level, which yielded a good F1 Score of 0.8. Although this evaluation is hardly rigid and objective as it was carried out by the authors, it suggests that an automated analysis may still deliver justifiable codings, even if these codings were not preferred by human coders.

As argued by Krippendorff (2004, p.292), something that is more important than the agreement at the level of the units of measurements (i.e. sentences) is the agreement at the level of the units of analysis, which can be called construct validity: will the theoretical constructs be measured correctly? Since this requires a theoretical construct, we performed a case study investigating media authority. The construct validity was then determined by computing the agreement on the media authority for the units of analysis, being combinations of parties, issues and media. This yielded a good correlation of .83 and .61 for two different definitions of authority, frequency of being quoted, and proportion of being acting rather than acted upon. Finally, we determined the predictive validity of the method by comparing the conclusions on three substantive hypotheses regarding authority based on the automated coding and the manual coding. The predictive validity was high, since the test results starting from either manual codings or automated codings were identical, with almost identical regression estimates. Substantively, both a human and the automated Semantic Network Analysis confirm that media authority is linked with power (indexing theory; Bennett, 1990) and that newspapers lend more authority to politicians in the traditional ideological 'pillar' of the newspaper (opportunistic witnesses; Hagen, 1993; Ruigrok, 2005). Moreover, human and computer coding both reject the link between issue ownership and authority, confirming the findings of Sigelman and Buell (2004). Even in a multivariate test, which adds the risk of the test results differing because of a different assessment of the multicollinearity between the independent variables, automated and human semantic network analysis obtained the same results. From this, we conclude that the automated Semantic Network Analysis as described in

this chapter is immediately useful for extracting semantic relations that can be used for political analysis of questions that are difficult to answer with thematic content analysis.

CHAPTER 7

Determining the valence of relations using Sentiment Analysis

'Press subtly tries to demonise Geert Wilders'

(Pers wil Geert Wilders subtiel demoniseren; Metro, August 21, 2007)

'All were about media logic and mediacracy.

And all were negative: the media did it.'

(Allemaal over medialogica of mediocratie. En allemaal negatief: de media hebben het gedaan.
de Volkskrant, Communicatiestaat, September 18, 2007)

It is often said that news is becoming more negative; that politicians are taking opportunistic positions on issues; and that the media are biased in their coverage of political news, supporting their favored party or demonizing newcomers. Analyzing any of these claims requires measuring whether an aspect of the discourse is positive or negative. This chapter presents a system using Sentiment Analysis methods to automatically determine this, extending automatic Semantic Network Analysis to positive or negative relations between actors and issues.

This chapter is an edited version of:

Wouter van Atteveldt, Jan Kleinnijenhuis, Nel Ruigrok, and Stefan Schlobach (2008), *Good News or Bad News? Conducting sentiment analysis on Dutch text to distinguish between positive and negative relations*, in C. Cardie and J. Wilkerson (eds.), Special Issue of the Journal of Information Technology and Politics on "Text Annotation for Political Science", vol. 5 (1), pp. 73–94

7.1 Introduction

The previous chapters discussed techniques to extract and interpret undirected and directed relations between concepts. This chapter will show how Sentiment Analysis techniques can be used to determine the polarity of relations and descriptions, that is, whether these relations or descriptions are positive or negative.

Automatically determining the polarity of relations and descriptions is not an easy task. Polarity can be expressed using verbs like ‘support’, adjectives like ‘good’, or nouns like ‘a winner.’ Often, whether a word has a positive or negative meaning is dependent on context, such as ‘cool relations’ versus ‘cool plans’; a special case of this is multi-word units such as ‘to push one’s buttons’ or ‘to lead up the garden path,’ which contain a negative sentiment even though the individual words are generally neutral. To make matters worse, positive and negative expressions contain more infrequently used words than non-polarised text (Wiebe et al., 2004), making it very difficult to create word lists for such expressions: Manually created word lists generally contain omissions, and the more rare words a category contains, the more difficult it is to create an exhaustive list. Infrequent words are also likely to be excluded from the manually coded training material that can be used for creating word lists automatically. As a consequence, there are currently no automatic Semantic Network Analysis methods that extract a polarised network from text. Approaches to automating the extraction of positive or negative relationships are often based on counting positive words on the one hand, and negative words on the other, such as in extracting issue positions from texts (Laver et al., 2003), in extracting evaluations (Fan, 1996), and in extracting real-world developments, such as attributions of economic success or failure (Shah et al., 1999). Schrodtt and Gerner (1994) use syntactic information for extracting relations, but restrict themselves to conflict and cooperation between actors in sentences with a limited syntactic complexity, such as headlines. Network content analysis methods inspired by Social Network theory (Wasserman and Faust, 1994) largely focus on the attention for specific relations between actors rather than their polarity (Diesner and Carley, 2004; Corman et al., 2002).

Within Computational Linguistics, recent years have seen the emergence of Sentiment Analysis, a field that aims to identify and classify subjective aspects of languages, especially expressions of positive or negative sentiment (Wiebe et al., 2004; Choi et al., 2006; Shanahan et al., 2006; Kim and Hovy, 2006). Sentiment analysis uses a variety of linguistic means such as elaborate thesauri, Part-of-Speech-taggers, lemmatizers, syntactic parsers, and Statistical Natural Language Processing methods to assess whether a text contains subjective sentiment and whether that

sentiment is positive or negative.

This chapter uses Sentiment Analysis techniques to automatically determine the polarity of relations between actors and issues in Dutch political newspaper articles. Specifically, we use a Machine Learning strategy with a number of lexical features based on an existing Dutch thesaurus and extracted automatically from a large uncoded corpus, using syntactic analysis to focus the Machine Learning on the predicate rather than the whole sentence. We use an earlier manual Semantic Network Analysis of the Dutch 2006 parliamentary elections (Kleinnijenhuis et al., 2007a, see section 2.5) to train the Machine Learning model and test the model at the sentence level. Finally, we validate the usefulness of the method for answering political research questions by replicating a number of analyses from the original study, comparing the results derived from the automatically extracted network with that derived from the manual coding on the level of analysis (e.g. a week of news about an actor) rather than the level of sentences.

The contribution of this chapter is threefold: Firstly, we present a method for automatically determining the polarity of relations between, and descriptions of, actors and issues in text. This is an important step in automating Semantic Network Analysis, and allows the automatic extraction of the data needed for many interesting political research questions. Secondly, we show that existing Sentiment Analysis methods can be used for extracting data that is relevant for answering political science questions. The existing Sentiment Analysis literature focuses on the ability to extract a linguistic phenomenon at the level of sentences, and we show that this can be used for analysing political phenomena at the level of political analysis. This serves as a use case and external validation for Sentiment Analysis techniques, and gives the political analysts an indication of the utility of these techniques for their research. Finally, Sentiment Analysis is generally focused on the English language, and although a number of papers apply these methods to other languages, this is the first explicit Sentiment Analysis study conducted on Dutch, showing that the methods developed for English can be translated to that language.

In the next section, we describe the relational content analysis method that we want to automate and formulate the tasks that the method needs to perform. This is followed by a brief summary of the relevant techniques in Sentiment Analysis, and an explanation of how we used these techniques to create our method. Subsequently, we present the performance at the sentence level and give a short analysis of which techniques performed well. Finally, we conduct the four case studies mentioned above, analysing the performance of the method at the level of analysis and showing its usefulness for political research.

7.2 Polarity in Political Communication

There are different ways in which polarity is important in political communication, having to do with the *relations* between actors and issues and their *evaluative* and *performance descriptions*. These relations and descriptions surface in a number of political studies. For example, the polarity of *relations between actors*, ranging from war to cooperation, is at the heart of international events research, starting from the seminal COPDAB-project (Conflict and Peace DATA Bank; Azar, 1980; Schrodtt and Gerner, 1994).¹ The polarity of *relations between issues*, ranging from negative to positive causation, or from dissociation to association, is the core of the cognitive map approach (Axelrod, 1976; Dille and Young, 2000). The polarity of *relations between actors and issues* is used to determine the issue positions of political actors (Kleinnijenhuis and Pennings, 2001; Laver et al., 2003; Laver and Garry, 2000).² The polarity of *evaluative descriptions* or (moral) judgements, such as 'Republicans are trustworthy' or 'Unemployment is awful' is the topic of evaluative text analysis (Hartman, 2000; Janis and Fadner, 1943; Osgood et al., 1956).³ *Performance descriptions* are statements about real-world developments or attributions of success and failure such as 'John is gaining in the polls', 'Unemployment is rising.' Although not always distinguished from evaluations, performance descriptions are in fact different from evaluations since the latter can often be seen as indirect expressions of relations between actors or issues: if John thinks Republicans are trustworthy, one can deduce a positive relation between John and Republicans, while John stating that the Republicans are winning in the polls has no implication for his opinion about that party. Although performance descriptions do not imply judgements, they can have an influence on opinions about the described or responsible actor. The polarity of the *performance descriptions* of actors, the attribution of success and failure, constitutes the core of attribution theory in psychology, and of the bandwagon effect (Lazarsfeld et al., 1944), political momentum (Bartels, 1988) and horse race news coverage (Iyengar et al., 2003) in political science.⁴ *Performance descriptions* of issue developments in the real world have been studied especially with regard to media reports of economic developments, such as reports about an increase or a decrease in employment (Hetherington, 1996; Soroka, 2006).

Most of these studies focus on a specific kind of relation or description. In some cases, multiple aspects of relations between actors and is-

¹See section 2.2.5 on page 22

²See section 2.2.4 on page 22

³See section 2.2.2 on page 20

⁴See section 2.2.3 on page 21

sues need to be considered. Diehl (1992), for example, argues that studying pro-con positions on salient issues can enhance the understanding of conflict and cooperation between states. Monge and Contractor (2003) argue for simultaneously testing theories at different levels — actors, dyadic and triadic patterns, and the whole network — to arrive at a better understanding of how these theories complement each other. Semantic Network Analysis, a branch of Content Analysis (Krippendorff, 2004, p.292), explicitly extracts both the attention for, and the polarity of, relationships between both actors and issues to arrive at a single network (Popping, 2000; Roberts, 1997; Van Cuilenburg et al., 1986). This yields content analysis data that are useful for studying a large variety of aspects of the coded texts. For example, Kleinnijenhuis et al. (2007b) show how news about issue positions, news about relations between political parties, and news about party performance each exerts a differential effect on the shift in party preferences during a political campaign.

Extracting the network of positive and negative relationships between actors and issues can be done manually, either with the text of the unit of observation, by asking coders what these relationships are after a careful reading of a text (Azar, 1980), or with the sentence as the unit of observation, by dissecting each sentence as one or more positive or negative relations between objects (Osgood et al., 1956; Van Cuilenburg et al., 1986). These processes are time-consuming and expensive, making it difficult to obtain large data sets, thereby impeding data-intensive research such as internationally comparative and longitudinal research. Typically, concessions are made through analysing only part of the texts, although such limitations may result in a loss of validity in the case of detailed research questions (Althaus et al., 2001). This makes it attractive to automate the extraction of positive and negative relations.

7.3 Task: Classifying NET relations

Given a relation or description expressed in a text, the task of the method presented in this chapter is to determine whether the relation is positive, negative, or neutral. Hence, this method is designed to work with the output produced by the method described in chapter 6. However, since we have a corpus manually analysed with Semantic Network Analysis, we can study the polarity separately from the identification of the relations by using the manually identified relations as input for the method described in this paper.

We define three different tasks on the basis of the different NET sentence types described in section 2.4. The *Relation* task is based on the sentence types that connect actors and issues: conflict, issue positions,

and issue causality. The *Performance* task uses the Success/Failure and (real-world) Developments sentences, which are both indicators of how well a person or issue is doing. The issue and actor evaluation sentences are used in the *Evaluation* task, in which a (moral) judgement is passed on an actor or issue. More formally, the three tasks are:

Relation Given two concepts located in a sentence, is their relation positive (cooperative, supportive), negative (conflictive, critical) or neutral?

Performance Given a concept located in a sentence, is it described as successful (increasing, winning), failing (decreasing, losing), or neither?

Evaluation Given a concept located in a sentence, is it evaluated positively (as good, sincere, beautiful), negatively (as evil, wicked, or ugly), or neutrally?

7.4 Sentiment Analysis

The work conducted in Sentiment Analysis (Wiebe et al., 2004; Shanahan et al., 2006) is highly relevant to our task. This field of Computational Linguistics aims at identifying and classifying subjective language, defined as the “language used to express *private states* in the context of a text or conversation” (Wiebe et al., 2004, p.5). This section will survey some of the techniques used in this field, on which we base the system described in the next section.

A number of Sentiment Analysis papers try to create lists of subjective words by starting with a ‘seed set’ of words with a known evaluative value, and then expanding this set. For example, Hatzivassiloglou and McKeown (1997) use “*Adj*₁ and/but *Adj*₂” patterns on a large corpus to cluster adjectives, assuming that *and* connects similar adjectives while *but* conjoins adjectives with opposite polarity. Hatzivassiloglou and Wiebe (2000) expand this system by adding gradable adjectives (adjectives that can be modified with a grading adverb such as *very*) as an indicator of subjectivity, and test whether these adjectives help in identifying subjectivity at the sentence level. Wiebe (2000) uses distributional similarity of syntactic relations to further expand this set. Two adjectives are distributionally similar if they appear in the same contexts, which in this case means having the same syntactic relations with other words (Lin, 1998). Wiebe et al. (2004) test various subjectivity clues, including unique words, N-grams, and distributional similarity, on a number of data sets. Baroni and Vegnaduzzo (2004) use “A near B” patterns using an Internet search engine to expand a seed set based on co-occurrence.

Finally, Riloff and Wiebe (2003) learn ‘extraction patterns’ from sentences containing known subjective words, creating lists of syntactic patterns such as a specific *verb – infinitive* or *active verb – preposition* combinations.

Word lists suffer from the inability to consider the specific context in which words are used. An alternative approach is to use a Machine Learning algorithm to discover patterns in large sets of training sentences, whose polarity has been manually annotated. In Machine Learning, two important choices are the learning algorithm and the characteristics of the text to use as features (or independent variables). For example, Wilson et al. (2005) use a program called BoosTexter, which uses decision rules to determine the polarity of words in context. As features, they use a thesaurus, words from the General Inquirer (Stone et al., 1962)⁵, the patterns from Riloff and Wiebe (2003), and a number of syntactic features such as whether a word is in the subject or object clause. Breck et al. (2007) identify and classify subjective statements using a Support Vector Machine using words, the verb categories defined by Levin (1993), and the word lists derived by Wilson et al. (2005) as input features. Choi et al. (2006) train two different Conditional Random Field (CRF) models, one for extracting opinions and opinion sources, and one for determining the relation between the two. This second model is a zero order CRF (which is equivalent to the Maximum Entropy model used in this chapter) trained on a number of lexical and syntactic features, such as whether the sentence is active or passive, the syntactic path between the opinion and its possible source, and a number of specific patterns called ‘syntactic frames’ that can match the grammatical structure.

These papers all focus on the English language. Mihalcea et al. (2007) try to directly translate subjectivity clues from English to Romanian using an online dictionary, but this has limited success. Mathieu (2006) presents and evaluates a computational semantic lexicon of French emotive verbs. The NTCIR Information Retrieval workshop in 2006 had an opinion extraction task in Chinese and Japanese as well as English (Seki et al., 2007), leading to a number of papers focussing on these languages such as Kanamaru et al. (2007) and Xu et al. (2007), who use Machine Learning methods for subjectivity in Japanese and Chinese texts, respectively. To the knowledge of the authors, no explicit Sentiment Analysis work has been performed on Dutch, although there is related work such as an investigation of subjective verbs (Pit, 2003) and work on automatically expanding lexical resources (Tjong Kim Sang and Hofmann, 2007).

⁵See section 2.6 on page 35

7.5 Method

The method described in this chapter will use a Machine Learning approach similar to the work discussed above (e.g. Wilson et al., 2005; Choi et al., 2006; Breck et al., 2007). In section 7.3, we defined three polarity classification tasks: classifying *relations*, classifying *performance* descriptions, and classifying *evaluative* descriptions. For each of these tasks we train and test a Maximum Entropy model (Berger et al., 1996), based on lexical and syntactic features. Maximum Entropy models are log-linear models built to maximise the entropy in the model within the constraints set by the training data that have been used successfully for a number of Natural Language Processing tasks (e.g. Abney, 1997; Ratnaparkhi, 1998), including the work by Choi et al. (2006) described above. Nonetheless, other Machine Learning methods such as Support Vector Machines or higher order Conditional Random Fields could have been used as well, and it would be interesting to test whether higher performance can be attained with other methods.

In Machine Learning, the learning algorithm is presented with a set of cases together with their actual class (the polarity according to the manual analysis). From these cases, called the training data, the learning algorithm creates a model of the relation between characteristics of the input data and the class. This model is subsequently tested on the test data: a set of cases not used in training with the actual class hidden from the model. Comparing the class assigned by the model with the actual class gives an indication of the performance of the model. The remainder of this section describes which features (characteristics of the input data) are considered, the strategy for collecting these features from the text, the procedure and measures used to test performance, and the corpus that is used for training and testing.

7.5.1 Features

An important choice in using Machine Learning models such as Maximum Entropy is which characteristics of the text, called features, are given as input to the model. Since the model can only use information contained in these features, the choice of features strongly influences the performance of the model. Model features are similar to the independent variables in the statistical modelling such as regression analysis, but the focus of Machine Learning is on finding the best model, not on understanding the underlying phenomenon. Consequently, the value of the parameters attached to the features are generally not of interest and Machine Learning models can have very large numbers of features.

Below, we list the features that are used in our model. The first two

feature groups are based on the output of linguistic preprocessing of the text, such as lemmatizing and parsing. A problem with such features is the *data scarcity problem*: it is quite likely that a word used for expressing polarity has not been encountered in the training data. To overcome this problem, we included lexical information to group words with similar meaning together: the third feature group is based on existing lexical information in the form of a thesaurus, while the last three feature groups are based on finding clusters of similar words in a reference corpus that has not been manually analysed. In each of the descriptions below, the feature set representing the example description ‘the young senator’ will be used to illustrate the discussed features.

Lexical and POS Features

Similar to Choi et al. (2006), we use the frequencies of lemmata and Part Of Speech (POS) tags as reported by the Alpino parser as features (Van Noord, 2006)⁶. For the example description, this would yield {lemma:the, lemma:young, lemma:senator, pos:Determiner, pos:Adjective, pos:Noun}

Syntactic and Surface bigrams

We use all adjacent lemma pairs in the selected part of the sentence as features. Moreover, we include the syntactic dependency relations between words reported by Alpino (Van Noord, 2006) as features: in a sentence like ‘John trusts Republicans,’ John is the subject of trusts and Republicans are the object of trusts, yielding the dependency relations *John-subject-trust* and *Republicans-object-trust*. In the example description ‘The young senator’, this surface and syntactic bigrams are {bigram:the_young, bigram:young_senator, dependency: the-determiner-senator, dependency:young-modifier-senator}

Brouwers thesaurus

Brouwers (1989) thesaurus is a manually created general-purpose Dutch thesaurus, comparable with Roget’s Thesaurus for English (Kirkpatrick, 1998).⁷ We look up all lemmata in the thesaurus, and use the resulting categories as features. For example, the word ‘young’ falls into Brouwers’ category *age* and *incompetent*, and ‘senator’ is categorised as *authority* yielding the features set {brouwers:age, brouwers:incompetent, brouwers:authority} for the example description.

⁶See section 3.2 on page 43

⁷See section 3.3 on page 48

Mutual Information on an uncoded Corpus

The intuition behind co-occurrence based methods such as mutual information is that words that frequently occur together probably have similar meanings. For this feature, we create clusters of words based on pointwise mutual information on an uncoded corpus similar to the work by Grefenstette et al. (2006) and Baroni and Vegnaduzzo (2004). Specifically, for each pair of words belonging to the same category (noun, verb, adjective/adverb), we determined the number of documents containing either or both terms, and calculate the mutual information as the log of the intersection divided by the product of the individual document counts. We then transform this to a distance metric by subtracting from the theoretical maximum $\log(|D|)$, yielding:

$$\begin{aligned} dist_{MI}(w_1, w_2) &= \log(|D|) - \log\left(\frac{|D| \cdot |D_{w1} \cap D_{w2}|}{|D_{w1}| \cdot |D_{w2}|}\right) \\ &= \log\left(\frac{|D_{w1}| \cdot |D_{w2}|}{|D_{w1} \cap D_{w2}|}\right) \end{aligned} \quad (7.1)$$

Using this distance metric, we created 500 word clusters using a K-means clustering algorithm, and used each of these clusters as a feature. In the example description, suppose ‘young’ is contained in cluster 131 and ‘senator’ is in cluster 265, we would get the feature set {mutual:131, mutual:265}.

Distributional Similarity based on Syntax Trees

Whereas co-occurrence is based on two words appearing in the same document, distributional methods are based on two words appearing in similar contexts. Following Lin (1998) and Wiebe (2000), we constructed a classification using the distributional similarity of the syntactic relations entered into by adjectives. In particular, we computed the distance between pairs of adjectives based on the cosine of the relationship frequency vectors for each adjective. Similar to the mutual information feature, we used this distance to create 500 clusters that are used as features.

$$dist_{DS}(w_1, w_2) = \frac{\sum_{r \in relations} fr(w_1, r) \cdot fr(w_2, r)}{\sqrt{\sum_r fr(w_1, r)^2 \cdot \sum_r fr(w_2, r)^2}} \quad (7.2)$$

(where *relations* is the set of all (syntactic relation, object) pairs, and $fr(w, r)$ is the frequency with which w is the subject of the relation r)

Conjunction patterns on an uncoded corpus

A problem with using distributional methods for determining polarity is that antonyms often occur in similar contexts. Similarly, subjective texts often contain a large number of both positive and negative words, making co-occurrence based methods difficult. Hatzivassiloglou and McKeeown (1997) explicitly look for words with the same polarity in an uncoded corpus by looking for conjunctions of adjectives using ‘and’ or ‘but’, relying on the fact that words of different polarity cannot be conjoined by ‘and’ (**a corrupt and legitimate regime*) and vice versa for ‘but’. We applied this to a corpus of uncoded text, looking for “.. en ..” (*and*) and “.. maar ..” (*but*) for all pairs of adjectives and verbs. From this we compute a distance metric as follows:

$$dist_{CP}(w_1, w_2) = \frac{1}{1 + e^{\frac{1}{10} \cdot (|w_1 en w_2| - |w_1 maar w_2|)}} \quad (7.3)$$

7.5.2 Strategies for Feature Collection

The features described in the previous section are all collected from words and word pairs in the text containing the relation or description to be classified. Since a sentence can contain multiple relations or descriptions, it might be better to collect features only from the part of the sentence containing the relation or description rather than the whole sentence. This section describes three strategies to focus the feature collection on the relevant part of the sentence.

Strategy 1: Sentence The first strategy simply collects features from the whole sentence, functioning as a baseline.

Strategy 2: Predicate In the second strategy, feature collection is restricted to the predicate expressing the relation or description. For the *relation* task, we define the predicate as being all nodes on the direct path between the subject and object in the dependency tree, and all modifiers and related verbs of these nodes. For the *performance* and *evaluation* tasks, the predicate comprises all nodes directly connected to the target node, and all modifiers of these nodes. As an example, consider the fictive sentence “De jonge senator lanceerde een persoonlijke aanval op de premier” (*The young senator launched a personal attack on the Prime Minister*), of which the dependency graph produced by Alpino (Van Noord, 2006) is given in figure 7.1. ‘Senator’ and ‘premier’ (*Prime Minister*) are identified as actors by the pre-processing (in this case the manual coding). To determine the predicate expressing the relation between these actors, we take the short-

est path between them through the dependency graph: ‘lanceer aanval op’ (*launch attack on*). Subsequently, this set of words is expanded by adding all their modifiers and auxiliary verbs, yielding “lanceer een persoonlijke aanval op” (*launch a personal attack on*). For the evaluation or performance description of the first concept, senator, we first select all direct parents and children: “lanceer jong de” (*launch young the*). This set is then expanded with all modifiers of these words, in this case none.

Strategy 3: Combination The third strategy is a combination of the other two strategies. It creates two distinct sets of features: one for the predicate and one for the remainder of the sentence. For example, the combination strategy applied to the performance description of senator in figure 7.1 would have separate features for ‘lemma young inside predicate,’ which would have value one, and ‘lemma young outside predicate,’ which would be zero. The lemma ‘personal,’ which is excluded in the *predicate* strategy, is included in the ‘out of predicate’ set in this strategy. This gives the Machine Learning algorithm access to the part of the sentence outside the predicate while still allowing the model to focus on the features in the predicate, for example by giving a higher weight to specific words in the grammatical context of the evaluation, performance description or relation.

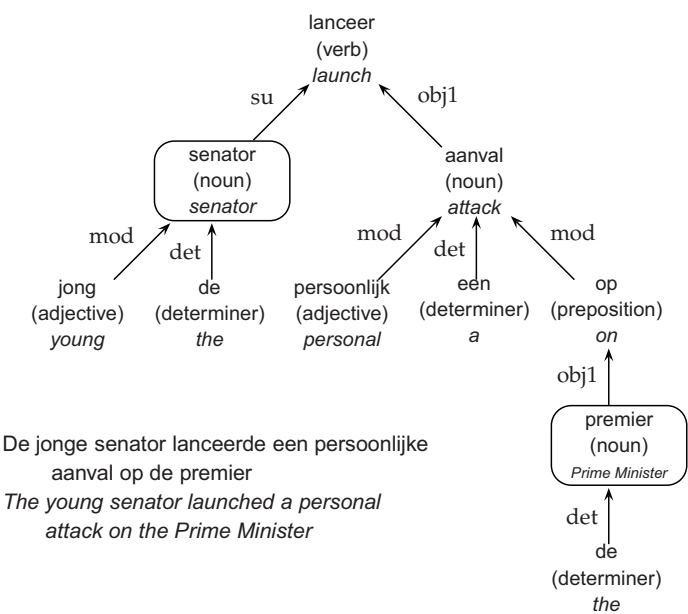


Figure 7.1: Grammatical analysis of example sentence

7.5.3 The Corpus

The main corpus used in the chapter consists of the manual Semantic Network Analysis of the coverage of the Dutch parliamentary election campaign of 2006 described in section 2.5. Unfortunately, although the relations in the manual analysis were connected to a specific sentence, they did not specify which words in a sentence represent the concepts used in the coding. In order to extract features from a relation or description, we need to know which words represent the used concept(s). Therefore, we used only the concepts where we could find a word in the sentence that matched the label of the concept. The precision⁸ of this matching is high, a manual evaluation of a small sample (N=61) indicates a precision of 97%. Unfortunately, the recall is low (59% on the same sample), especially since many different words can be used to refer to the same issue (recall on political actors was 77%, while on issues and other actors it was 51%). As a result of this limitation, our actual corpus consists of 5,348 out of 16,455 relations where both subject and object could be matched, and 3,025 (out of 5,816) performance descriptions and 2,316 (out of 4,722) evaluations where the described concept could be matched. Although this is a small sample, we have no reason to assume that it is biased except for the fact that actors will be overrepresented. Hence, we expect the results on this sample to generalise fairly well.

In the coding instructions for the manual analysis, coders were told to omit neutral relations, so the corpus does not contain explicit neutral cases. Since it is important that the method can distinguish between neutral and polarised statements, we created these statements from the polarised statements as follows: For the *relation* task, we added a neutral relation between every pair of identified concepts between which no relationship was coded. For the *performance* and *evaluation* tasks, we added a neutral statement for a random subset of concepts that were not coded with a performance or evaluation. This resulted in 3,359, 3,292, and 3,466 neutral statements for the relations, performance descriptions, and evaluative descriptions, respectively.

Reference Corpora

In addition to this coded corpus, we use two reference corpora as sources of uncoded material for the last three features described above. The first reference corpus consists of the fully parsed sentences of the non-lead paragraphs of the 2006 election corpus, which were not manually analysed. This corpus is used for the *Distributional Similarity* feature.

⁸See section 3.4 on page 49

In order to use the *Mutual Information* and *Conjunction Patterns* features described above, we also use an unparsed, uncoded reference corpus. This corpus consists of articles from the Dutch elections in 1994, 1998, 2002, and 2003 (600,000 articles), a broad range of articles on government policy and media exposure in 2003 and 2004 (1.4 million articles), a stratified sample of all news in 2006 and 2007 (60,000 articles), and the news surrounding two recent events (the referendum campaign for the European Constitution and the news on two anti-immigration populists, Geert Wilders and Rita Verdonk; 63,000 articles). In total, this corpus consists of around 750 million words in 2 million articles.

7.6 Results

Table 7.1 lists the performance of the models trained for each of the three tasks defined in the Task section. The *Full model* row gives the F1 Score of the model using all features described above and using the best strategy for that task. As will be shown below, for the *relation* task this was the *combination* strategy, while surprisingly the *sentence* strategy was best for the evaluation and performance tasks. Apparently, the predicate definition for descriptions does not contain the relevant information for determining the sentiment. For comparison, the performance of two baseline models is also given. The *Guess baseline* is the performance attained by blindly guessing each category with a chance proportional to its frequency in the whole corpus. The *Lemma baseline* is a more severe baseline, consisting of a Maximum Entropy model trained using only the lemma frequencies in the whole sentence as features. Precision and recall were very close to the F1 Score and to each other for each of the results presented in this section, so to avoid redundancy they are not reported.

The full model improves on both baseline models for each task, performing between .09 and .025 better than the Lemma baseline. Although this improvement is not very high, it is highly significant according to a

Table 7.1: Performance of the full model and baselines (F1 Scores)

| Model | Relation | Performance | Evaluation |
|----------------|----------|-------------|------------|
| Guess baseline | 0.114 | 0.134 | 0.154 |
| Lemma baseline | 0.541 | 0.534 | 0.534 |
| Full model | 0.631*** | 0.559*** | 0.580*** |
| N | 8,681 | 5,988 | 5,773 |

***Significantly better than ‘Lemma’ baseline at $p < 0.001$ (t-test on 20 folds)

t-test on difference of means. Total performance on the *Relation* task is 0.631, which is comparable to the results reported by Wilson et al. (2005): they report an average F1 Score of .728 on identifying whether a relation is polarised, and .628 on classifying the resulting relation (ignoring their extra category of sentences that contain both positive and negative polarity).

7.6.1 Feature Contributions

As stated in section 7.5, we use three different strategies for collecting features from the sentences: the *sentence*, *predicate*, and *combination* strategy. In order to test which of these strategies performed best for each task, we calculated the performance of the model using all features from the three different strategies. The top three rows in table 7.2 presents the results of this comparison. For each task, the F1 Score of the best performing strategy is set in boldface, and for the other strategies the performance is listed along with the significance of the difference from the best model. For the relation task, the model using separate features for the predicate and the remainder of the sentence is significantly and substantially better than using the whole sentence. This indicates that the syntactic analysis is useful to determine the predicate that expresses the relationship using this information. Looking at the performance and evaluation statements, the model using the whole sentence is the best model. For performance statements, this difference is not significant with respect to the ‘combined’ model, while for the evaluation statements both differences are significant. From this, we conclude that our definition of the predicate for these descriptive statements is probably suboptimal, so improving this definition can be beneficial for the classification of the descriptive statements.

The lower part of table 7.2 lists the contribution of the different feature sets to the full classifier. For each feature, we have calculated the performance decrease if we leave that feature out, and list this difference and the significance of this difference. This is a rather strict test because of the overlap between the different features, so the individual scores are quite low.

The first five features do not paint a very clear picture. For the *relation* task, the individual lemmata can be left out without decreasing performance, indicating that the information in the lemmata is captured in the other features. For the other tasks, however, the lemmata are the largest contributors. The reverse holds for the Part-of-Speech and surface bigram features: for the *relation* task they are the strongest contributors, while for the other tasks they contribute less or not at all. Interestingly, Brouwer’s thesaurus works well for the *performance* and *evaluation* tasks,

Table 7.2: Performance of strategies and features (F1 Scores)

| Strategy | Relation | Performance | Evaluation |
|--------------------------|--------------|--------------|--------------|
| Sentence | 0.562*** | 0.559 | 0.580 |
| Predicate | 0.603*** | 0.521*** | 0.479*** |
| Combined | 0.631 | (0.556) | 0.546*** |
| Feature | Relation | Performance | Evaluation |
| Lemmata | (0.003) | 0.016*** | 0.016*** |
| POS | 0.022*** | (−0.004) | 0.010** |
| Bigram | 0.015*** | 0.009* | (−0.003) |
| Dependency | 0.012*** | (0.007) | 0.006* |
| Thesaurus | (−0.003) | 0.016*** | 0.011*** |
| Distr. Sim. (Adjectives) | (0.004) | (−0.007) | (0.000) |
| Distr. Sim. (Nominals) | 0.007** | (−0.004) | (−0.006) |
| Distr. Sim. (Verbs) | 0.007* | (0.006) | (0.001) |
| Mutual Information | 0.014*** | −0.010* | (−0.002) |
| Conjunction Patterns | 0.012*** | (−0.007) | 0.006* |
| N | 8,681 | 5,988 | 5,773 |

The best model is set in boldface in the strategy rows. Cells in the features rows indicate performance degradation when leaving that feature out. Significance of differences based on t-test with 20 folds.

***Significant at $p < 0.001$ level; **Significant at $p < .01$ level; *Significant at $p < .05$ level; (..) not significant

but does not contribute significantly to the *relation* task. Looking at the last five features, which are all based on word similarities using the uncoded corpus, we see that for the *relation* task all features contributed significantly. The largest gain came from the simplest method, the mutual information based on proximity queries, and the conjunction patterns also scored well. For the distributional similarity features, which are specified per Part Of Speech, we can see that the contribution of the adjectives is barely significant, while the contributions of the verbs and nouns are highly significant. This is not surprising, since relationships between concepts will often be expressed using verb phrases and noun-verb combinations such as “give support” or “pick a fight”. It does underscore that the traditional focus on adjectives in Sentiment Analysis might not be suited for determining the polarity of relations. For the descriptive tasks, the clustering methods perform worse: for evaluations only the conjunction patterns improve significantly, while for the performance task the mutual information even decreases performance significantly.

Table 7.3: Performance of the model on the different classes (F1 Scores)

| Class | Relation | Performance | Evaluation |
|----------|----------|-------------|------------|
| Negative | 0.818 | 0.466 | 0.591 |
| Neutral | 0.697 | 0.759 | 0.776 |
| Positive | 0.576 | 0.473 | 0.362 |
| N | 8,681 | 5,988 | 5,773 |

7.6.2 Error analysis

In order to improve the performance of the model, it is interesting to see whether we can detect patterns in the errors made by the model. Table 7.3 lists the performance of the model on each target class. This paints an interesting picture: for the *relation* task, the conflicts or negative issue positions are easier to detect than the neutral or positive ones. Possibly, the language in which criticism is expressed is less ambiguous than that in which positive sentiments are expressed. For the *performance* and *evaluation* tasks, the picture is different: the performance on the neutral category is much higher than that on the other categories. Apparently, detecting sentiment in these descriptions is easier than classifying the sentiment. Worst performance was attained on positive evaluations, which is probably due to the low frequency of these statements (11%, see table 7.4 below).

Table 7.4 lists the confusion matrix per task in table percentages. Each cell contains the percentage of cases that belonged to a certain class according to the manual coding and were assigned a certain class by the model. The bottom row and last column for each task show the total per-

Table 7.4: Confusion matrix per task (table percentages)

| Manual | | Model | | | | | | | | | | | |
|--------|----------|----------|----|----|-----|-------------|----|----|-----|------------|----|---|-----|
| | | Relation | | | | Performance | | | | Evaluation | | | |
| | | – | ± | + | Σ | – | ± | + | Σ | – | ± | + | Σ |
| – | Negative | 18 | 6 | 6 | 30 | 10 | 8 | 5 | 23 | 15 | 12 | 2 | 29 |
| ± | Neutral | 6 | 29 | 7 | 42 | 6 | 43 | 5 | 55 | 9 | 49 | 2 | 61 |
| + | Positive | 5 | 7 | 17 | 29 | 5 | 8 | 10 | 22 | 3 | 5 | 3 | 11 |
| Σ | Total | 29 | 42 | 29 | 100 | 21 | 59 | 20 | 100 | 27 | 66 | 7 | 100 |

centage of cases assigned to a class according to the model and according to manual coding, respectively. For example, the first data column shows that 29% of all cases in the relation task were assigned the ‘negative’ class by the model, out of which 18% were also assigned that class by the manual coding. The remaining 11% were divided over classes assigned the neutral (6%) and positive (5%) class by the manual coding. This table shows that there is no systematic bias in the errors made by the model: the marginal distributions of the predicted classes are very similar to the marginal distributions of the actual classes. Moreover, for every task and target class, the mistakes seem divided over the other classes according to the marginal distribution.

Finally, table 7.5 lists the performance of the model for each statement type, following the NET statement types described in section 2.4 with an additional distinction between political actors (ministers, parliamentarians) and other actors (such as citizens and pressure groups). In each task, the performance on statements involving political actors is highest followed by the performance on issues; performance on other actors is worst. Possibly, the language used in these cases is simply less explicit or more diverse, but it is also possible that it is an effect of the higher frequency of political actors, caused in part by the overrepresentation of actors due to the label matching problem described in the section 7.5.3. This suggests that it could be interesting to either train separate models for the different actor types, or use features to allow the model to distinguish between them.

Table 7.5: Performance on different statement types

| Task | Statement type | N | F1 Score |
|-------------|----------------------------------|-------|----------|
| Relations | Conflict (politicians) | 3,420 | 0.680 |
| | Conflict (other actors) | 2,258 | 0.592 |
| | Issue positions (politicians) | 1,276 | 0.552 |
| | Issue positions (other actors) | 836 | 0.477 |
| | Issue causality | 627 | 0.544 |
| | Other | 264 | 0.526 |
| Performance | Success / Failure (politicians) | 2,962 | 0.581 |
| | Success / Failure (other actors) | 608 | 0.467 |
| | Real world developments (issues) | 2,416 | 0.554 |
| Evaluation | Evaluation of politicians | 3,179 | 0.577 |
| | Evaluation of other actors | 1,227 | 0.523 |
| | Evaluation of issues | 1,366 | 0.524 |

7.7 Validation

In the previous section, we calculated the performance of our method by comparing the outcome from the trained model with the manual codings at the level of measurement. For political analysis, a much more important question than how well it performs on individual sentences, however, is how well it answers the questions it was designed for (cf. Krippendorff, 2004, p.243): measuring how actors and issues are framed and portrayed. Since the precise answer depends on the (political) research question, we will take a number of analyses performed previously on the Dutch 2006 campaign data (Kleinnijenhuis et al., 2007a), and test how well the results of these analyses based on the outcome of the model match the results based on manual analysis. Specifically, we will look at the overall tone of the news during the campaign, the issue positions taken by political parties, the patterns of conflict and support between parties in different periods, and whether newspapers differ in their attribution of success to the different parties.

7.7.1 *Overall tone of the news*

It is often claimed that news is becoming more negative, especially during campaigns (Patterson, 1993). In that light, it is interesting to look at the tone of the news operationalized as the average polarity of all statements. Figure 7.2 shows the graphs of the tone of the news for three news types: evaluations, issue positions, and conflict. For each graph, the two lines represent the results computed on the basis of the manual codings and on the output of the model.

In the topmost graph we can see that direct evaluations become more negative very slowly after the second week, going from -0.11 to -0.16. The issue positions also become less positive over time, meandering from around .3 in the first weeks to .1 in the last. Interestingly, the conflict news, defined as all relations between actors, seems to become more positive, going from -0.25 to around neutral. This is probably because the beginning of the campaign was characterized by a strong clash between the leaders of the PvdA (Social Democrats) and CDA (Christian Democrats), while in the last weeks there was a *détente* between the left-wing parties (see section 2.5 for a brief description of the 2006 elections). The lines of the manual coding and the model output follow each other very closely. In fact, the two are correlated with a coefficient of .9, albeit based on only three scores for 17 weeks ($N=51$). From this we can conclude that the current model is certainly capable of finding patterns in the overall tone of the news.

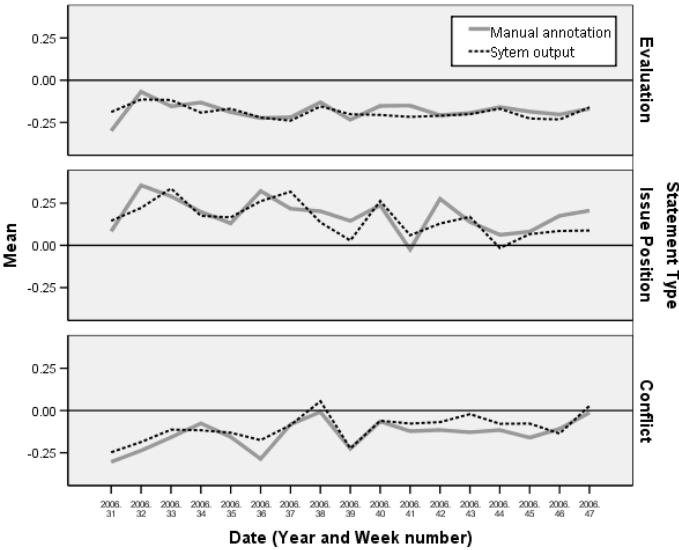


Figure 7.2: The overall tone of the news according to the manual coding and system output

7.7.2 Issue positions

Another aspect of political campaigns is the issue positions taken by different actors. There are a number of theories that predict a relation between (perceived) issue positions and voting behavior, such as Directional Voting (Rabinowitz and MacDonald, 1989) and Spatial/Proximity Voting (Westholm, 1997). This section analyzes party issue positions, replicating table 4.2 in Kleinnijenhuis et al. (2007a, p.75), although the actual results may differ since this is based on the smaller selection of matched sentences as described in the Corpus section.

Table 7.6 shows the average issue position of three parties on three issue categories, Leftist Issues (such as job security, welfare), Rightist Issues (such as taxation, defense), and Administrative Reforms (such as the referendum and elected mayor). For each issue category, three columns are given: the number of issue statements from the party on that issue, and the average polarity according to the manual analysis and according to the model. On the left hand side of the table, we see that the PvdA (Social Democrats) are in favor of leftist issues while the conservative VVD are against, with the CDA (Christian Democrats) taking the middle ground. Looking at rightist issues, the reverse happens: the VVD

Table 7.6: Issue positions on a number of issues (-1..1)

| Party | Leftist | | | Rightist | | | Reforms | | |
|-------|---------|------|-------|----------|------|-------|---------|------|-------|
| | N | Man. | Model | N | Man. | Model | N | Man. | Model |
| PvdA | 71 | 0.3 | 0.3 | 37 | -0.1 | 0.0 | 4 | 0.8 | 0.5 |
| CDA | 72 | 0.2 | 0.1 | 111 | 0.3 | 0.2 | 23 | 0.3 | 0.5 |
| VVD | 56 | -0.1 | 0.1 | 101 | 0.1 | 0.3 | 34 | -0.1 | 0.1 |

Example: There were 72 statements in which the CDA took a position on leftist issues; according to the manual analysis, the CDA was somewhat positive (+0.2), while according to the model output this is only +0.1.

is in favor and the PvdA against. There is a discrepancy here between the manual coding and the model: according to the manual coding, the CDA is slightly more rightist than the VVD, while the model places the VVD to the right of the CDA.

On the last issue presented here, Administrative Reforms, the PvdA is strongly in favor according to the manual coding while the more conservative CDA and VVD are slightly above and below zero. According to the model, however, the CDA is also quite positive, placing them side by side with the PvdA. An important example of a reform issue during the election was whether to hold a referendum on the new EU constitutional treaty, of which the PvdA was in favor. The greater divergence on this issue between manual coding and model output is probably due to the lower number of statements on which this is based, allowing for less room for individual errors.

Comparing the manual coding and our model, we see that they diverge, although they are generally in the same range. This is confirmed by a correlation analysis on the full selection of 7 parties and 14 issues, which shows a weighted correlation coefficient of 0.72 (N=98). Generally speaking, the results of the model will be similar to the results based on manual coding, but there will be small errors if one looks at the details, especially for the parties and issues receiving little attention; this latter point is reflected in the fact that the correlation coefficient that is not weighted for frequency is only 0.58. If we were to trace the development of issue positions over time, for example per week, the weighted and unweighted correlations also drop, to 0.58 and 0.54, respectively. Thus, in general the model is good enough to answer questions on party issue positions, but it performs less well on smaller parties and issues or for smaller time periods.

7.7.3 Political Conflict

An interesting part of campaigns in multi-party systems is the pattern of support and criticism between the parties, as the various parties balance discrediting and ignoring the other parties while also keeping possible future coalitions in mind. Here, we shall replicate the analyses presented graphically in figures 5.1-5.3 of (Kleinnijenhuis et al., 2007a, p.84-89), which show the network of party relations for three periods.

Table 7.7 shows a sample of this network based on the manual codings and output of the model. Each row represents the relation between two specific parties. For each relation, the number of statements (N) and average polarity according to the manual coding and according to the model is given for three time periods. The top two rows show the mutual relations between the CDA and PvdA, which were seen as the main contestants for becoming the largest party and had strong negative relations during the whole period. The third row shows the internal support and criticism within the PvdA. In the first period, the PvdA has internal problems after forcing two Turkish candidates to withdraw because they refuse to acknowledge the Armenian genocide. According to the manual codings, they continue to have some internal problems, although the model actually measures a moderate positive internal relation in the second period. The bottom rows show the relation between the PvdA and more extreme Socialist Party. The PvdA completely ignores the SP in the beginning while there is some sharp but low-frequency criticism during the second period. In the last period, relations turn positive, although the model measures it as being slightly less positive than the manual coding.

If we compare the full network based on the manual coding and out-

Table 7.7: Support and criticism in three periods (from -1 .. +1)

| Subject | Object | 1 Sept - 15 Oct | | | 16 Oct - 13 Nov | | | 14 Nov - 22 Nov | | |
|---------|--------|-----------------|------|------|-----------------|------|------|-----------------|------|------|
| | | N | Man. | Mod. | N | Man. | Mod. | N | Man. | Mod. |
| CDA | PvdA | 65 | -0.6 | -0.6 | 79 | -0.8 | -0.5 | 23 | -0.8 | -0.7 |
| PvdA | CDA | 78 | -0.8 | -0.6 | 70 | -0.6 | -0.5 | 34 | -0.7 | -0.4 |
| PvdA | PvdA | 47 | -0.2 | -0.3 | 25 | -0.1 | 0.6 | 7 | -0.1 | 0.2 |
| PvdA | SP | 0 | - | - | 10 | -0.7 | 0.3 | 17 | 0.6 | 0.4 |
| SP | PvdA | 3 | -0.3 | -0.3 | 3 | -1.0 | -1.0 | 7 | 0.7 | 0.4 |

Example: In the second period there are 79 statements in which the CDA expresses an opinion about the PvdA. According to the manual analysis, this opinion was strongly negative (-0.8), while according to the system output this was slightly less so (-0.5).

put of the model for these three periods, we find a weighted correlation of 0.77 (N=21), which is certainly acceptable. If we do this comparison per week rather than in three periods, this drops to 0.66 (N=119). From this we conclude that the performance of the model is certainly good enough to analyze party relations over fairly large time spans, and probably good enough for more detailed analysis.

7.7.4 Party Performance and Newspaper Preferences

During campaigns, a significant portion of the news is always devoted to the horse race: who is winning in the polls, who won the last debate, who has the best chances of becoming Prime Minister? Because of the bandwagon effect (Lazarsfeld et al., 1944), to be portrayed as being successful is often a self-fulfilling prophecy (Bartels, 1988). Hence, it is interesting to investigate which newspapers portray which parties as being successful or failing. This replicates the analysis presented in Kleinnijenhuis et al. (2007a, table 6.4 p. 104).

Table 7.8 shows the performance of the main parties according to three newspapers. According to the manual codings, *de Volkskrant*, a left-wing quality newspaper, portrays the CDA and PvdA as somewhat successful and neutral, respectively, while the model classifies in the reverse order. According to manual coding and the model, the VVD is depicted as failing. The popular conservative newspaper *De Telegraaf* portrays the CDA as being fairly successful and the PvdA as failing. The model completely misses this, portraying the CDA as neutral and the PvdA as successful. The left-wing confessional newspaper *Trouw* also portrays the PvdA as failing, but is also negative about the performance of the other parties. The model measures a less extreme failing for PvdA, and a slight success for the VVD.

It seems from this table that the performance of the model on this case

Table 7.8: Performance of the main parties according to three newspapers

| Party | de Volkskrant | | | De Telegraaf | | | Trouw | | |
|-------|---------------|------|-------|--------------|------|-------|-------|------|-------|
| | N | Man. | Model | N | Man. | Model | N | Man. | Model |
| CDA | 91 | 0.2 | 0.0 | 82 | 0.3 | 0.0 | 81 | -0.2 | -0.2 |
| PvdA | 82 | 0.0 | 0.2 | 40 | -0.5 | 0.2 | 34 | -0.6 | -0.2 |
| VVD | 79 | -0.2 | -0.3 | 67 | -0.1 | 0.0 | 50 | -0.3 | 0.1 |

Example: *De Volkskrant* contained 91 evaluative statements about the CDA. According to the manual analysis, these statements were on average slightly positive (+0.2), while according to the system output they were neutral (0.0).

is worse than for the other use cases, which is confirmed by a correlation coefficient of only .46. The most likely explanation for the difficulty the model has with this task is that the classifier for Performance statements was the weakest. This underscores the importance of optimizing the used classifier, even if performance on higher levels of aggregation can be higher than on the sentence level.

7.7.5 F1 Score and usefulness

The main conclusion of the results section is that the full model improved significantly over the simpler *lemma* baseline model, but left room for improvement. An interesting question is whether the .03 to .07 increase in F1 Score gained by adding all the features of the full model translates into a better answer to the political research questions. To answer this question, we have calculated the same correlations as presented above using the output of the *lemma* baseline (see section 7.6 above). The results of this comparison are presented in table 7.9. For each of the subsections in this section, we list the correlations as reported above and the analogous correlation achieved using the baseline model.

For the tone of the news, the full model outperforms the *lemma* baseline but both score >0.9. For the analyses of issue positions and political conflict, the performance of the baseline is substantially lower, and for the more coarse-grained analyses the correlation drops from a very acceptable 0.72 and 0.77 to a meagre 0.52 and 0.59. Interestingly, the baseline model actually outperforms the full model on classifying the success of parties per newspaper, although the performance of the full model on that use case was already fairly poor. These results show two things: firstly, that the full model presented here is an improvement over the Lemma baseline model even if compared on a higher level of analysis. Secondly, and more importantly, it shows that a relatively modest in-

Table 7.9: Correlations with gold standard of full model and baseline

| § | Analysis | Full model | <i>Lemma</i> Baseline |
|-----|---------------------------------|------------|-----------------------|
| 6.1 | Tone of the news | 0.933 | 0.907 |
| 6.2 | Issue positions (whole period) | 0.718 | 0.516 |
| | Issue positions (per week) | 0.575 | 0.423 |
| 6.3 | Political conflict (per period) | 0.774 | 0.591 |
| | Political conflict (per week) | 0.656 | 0.527 |
| 6.4 | Party performance per newspaper | 0.460 | 0.603 |

crease in F1 Score of .07 can lead to increases in correlation of 20 percentage points on a higher level of aggregation and can make the difference between being good and being not quite good enough.

7.8 Conclusion

This chapter presented a method that automates an important step in Semantic Network Analysis: the determination of the polarity of relations and descriptions, that is, whether those relations and descriptions are positive or negative. We analyzed three types of statements: relations between actors and issues; evaluations of actors and issues; and performance descriptions of actors and issues. Using techniques from Sentiment Analysis, we trained a Machine Learning model using a number of lexical features, and using syntactic analysis to focus the model on the specific relation or description rather than the whole sentence. The model was trained and tested on a manual Semantic Network Analysis of the Dutch 2006 parliamentary elections (Kleinnijenhuis et al., 2007a).

The direct comparison of the automatic analysis with the manual codings on the level of sentences showed that the method can reproduce these codings reasonably well, and significantly better than a baseline model based on only the lemma frequencies. The syntactic analysis improved the classification of relations but does not improve the classification of evaluations and performance descriptions.

In order to show that the method is useful for political research, we compared the automatic analysis with the manual Semantic Network Analysis at the level of analysis of politically interesting phenomena. This was performed for four different use cases taken directly from the original election study. We found that the method can immediately be used for determining the overall tone of the news, overall issue positions, and party conflict patterns in the periods used in the original analysis. For determining party performance according to the different newspapers, and for analyzing issue positions and conflict patterns in smaller periods, the performance of the method was lower, due to the underlying model performance and the high granularity (the low number of statements within one unit of analysis). These tests showed that it is important to validate automatic content analysis methods on the actual task to be performed rather than relying on sentence-level performance. Moreover, we showed that a modest increase in F1 Score at the level of measurement can lead to substantially better performance at a higher level of analysis.

To further improve the method, we analyzed the errors made by the method and suggest a number of ways to alleviate these. First, we can

increase the size of our training corpus by resolving the matching problems described above and by combining the corpus with other existing manually analyzed corpora, such as the other Dutch election campaigns, which have been analyzed using the same method since 1994. Second, the methods used to create clusters of similar words based on the reference corpus did not perform as well as we had hoped. Using a ‘seed set’ of words that are good indicators of polarity, we can create these clusters in a more focused manner, hopefully leading to better performance. Finally, since using the syntactic analysis to determine the predicate improved performance for the relations but not for the descriptions, we can probably improve performance by modifying which part of the sentence is contained in the predicate for the descriptions.

Although the performance of the method presented here leaves room for improvement, this chapter presents two meaningful contributions. First, this is the first study that applies Sentiment Analysis techniques to the Dutch language, showing that the techniques for English also work for Dutch and providing a baseline and infrastructure for future Dutch Sentiment Analysis research. Second, by duplicating a number of analyses from an existing election study, we provide external validation of the Sentiment Analysis techniques used, and give the Political Scientist insight into how well these techniques really perform at answering his or her research question.

Taken together, chapters 5 – 7 present the technical foundations for automating Semantic Network Analysis: chapter 5 shows how simple associative relations can be extracted and interpreted; chapter 6 uses syntax to differentiate between source, subject, and object, enriching the relations with directionality. This chapter further enriches the relations with valence or polarity, creating a directed and signed network between concepts. Each of these chapters shows how the increasingly complex network can be used to answer substantive communication research questions, and chapters 6 and 7 also test the performance of the method, both on the level of measurement and the level of analysis. This gives us confidence that the techniques are reliable, valid, and relevant.

Part III

Reasoning with Media Data

Using RDF to store Semantic Network Data

'Ayaan Hirsi Ali switches to the liberals; Who's next?'

(Wie volgt Ayaan Hirsi Ali naar de liberalen?; *de Volkskrant*, November 18, 2002)

'One day a socialist, the next day a liberal, it has happened before'

(De ene dag socialist, de andere dag liberaal, 't is meer vertoond; *Dagblad Rivierenland*, November 2, 2002)

To efficiently analyse Semantic Network data and combine different data sets requires formalising both the actual Semantic Network data and the background knowledge that connects the concrete objects in the text to more abstract objects used in research questions. This background knowledge is characterised by political roles that change over time and different categorisations needed by different research questions. This chapter shows how RDF can be used to formalise the media data and background knowledge.

This chapter is an expanded version of:

Wouter van Atteveldt, Stefan Schlobach, and Frank van Harmelen, *Media, Politics, and the Semantic Web: An experience report in advanced RDF usage*, In: E. Franconi, M. Kifer, and W. May (Eds.): ESWC 2007, LNCS 4519, pp. 205-219, Berlin: Springer

8.1 Introduction

An essential feature of Semantic Network Analysis is the separation of extraction and querying. In the extraction phase, whether conducted manually or automatically, a network is created that links concrete actors and issues, staying as close to the text as possible. In the querying phase, the communication research questions, which often contains abstract and complex, is answered by querying the extracted network. This separation makes it easier to share and combine data sets: On the one hand it allows different research questions to be answered using the same data, by performing different aggregation and combination steps in the querying phase. On the other hand it allows differences in domain — and hence in the concrete objects used in the networks — to be bridged by aggregating to the same abstract concepts. Ideally, we can create large shared Semantic Network repositories, for example containing election campaigns or other political news from different countries and time periods. Such repositories could enable different researchers to test different models on the same underlying data and without incurring the costs of data gathering for each study.

It is unlikely that such a data set will be created in a single effort. Even using automatic Semantic Network Analysis, for example using the techniques presented in chapters 5–7, gathering the data requires significant effort as well as local domain and linguistics expertise. Therefore, creating a large Semantic Network repository will require combining Semantic Network data obtained in different projects and presumably different research groups. This means that it has to be possible to combine network data from different sources containing different actors and issues, and analyse the resulting network in meaningful ways.

In this chapter and the following one, we show how the Semantic Web techniques described in chapter 4 enable us to represent Semantic Network data to allow combining, sharing, and analysing the data efficiently. This chapter describes a method for representing media data and background knowledge using RDF, a Semantic Web graph description language. This representation is split into two parts: the media data and the background knowledge. The main challenge in representing the *media data* — the extracted Semantic Networks between concrete actors and issues — is that these data consist of triples themselves, which means that we use RDF triples to describe Semantic Network triples. Section 8.2 describes possible solutions for this problem and presents the resulting data model used for representing media data. The *background knowledge* forms the link between the concrete actors and issues extracted from the text and the more general concepts used in research question: a text might mention Balkenende proposing a method to decrease youth unem-

ployment, while we are interested in statements about the Prime Minister on Economic Issues. Formalising the background knowledge to make this link involves dealing with the dynamic nature of politics: politicians switch political functions and start new parties. Moreover, depending on the research question, different ways of abstracting from the specific to the general will be needed, as for some questions youth unemployment needs to be considered a negative valence issue, and for other studies it is an economic issue. Section 8.3 describes how these requirements can be met in a political ontology. Section 8.4 describes the actual ontology that was used for the election study described in Kleinnijenhuis et al. (2007a) that supplied the corpus used in chapters 6, 7 and 9. Finally, section 8.5 will survey some possibilities in which the richer Web Ontology Language OWL could be used to extend the political ontology.

8.2 Representing Media Data: Statements about Statements

This section investigates whether RDF can be used to represent the Semantic Network data extracted using the NET method, the Semantic Network Analysis method described in section 2.4. From that description of the NET method, we can formulate the following requirements for a representation of NET Semantic Networks data:

Quantitative value In addition to different statement types, Relational Content Analysis often includes a quantitative indicator of the direction and strength of a relation. For example, The statement “.. *we should invest more*” is positive (+0.7) while the statement “*Hard confrontation Left and Right*” is strongly negative (-1).¹ Other examples are adding a measure of the weight and ambiguity of a relation. In Social Networks terms, graphs labelled with values are called signed and/or valued networks (Wasserman and Faust, 1994).

Article Metadata To trace the evidence for an analysis and for time-based analyses, it is necessary to attach metadata to coded sentences, including publisher and publishing date, location in the newspaper, and a link to the original article.

Extra Arguments Sometimes we need to code certain additional aspects of a relation. For example, in the sentence “*Bos and Balkenende attacked each other over poverty*”, we want to capture the topic of the disagreement as well as the fact that these actors disagree.

¹See section 2.4 on page 29 for the example article from which these sentences are drawn.

Quoted Sources The sentence “*Bos: ‘good health care costs money, so we should invest more’*” contains a positive causal relation between Investing and Health Care, but this relation is not directly stated by the newspaper but rather by a quoted source. These sentences need to be placed in different subnetworks, where they are distinct but accessible from the main network.

8.2.1 Enriching Triples

The problems described above all require enriching Semantic Network Analysis triples by adding extra information. This is difficult, as RDF is intended for describing resources, not triples: triples do not have URIs and hence cannot be used as the subject or object of other triples. We are not the first to signal this difficulty: MacGregor and Ko (2003) cite the need to enrich triples to describe event data, and a number of authors encounter the problem of enriching triples when they want to use RDF to describe RDF documents, for example for reasoning about provenance and trust (Carroll et al., 2005).

RDF(S) allows some form of adding information to existing triples. Trivially, we can replace each of the nodes in a triple by a node carrying more information, and point back to the original node. In RDFS, it is possible to do so transparently by making the new node a subclass or subproperty of the original node. Additionally, the RDFS specification includes a reification mechanism (Brickley and Guha, 2004). Essentially, an anonymous instance is made to represent the statement, and standardised vocabulary is used to define the subject, object, and predicate of the statement. The anonymous instance, being a normal node in the graph, can then be used in further statements. According to the definition, a reified statement is hypothetical, i.e. it does not imply the original statement.

Another solution is using the n-ary relation design patterns described in Noy and Rector (2005). This is similar to reification in that a new node is created that represents the relationship, but the reification vocabulary is eschewed since “in n-ary relations [...] additional arguments in the relation do not usually characterise the statement but rather provide additional information about the relation instance itself” (Noy and Rector, 2005). This has the same disadvantage as reification (the original triple semantics are lost) but additionally it has no formal meaning or standardised vocabulary.

To overcome these problems, a number of authors have suggested extending the notion of a triple to include a fourth place, often seen as a context marker (MacGregor and Ko, 2003; Carroll et al., 2005; Dumbill, 2003; Guha et al., 2004; Sintek and Decker, 2002). For example, Guha

et al. (2004) propose a context mechanism that explicitly assumes the context marker to indicate provenance, and they include a complicated system of lifting and aggregating mechanisms to combine RDF documents from different sources. At the other extreme, MacGregor and Ko (2003) and Dumbill (2003) support replacing triples by quadruples without restricting the interpretation of such triples.

A proposal that seems to be gaining support is Named Graphs (Carroll et al., 2005). This proposal also adds a fourth place to the triple and defines the semantics of this added element but does not prescribe the interpretation in the way that Guha et al. (2004) does. Named Graphs semantics allow for nested graphs and propose a predicate to indicate nesting. The main disadvantage of this method is that it is not standardised, lacking tool support and declarative semantics. Also, since the intended meaning of the context is the containing graph, Named Graphs add extra information to the whole statement rather than to the predicate, similar to reification.

The proposals for adding information to triples in the literature are diverse in nature. Part of the reason for this diversity is that the problems they are trying to solve are diverse. Specifically, there are two main factors concerning which the proposed solutions diverge: the *meaning* of the extra information with respect to the original triple; and the *opacity* of the enrichment.

Meaning with respect to original triple The problems reviewed above all require enriching an existing triple with extra information, and the techniques and proposed additions surveyed above all facilitate adding information to (existing) triples. However, the added data can mean different things with respect to the original triple. If we assume that we want to add information x to an existing triple Rab , we can schematically distinguish four different meanings of the new information:

R^xab Adding information about the predicate of the triple;

Ra^xb Adding information about the subject or object of the triple;

$(Rab)^x$ Adding information about the whole triple; and

$Rabx$ Adding an extra argument to the triple on equal footing with the subject and object.

Transparency of the enrichment The reason for desiring declarative semantics is that we want our data to be interpretable by third party applications that do not (completely) share our data model: if our only concern is the internal use of our data we might as well use a custom RDF structure, such as a dummy node, for all enrichments, leaving the semantics implicit. In this respect, an important question to answer is: What do we want applications to do with the original triple if they do not

understand the enrichment: Assume it to be part of the RDF graph, or ignore it? In other words: do we want the enrichment to be transparent or opaque?

Transparent additions preserve the original meaning of the triple in the graph, meaning that applications that do not interpret the richer relation can still see the original relation; while

Opaque additions remove the original triple from the graph, meaning that it will not be visible to an application that does not (or cannot) interpret the enrichment technique.

Depending on the modelling requirements, we want to add certain information to a triple in a certain way. For example, a quoted source in a newspaper should be an opaque statement about the whole triple, while quality should be a transparent addition to the predicate. Thus, rather than looking for a single 'correct' solution, we think that multiple options are needed to express these differences in enrichment. Table 8.1 categorises the various proposals in these terms and serves as the basis for making the appropriate modelling choices. The proposal by Guha et al. (2004) is omitted from this table because its purpose is strictly describing graphs rather than enriching triples.

We will now reconsider the requirements listed above in terms of table 8.1. As listed in the previous sections, the basic unit of information is a triple representing a media relation (e.g. Bos dislikes Balkenende). To this triple we add information to *quantify* the predicate, add *extra arguments* to the relation, *specify the source of a quoted statement*, and link the media statement to *metadata* such as publisher and publishing date. As stated above, quoted sources should be opaque as the quoted statement is not directly asserted by the newspaper. The other additions should all be transparent: the original triple is a valid part of the graph with or without the extra information. The *quantification* is an enrichment of the predicate, but very difficult to represent using subproperties because of the quantitative and unrestricted nature of the information. The *extra arguments* and *quoted source* both add extra arguments that are subordinate to the main triple, falling somewhere between the intended meaning of reification (statements about triples) and n-ary relations (multiple arguments of equal weight). The *metadata* is adding information to the whole triple, and fits in the use case of reification and named graphs.

Surveying the table 8.1, there is no perfect method for adding information to triples. Named graphs have the desired transparency but offer no solution for distinguishing between extra arguments and metadata. Quadruples allow extra arguments in a natural way but this comes at the expense of flexibility and semantic clarity.

Table 8.1: Suitability of discussed mechanisms for expressing different triple enrichments

| | Transparent | | | | Opaque | | | |
|--------------|-------------------------------------|---------------------------------------|--------------------------------------|-----------------------------|-------------------------------------|---------------------------------------|--------------------------------------|-----------------------------|
| | Enriching an argument Ra^xb | Enriching the predicate R^xab | Enriching the triple $(Rab)^x$ | Extra argument $Rabx$ | Enriching an argument Ra^xb | Enriching the predicate R^xab | Enriching the triple $(Rab)^x$ | Extra argument $Rabx$ |
| RDF | | | | | \pm^1 | + | | |
| RDFS | + | \pm^1 | | | \pm^1 | + | | |
| N-ary | | | | | | | \pm^2 | + |
| Reification | | | | | | | + | \pm^2 |
| Quadruples | | | \pm^3 | \pm^3 | | | | |
| Named Graphs | | | + | | | | + | |

¹Adding a discrete categorisation is possible, but adding quantitative information is very difficult.

²N-ary patterns are explicitly intended to express an extra argument to a statement, while reification is intended to express information about a statement, making other uses of these solutions difficult to interpret.

³Since there is no specified interpretation of the extra argument, it is not possible to distinguish between these two cases.

Within the existing standards, reification covers adding metadata, and a case could be made for using reification to represent additional arguments. N-ary relations are better suited for the additional arguments but suffer from the lack of a standard vocabulary. It is possible to mix and match mechanisms, but this comes at the expense of increasing complexity, and if multiple non-standard mechanisms are mixed it will be difficult for third parties to understand what we mean.

8.2.2 *The Data Model*

Given the considerations above, we decided to stick to one representation for all enrichments. Since tool support for the proposed extensions is still limited, and the intended meaning of our enrichment is closer to meta-statements than to adding arguments, we decided to use RDFS reification. Figure 8.1 visualises the model resulting from the decisions described above. The main element of the data model, the original triple of the relational method, is now a reified `net:triple`, (a subclass of `rdf:Statement`). Triples have a subject, predicate, and object as required for reification, and also have the quantitative value ‘connection’ and an angle. On the left-hand side, triples are connected to textual units (sentences) from an article using the `dc:subject`, and metadata about this article and the coder are recorded.

8.3 Representing Political Background Knowledge

The previous sections showed how RDF can be used to represent the relations between objects extracted in Semantic Network Analysis. These relations connect the concrete objects directly mentioned in the text, such as politicians, parties, and specific issues. In order to use these relations in analyses and to combine networks from different countries or time periods, it is necessary to link these concrete objects to the more abstract concepts that are used in the communication research question and that are more stable across different domains. As described in chapter 4, *Ontology* is the study of the ‘things that are’, in other words, of the vocabulary used in the descriptions of the world. By describing the concrete objects such as actors and issues in terms of the more abstract concepts, the ontology provides a semantic context in which to interpret these objects. Different research questions will often require different contexts, categorising and interpreting the objects in a specific way. For example, sometimes we need to categorise politicians according to political function (parliamentarian, minister), while at other moments we need to use their party affiliation. On the other hand, data from different time periods or countries will often contain different actors even if the political

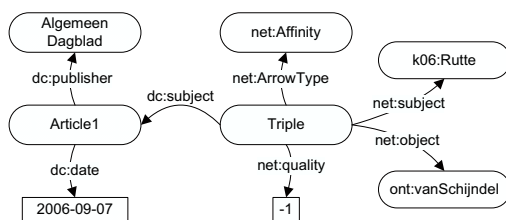


Figure 8.1: The data model used

roles remain constant. To make matters even more complex, the same actor might be present in multiple data sets from different periods, but because the actor changed roles over time, he or she needs to be categorised differently. Hence, in order to combine different data sets into a repository that is useful for answering a diverse range of research questions, it is necessary to create a formalisation that connects the concrete and abstract concepts in a flexible way, accounting for different possible categorisations and for the dynamic nature of political relations.

8.3.1 *Dynamic Political Roles*

There are a number of political roles and functions, such as being president, a member of parliament, or a member of a party, that are fulfilled by politicians for only a certain time period. The functions are social roles in the sense that they are anti-rigid — meaning that the existence of the politician does not depend on his playing a role — and dynamic (cf. Masolo et al., 2004). As surveyed by Steimann, treating such roles in frame-based models such as RDF, is not trivial, and has received considerable attention in the literature (Steimann, 2000; Sowa, 1988, 2000; Guarino, 1992). The simplest approach — treating a role as a predicate — does not allow for specifying temporal bounds or other information for the reasons discussed above. Another possibility is to make each occurrence of a role being played a subproperty of the role property. As argued by Steimann (2000), this leads to a number of complications and does not really solve the problem. It is also possible to reify the role relation, such as done by Mika and Gangemi (2004), putting the treatment of roles in line with the treatment of the NET statements as discussed above. However, the extra node created by this approach is on the meta-level rather than the object level. As will be shown below, this makes it more difficult to reason with the role instances. Gutierrez et al. (2005) propose an extension to the RDF model that can be expressed using a form of reification for reasoning with temporal validity. Since this requires an extension of

the RDF standard, this approach was not considered here.

The approach chosen here is to create an *adjunct instance* representing one occurrence of the role (Wong et al., 1997). For example, we would create a node `role-Bos-PvdA` representing the fact that Bos is a member of the PvdA party. This instance is linked to the role player (e.g. Bos), and the statements that define the role (*partyMemberOf* PvdA) are made with this role instance as subject. Subsequently, we can specify the *From* and *To* dates as properties of the role instance. This allows normal reasoning to occur on the role instance, leaving the inference system to determine whether a specific person is a member of a role for a certain coded triple (see below).

RDF Reasoning The advantage of the adjunct instance approach chosen here for temporal roles is that querying can be simplified by adding a custom reasoning step: if the subject or object of a coded triple is a role instance, and the publication date of the triple lies between the from and to date of the role, the role instance (also) becomes the subject (or object) of that triple. More formally, the antecedents and consequent of the rule can be given as follows:

```

1  IF    ?triple dc:date ?ADate
2  AND  ?triple ?rel ?object
3  AND  ?roleInstance amcat:roleSubject ?object
4  [AND ?roleInstance amcat:roleFrom ?FDate ]
5  [AND ?roleInstance amcat:roleTo ?TDate  ]
6  AND  (FDate is null OR FDate <= ADate)
7  AND  (TDate is null OR TDate >= ADate)
8  THEN ?triple ?rel ?roleInstance

```

This is illustrated in figure 8.2, where it is concluded that ‘Rutte as a VVD member’ is the subject of the triple from the publication date of the article and the *from* date of the role. This reasoning makes it possible to query for ‘a triple whose subject is a member of the VVD’, without having to process the roles or even knowing that the *memberOf* role is a temporal role.

This highlights the advantage of using an adjunct instance over using reification. Using reification, the ‘dummy’ node is the `rdf:Statement`

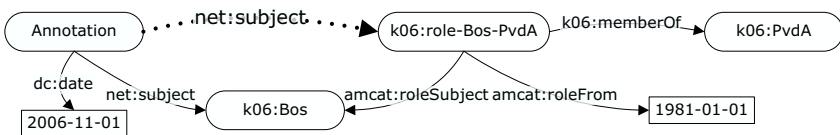


Figure 8.2: Inference of a Role played by a Politician

node. This node is on a different level of abstraction from the normal vocabulary (e.g. the politician and his party), and there is no *memberOf* relation between the dummy and the party. Hence, there is no easy way for retrieving (for example) all sentences about VVD-member without putting the temporal reasoning in the query or creating additional dummy nodes during the reasoning.

8.3.2 Representing Issues

Compared to the dynamic nature of the roles played by political actors, representing the issues is relatively simple since they are assumed to be static. The issue hierarchy plays an important role in Semantic Network Analysis by determining how concrete issues are related to the abstract overarching issues used in communication theory, and can be seen as a categorisation scheme or even as a high-level codebook, since it defines which topics to include in each category or code (cf. Krippendorff, 2004, p.132–135). Since different research questions will need different categorisations, we need to be able to create multiple parallel hierarchies. RDFS allows multiple inheritance, so this requirement is not problematic.

There is a problem with using an RDFS Subclass hierarchy for the issues, however: the issues higher up in the issue hierarchy need to be accessible for coding as well, as these more generic issues are often mentioned in texts directly. RDF allows using a class as a direct object in a relation, but doing so confuses the data and vocabulary definition. Since Description Logics require separating the A-Box (Assertions) and T-Box (Terminology), this means that the resulting graph is incompatible with OWL-DL. Although we do not currently use OWL, this is still a disadvantage as it limits our options and is semantically confusing.

Noy (2005) signals this problem and describes five approaches to overcome it. The first approach is by simply using the classes directly as described above, losing OWL-DL compatibility. The second approach is creating a distinguished instance of each class and using that instance, with the disadvantages of creating additional elements in the vocabulary (incurring storage and maintenance penalties) and not having a direct relation between the instance representing a subclass and the instance representing its superclass. The third approach involves creating a separate distinguished instance hierarchy, using a relation such as the *broader* relation from the SKOS thesaurus description vocabulary (Miles and Bechhofer, 2008). This instance hierarchy is then linked to the class hierarchy. This creates even more vocabulary elements, aggravating the maintenance and storage problems, while the links in the issue hierarchy still depend on non-RDFS vocabulary and reasoning: although the

RDFS subclass relation is defined to be transitive, the corresponding instance relation (e.g. *subissueof*) does not benefit from this and RDF does not have a standard way of declaring transitive relations. The fourth approach creates an anonymous restriction on the relation, stating that the object of the relation has to be some instance of the target class, without specifying which instance. This is equivalent to creating a blank RDF node as a class instance and using that as a subject. A downside of this approach is that it confounds the meaning of the relation: two relations about an issue will point to different blank nodes which may or may not refer to the same resource. Their last approach is to directly use the class as the object of a relation but declaring the relation to be an annotation property. Although the resulting graph will be within OWL-DL, these relations are ignored by DL reasoning and no further restrictions can be defined on the relations.

Given that these approaches to coding using a class hierarchy all have considerable disadvantages, we decided to create an instance hierarchy such as proposed in the third approach, but without creating the parallel class hierarchy. We defined the relation between the instances, *subIssueOf*, derived from the SKOS *broader* relation (Miles and Bechhofer, 2008). Using SKOS vocabulary has the advantage of specifying part of the semantics of our subissue hierarchy and allowing us to use SKOS tools, such as a thesaurus browser, to view the issue hierarchy. SKOS also allows multiple inheritance, so this requirement is still met.

8.4 A political ontology

The previous section outlined how to use RDF(S) to create such an ontology. This section will present the ontology created for political communication research; the ontology was created for Dutch politics but other political systems can probably also be accommodated.

8.4.1 Actors and roles

Political actors are an important part of the vocabulary in political discourse. As described in the previous section, political actors are characterised by their dynamic roles, such as party membership and ministry leadership. These roles are temporally bound and one actor may play multiple roles simultaneously.

Figure 8.3 shows the relevant part of the actor class hierarchy in the ontology. The grey open arrows show subclass (is-a) links. For example, *Person* is a subclass of *Actor*, meaning that every actor is-a person. The other subclasses of *Actor* are *Political Actor*, and *Institutional Actor*. *Politician*

is a subclass of both *Person* and *Political Actor*, meaning that a politician is both a person and a political actor. *Active Politician* is a subclass of *Politician*. Note that neither *Politician* nor *Active Politician* has explicit instances, since being a politician is not a permanent characteristic of a person, and RDF class membership is a static relation. In the upper right corner, *Party* and *Government Body* are both subclasses of *Institutional Actor* and *Political Actor*.

The black arrows ending in solid circles indicate *partOf* relations with the range and domain of the relation being the classes connected by the arrow. Each of these *partOf* relations is in fact a different subproperty of the *partOf* relation. This allows the domain and range of the relation to be specified in RDFS. For example, the relation between *Politician* and *Party* is *partyMemberOf*. Similarly, *Active Politician* has a *partOfBody* relation with *Government Body*.

The bottom of figure 8.3 shows how *Government Body* is further specified into the kind of body (*Legislative* and *Executive*) and the level of government (*Municipal* and *National*). Combining these gives the actual bodies of government. For example, *Cabinet* is a *National Executive* body, and *Parliament* is a *National Legislative* body, while *City Council* is *Municipal* and *Legislative*. Similarly, the *Municipal* can have an *Executive* body (in the Netherlands formed by the mayor and aldermen), and similar bodies exist at European and Provincial levels (not shown here to avoid redun-

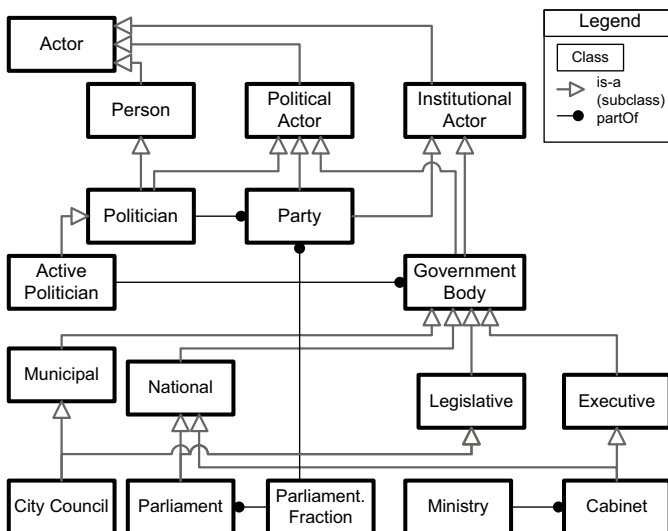


Figure 8.3: Actor class hierarchy

dancy). The last class in figure 8.3 is the Parliamentary Fraction, which is part of both a parliament and a party. Note that since an Active Politician can be part of a Government Body, he or she can be part of any of the subclasses of Government Body.

Figure 8.4 shows a concrete example of the network obtained when filling in these classes with instantiations. Rita Verdonk, shown in the bottom left corner, was a member of the VVD party and was elected to parliament and sworn in on the November 30, 2006. On September 14, 2008, she was expelled from the VVD Fraction after criticising the fraction leader Mark Rutte. In the graph, this is shown in the dates on the *partOf* relation from Verdonk to the VVD fraction. Interestingly, she chose not to give up her seat while remaining a member of the VVD. This situation lasted for a month until she left the VVD party on October 15 to start her own ‘political movement’ TON (Trots op Nederland; *Proud of the Netherlands*). The figure shows how she became the leader of the fraction ‘Member Verdonk’ in September, but only switched her VVD membership for the leadership of the TON party a month later.

The top part of figure 8.4 shows the parties and fractions. VVD and TON are both of type Party, and Member Verdonk and VVD Fraction are of type Parliamentary Fraction. Both fractions are part of their respective parties, and both are part of the Tweede Kamer, the Dutch Parliament.

An example of the dynamics of political functions is given in Figure 8.5. The bottom left corner contains the politicians: Maxime Verhagen and Jan-Peter Balkenende. The left-hand side shows their relations with the CDA party and fraction: Both politicians are members of the CDA, and Balkenende has been leading the CDA since October 1, 2001. Balkenende was elected to parliament in 1998 and left parliament to become Prime Minister in July 2002. Verhagen was a member of parliament from 1994

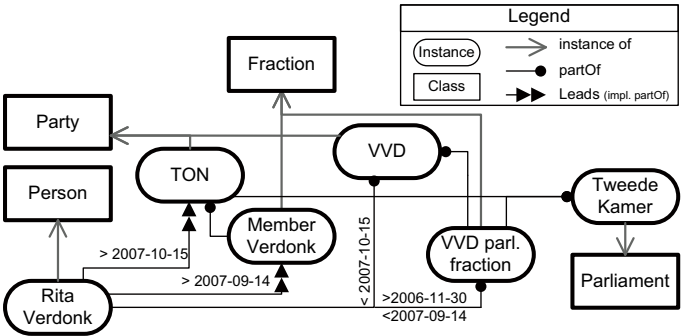


Figure 8.4: Dynamic roles and party membership

(not shown), and became leader of the CDA Fraction in July 2002. He remained fraction leader (with a short break after the 2003 elections) until the 2006 elections, when Balkenende returned to parliament and became fraction leader during the formation of the Cabinet Balkenende IV.

On the right-hand side, the relations with the Cabinet and Ministry instances is shown. Balkenende was leader of the four Cabinets Balkenende I–IV. In February 2007, Verhagen became Minister of Foreign Affairs in the Cabinet Balkenende IV, shown by the relations from Verhagen to those instances.

8.4.2 Issues

As described in section 8.3, issues are categorised in multiple issue hierarchies consisting of instances linked by *subIssueOf* relations. Table 8.2 gives an overview of two such categorisations: by *political direction* and by *departmental topic*. Political direction follows the categorisation of issues in the political spectrum, such as leftist versus rightist issues. The position of a politician on the different directions determines the place of that politician in the political spectrum. The topics loosely follow the departmental structure of government and traditional areas of expertise, such as finance versus law. Although there is some semantic overlap between topics and directions, they are two orthogonal categorisations. For each of these categorisations, the first level of issues and a selection of subissues are shown. All shown subissues exist in both categorisations, creating the mesh of subissues shown in the table. We describe a subset of this table below.

Administrative Reforms are reforms to the democracy or functioning of

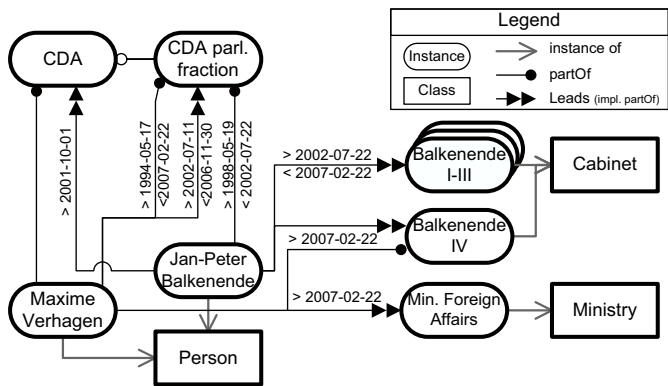


Figure 8.5: Dynamic roles and political functions

Table 8.2: Issues categorised according to direction and topic

| Direction | Subissue | Topic |
|------------------------|--------------------------|--------------------------------|
| Administrative Reforms | Elected Mayor | Administrative Reforms |
| Education | Education | Education & Science |
| Environment | Public Transport | Transportation |
| | Environment | Spatial Planning & Environment |
| Infrastructure | New Roads, Rail | Spatial Planning & Environment |
| | Residential Construction | Spatial Planning & Environment |
| Immigration | Immigration | Immigration |
| Crime | Safety | Public Order & Safety |
| | Crime | Justice |
| | Rule of Law | Justice |
| Valence | Economic Growth | Economic Affairs |
| | Economic Liberalisation | Economic Affairs |
| | Health | Health Care |
| Rightist | Budget, taxes | Finance |
| | Health Care | Health Care |
| | Agriculture | Agriculture |
| Health Care | Freedom of Speech | Norms & Values |
| | Food Safety | Norms & Values |
| Traditional Values | Abortion | Norms & Values |
| | Army, Defense Spending | Defence |
| New Leftist | Gay Marriage | Defence |
| | Peacekeeping Missions | Defence |
| Leftist | Welfare | Social Affairs |
| | Development Aid | Foreign Affairs |
| Europe | European Integration | Foreign Affairs |

government such as a directly Elected Mayor or Prime Minister and transparency of government. These issues have both a directly matching direction and topic, so they form a simple row in the table.

The (pro) Environment direction is slightly more complex: Environment itself is categorised under Spatial planning and Environment, but Public Transport, which also belongs to the direction Environment, is categorised departmentally as Transportation. The reasoning behind this is that a politician in favour of public transport is in favour of the environment (direction), yet the responsible minister will be that of transportation and a relevant expert will know about logistics and transportation rather than ecology.

Valence issues are issues that politicians will all agree on in general:

everybody wants economic growth and health and nobody wants inflation or unemployment. These issues are scattered across the departments: Economic Growth belongs to Economic Affairs, Health to Health Care, Rule of Law to the Justice Department, and Freedom of Speech to Norms and Values (as a constitutional right).

Rightist issues are those issues that conservative politicians are generally in favour of. A number of them, such as the government Budget and lower taxes, are part of the department of Finance. Economic Liberalisation, such as privatisation and deregularisation belong to Economic Affairs. Two other rightist issues are Agriculture, belonging to the department with the same name, and the Army, which belongs to the department of Defence.

New Leftist issues represent post-materialistic idealism. Where traditional leftist issues such as welfare and labour conditions are concerned with materialistic concerns, new leftist issues are about ethical or international concerns. New leftist issues can be found in different topics. For example, Gay Marriage is part of the topic Norms and Values, Peacekeeping Missions are a Defence topic and Development Aid belongs to Foreign Affairs.

The issues described here are further elaborated into subissues. For example, Taxes contains mortgage interest deduction and profit tax, and Peacekeeping Missions contains the Dutch missions in Uruzgan and Iraq. The advantage of using such a detailed categorisation scheme lies in postponing coding decisions from the extraction to the querying phase, and hence from the coder to the analyst. For example, suppose a text to be analysed using political directions mentions the military mission in Uruzgan. If this was directly coded as one of the main directions, the coder would have to decide whether the main focus is on peacekeeping or on fighting, and hence whether it is new leftist or rightist. By directly including Uruzgan as an issue in the issue list, the coder does not have to make this decision, and the categorisation can be inspected and changed afterwards. Moreover, different categorisations can be used depending on the research question, allowing the data to be used more flexibly.

Both the actors and issues described above are developed for and geared towards the Dutch political situation. However, most systems with a party democracy should fit in the actor hierarchy without too much adaptation, and the issue categorisation is applicable to different situations as well. Moreover, the advantage of formalising the concept hierarchy like this is that it allows for specification and adaptation of concepts using subclasses. This makes it possible for details of the hierarchy to be changed for a new country or situation while the data sets can still be compared and combined on a level of abstraction they agree on.

8.5 Using OWL for a richer ontology

The preceding sections described how the media data and background knowledge needed for Semantic Network Analysis could be represented using RDF(S). As such, there is no need for using a more expressive ontology language such as OWL. However, there are a number of ways in which a more expressive language might be beneficial.²

Ontology Checking One of the features of RDF is that it does not allow negation or disjunction, and hence cannot lead to inconsistencies. This means that RDF(S) cannot be used to detect ‘errors’ in the ontology: Suppose *partyMemberOf* has domain *Politician* and range *Party*, and we accidentally assert that the Party CDA is a member of the Politician Balkenende rather than the other way around. An RDFS inferencer will happily conclude that Balkenende is a party as well as a politician and *vice versa* for the CDA. OWL allows for the formulation of constraints to prevent this type of mistake. For example, the mistake described above can be prevented by defining the classes *Party* and *Politician* to be *disjoint*. *Cardinality Constraints* could also be used to enforce constraints, by for example demanding that there be 150 members of parliament and that a politician is a member of exactly one party. Since these roles are dynamic, however, this requires concrete domain reasoning about the role dates and is not as simple as formulating a constraint. More complex restrictions can also be made, for example stating that a person in the Netherlands cannot be minister and parliamentarian at the same time, or that a person can only be a member of the parliamentarian fraction of his or her party. Getting these constraints right might be difficult, however, as political reality can be complicated: Verdonk remained a member of the VVD party for a month after starting her own fraction, and after an election the demissionary ministers can sit in parliament for up to 100 days. These idiosyncrasies are problematic because an error signalled by an OWL inferencer cannot be simply ignored or treated as an exception: an inconsistency affects the whole ontology and has to be remedied.

Disjoint Categorisations As described above, in RDF it is possible to elegantly add background knowledge and use this knowledge to link ‘data level concepts’ to ‘theory level concepts’. A common use case in Content Analysis is to define a set of categories on the media data, for example statements with an opposition politician as subject, with a coalition politician as subject, and statements with a societal actor as subject. Counts of such statements per period are then used either in statistical

²See section 4.2.3 on page 59

analysis or presented in a table. Both uses require the categories to be disjoint and exhaustive with respect to the higher category, in this case ‘actor statements’. In other words, the higher category should be *partitioned* by the proposed categories. In RDF, this can be checked for real data, for example by querying for instances of two categories. In some cases, such as presenting data real-time on a web page, we would like to be able to *prove* that such a categorisation will always be a partitioning. In OWL, this can be achieved by proving that each pairwise conjunction of the categories is unsatisfiable. Exhaustiveness can be shown by proving that the higher category implies membership of one of the lower categories. More formally, proving that the categories $\{A_1 \dots A_n\}$ partition B in the ontology \mathcal{O} means proving $\mathcal{O} \models A_i \sqcap A_j = \perp$ for all $i \neq j$, $i, j \leq n$ and $\mathcal{O} \models B \sqsubseteq A_1 \sqcup \dots \sqcup A_n$.

Defining Concepts The background knowledge described in this section is used to bridge part of the ‘semantic gap’ between the concrete objects in text and the abstract concepts used in communication theory. As argued in chapter 2, there is also a complexity gap between the atomic theoretical variables and their realisation in complex textual structures: theoretical variables should often be operationalised as patterns in the Semantic Network rather than as single relations. Chapter 9 will show how RDF queries can be used to identify such patterns procedurally, but if we can use OWL to define such patterns declaratively it has the advantages of more explicit semantics, and a standard inferencer can be used to add the identified patterns to the Semantic Network. Whether this is possible depends on the complexity of the pattern to be detected. For example, the sentence types identified in section 2.4 are easy to formulate as OWL queries: a conflict sentence is any affirmative relation between two actors. A more complex pattern, such as a contested issue, can also be described as any issue which has two opposite incoming relations from politicians. A general restriction on such patterns is that the Description Logic, the logic behind OWL, does not allow variable binding. In practice, this means that patterns may only be star-shaped: it is impossible to create diamond patterns or other cycles since this requires the ability to constrain two nodes to be identical, which requires a form of variable binding. Van Atteveldt and Schlobach (2005) describe how the *Hybrid Logic* $H(@, \downarrow)$, an extension of description logic with variable binding, can be used for defining patterns in Semantic Networks. Unfortunately, reasoning in description logic extended in this way is no longer decidable, forcing the actual identification to be done using model checking (Areces et al., 1999; Franceschet and de Rijke, 2005). Since hybrid logic model checking is similar to RDF querying, this means that the usefulness of OWL definitions over RDF queries are limited to cases in which

the pattern can be defined in Description Logics, i.e. those cases where the pattern does not contain cycles.

8.6 Conclusions

This chapter showed how RDF can be used to formally represent Semantic Network data, both the concrete networks extracted from text and the background knowledge needed to link the concrete networks to the generic concepts used in analysis. In particular, we discussed three problems:

Statements about Statements Since Semantic Network data itself consists of triples, we need to use RDF triples to describe the network triples. We identified a number of different ways in which triples can be enriched with new data, distinguishing between opaque and transparent enrichment and investigating the relation between the new information and the existing triple. After surveying a number of proposals to solve the statements-about-statements problem, we decided to stay within the RDFS standard by using reification.

Dynamic Roles Politics is characterised by actors that frequently shift roles and allegiances: parliamentarians move into cabinet and afterwards become mayor, dissenters are expelled from their parties and start new parties, etc. For an ontology to be useful across data sets, these dynamics need to be captured in the background knowledge. We propose using adjunct instances to represent these roles. By using RDF reasoning, the role memberships can be resolved at the moment the data are entered into the repository, allowing for easy querying.

Multiple Instance Inheritance Since different research questions require different categorisations of the issues, the issue hierarchy is characterised by multiple inheritance. Moreover, the more abstract issues are also used in the Semantic Networks, making it difficult to use a normal class hierarchy for the issues. We showed how an instance hierarchy using standard SKOS vocabulary can be used to describe the issues.

Additionally, we presented the actual ontology that was used in the 2006 election study that was described in chapter 2 and used in chapters 6, 7 and 9. By elaborating on a number of actors playing different roles, we showed how the dynamic roles function in practice. Moreover, we showcase the multiple issue inheritance by looking at a selection of issues through two existing categorisation systems. Finally, we briefly discussed a number of ways in which the richer Web Ontology Language (OWL) can be used to test and enrich the political ontology.

CHAPTER 9

Querying, Analysing, and Visualising Semantic Network Data

'All this television time for this hype — why?'
(Al die zendtijd voor deze hype — waarom? *nrc-next*, 4 February 2008)

News frames and other concepts relevant to communication science can often be identified as patterns on Semantic Networks. This chapter presents a high-level query language in which such patterns can be defined, and a web application for querying, analyzing, and visualizing Semantic Network repositories using these patterns.

This chapter is based on:

Wouter van Atteveldt, Nel Ruigrok, Stefan Schlobach, and Frank van Harmelen (2008), *Searching the news: Using a rich ontology with time-bound roles to search through annotated newspaper archives*, to appear in the proceedings of the 58th annual conference of the International Communication Association, May 22–26, Montreal.

9.1 Introduction

As described in chapter 2, many of the concepts that are used in communication theory are complex. For example, the associations between issues and attributes in Second Level Agenda Setting (section 2.2.1) or frames such as the responsibility frame used by Valkenburg et al. (1999), which consists of a government actor being held responsible for a certain problem. Section 2.3 argued that there are two gaps between these concepts and the source text that Content Analysis has to bridge: an *abstraction gap* between the concrete objects in text and the abstract concepts in communication theory; and a *complexity gap* between the atomic variables in communication models and the complex textual structure they represent. In *Semantic Network Analysis*, a textual network is extracted containing relations between concrete objects. Formalised background knowledge, such as the political ontology presented in the previous chapter, bridges the abstraction gap by linking the concrete to the abstract. The *complexity gap* can be bridged by defining theoretical variables as patterns on the Semantic Network.

This chapter shows how such patterns can be defined in a high-level query language and automatically identified in Semantic Networks. A web application is presented that enables users that are not experts in knowledge representation to query the repository using these patterns and inspect and visualise the results. This makes it possible to define relevant communication science concepts as formal queries, and automatically search for those concepts in Semantic Network data. The requirements for this application are as follows:

Search for patterns in the extracted networks Many relevant Content Analysis questions can be reduced to identifying patterns between actors and issues.

Query on an abstract level The network extracted from the text is a network of concrete actors and issues such as Bush or CO₂. The research question a social scientist wants to answer is formulated in more abstract terms such as President or Environment. By querying the abstract level rather than the concrete objects, the same queries can be applied to data from different countries or time periods.

View results on an aggregate level and zoom in on details For quantitative analysis, the Social Science user wants the data on an aggregate level, for example total frequency of a pattern per newspaper per month. In order to make sense of the results and to check and refine these patterns, the user must be able to easily 'go back to the text' to see the underlying detailed data.

In order to query the resulting graph, we define a simple high-level query language which is translated into an RDF query language (section 9.2). A simple web interface allows a user to input the query, view the aggregate results, and zoom in on interesting articles (section 9.3). The system is tested by letting a Social Scientist use the system to reproduce a number of analyses from the 2006 election study described in section 2.5 (section 9.4).

9.2 Querying the Semantic Network

The previous chapter presented an RDF data model for representing Semantic Network Data. This model uses RDF reification to represent ‘high-level’ Semantic Network triples using ‘low-level’ RDF triples. If we want to define patterns on the Semantic Network, we want to do so in terms of the high-level Semantic Network rather than the low-level RDF triples. This makes the task of defining the patterns easier, and makes sure that the patterns express the intended meaning rather than artifacts of the chosen RDF representation, allowing us to change the representation without having to change the pattern definitions (for example, if context markers become a standard RDF feature).

This section presents a high-level query language that can be used to define patterns and can be automatically translated into standard RDF queries. Patterns in this language consist of triples, which can be either *network triples* or *background triples*. A network triple consists of two concepts or variables that are related in the Semantic Network. This relationship can be constrained to be positive (+) or negative (-), or can be unconstrained (_). A background triple links a variable to a concept or another variable using one of the relations in the ontology, such as is-a (`rdf:type`) or one of the *partOf* relations. For example, suppose we want to search for *internal criticism* within parties. This is formulated as follows:

```
1 ?ActorA - ?ActorB
2 ?ActorA ont:memberOf ?P
3 ?ActorB ont:memberOf ?P
4 ?P isa ont:party
```

The first line specifies a negative relation between two variables `?ActorA` and `?ActorB`; the following two lines require both variables to have a `ont:memberOf` relation with a third variable, and the last line specifies the node represented by this third variable to be of type `ont:party`. This query language defines relations between nodes, where the relations can be either background relations or (positive or negative) media relations, and the nodes can be resources (actors or issues) or variables.

These queries are automatically translated into SeRQL¹, yielding the following SeRQL for the example query:

```

1 SELECT DISTINCT ActorA, ActorB, P, Qua_1
2 FROM {} net:subject {ActorA_1};
3     net:object {ActorB_1};
4     net:quality {Qua_1};
5     dc:subject {} dc:identifier {ArticleID},
6 {ActorA_1} ont:memberOf {VAR_P_1};
7     net:roleSubject {ActorA} rdf:type {ont:Root},
8 {ActorB_1} ont:memberOf {VAR_P_1};
9     net:roleSubject {ActorB} rdf:type {ont:Root},
10 {VAR_P_1} rdf:type {ont:party};
11     net:roleSubject {P} rdf:type {ont:Root}
12 WHERE (Qua_1 < "0"^^xsd:float)

```

The media relation `?ActorA - ?ActorB` is translated into subject, quality, and object relations from the annotation node (lines 2–5). For the party membership of Actor A, this creates the `ont:memberOf` relation on line 6. Line 7 requires that the `ActorA_1` variable is a role played by ActorA, which derives from `ont:Root` to make sure we select the actual person rather than an adjunct instance. `net:roleSubject` is defined to be reflexive to allow normal relations to be dealt with using the same query. Lines 8–11 define similar restrictions for the other two variables, `?ActorB` and `?Party`. The last line makes sure that the relation is negative. If multiple media relations use the same variables in the simplified syntax, they would be translated into different variables (ie `ActorA_1` and `ActorA_2`) to allow different adjunct instances with the same role subject to be used.

9.3 Searching the News

The query language presented in the previous section was designed to make it easy to search for patterns in Semantic Network repositories. To test whether this query language actually helps in conducting analyses for communication research, a prototype web application was developed. In this application, queries can be performed on the Semantic Network Analysis of the 2006 parliamentary elections as described in section 2.5.

Figure 9.1 shows the query screen of the web application. In the top field, the user fills in a query in the language presented in the previous section, in this case the ‘internal criticism’ example described earlier. In

¹ At the time this research was done, the SPARQL query language (Prud’hommeaux and Seaborne, 2006) was not yet an official W3C recommendation; translating into SPARQL would be a trivial modification as SPARQL and SeRQL are quite similar.

| Evaluate Query | |
|---|---|
| Query | <div>?ActorA - ?ActorB ?ActorA ont:memberOf ?P ?ActorB ont:memberOf ?P ?P isa ont:party</div> |
| Instantiate | <input type="text"/> |
| Show results per | Month <input type="button" value="v"/> |
| Limit results to | Year/Month <input type="button" value="v"/> <input type="text"/> |
| Show | Frequency and Quality <input type="button" value="v"/> |
| <input type="button" value="Submit Query"/> | |









Figure 9.1: Querying Internal Conflict with the Web Application

the *instantiate* field, he or she can select the variables for which the instantiations are shown in aggregate results. The next two input fields allow the user to view per month, week, or day, and to limit the results to a specified time period. In the final box the user can select whether to display (average) quality and whether to show results as a table or as a graph.



Figure 9.2(a) shows the aggregate results per month without instantiations. In the example, the internal conflict frame occurred very often (155 and 161 times) in the first two months of the study due to a number of internal problems in the PvdA, CDA, and VVD. During the next two months this dropped to 64 and 94 times.

Often, the user will want to know more detail than just the frequency: in which parties did the conflict occur? This is shown by filling in the relevant variable, *?P*, in the *instantiate* field. Figure 9.2(b) shows this for the example query for the month of August: VVD had most internal conflicts (64), with all conflicts being close to maximally negative (-1.0).

By clicking on the icon on the left-hand side, the user can get a list of articles in which the pattern occurs. For each article, the relevant meta information is given, as well as the frequency and quality of the pattern. Inspecting the relevant headlines often gives a good idea of what is happening in a time period, and is very useful for understanding frequency peaks. Subsequently, the user can click on the magnifying glass in the article list. This shows the text of the article itself and a network visualisation of the annotation of the article. Figure 9.2(c) shows the text and network of the article from the earlier example. Rutte and Van Schijndel are both members of the VVD party; see section 2.4 for an explanation of the Ideal and Reality nodes. In the network, the edges that matched the

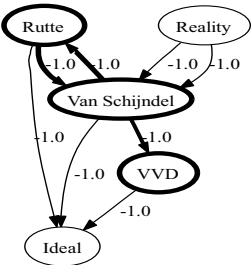
| | | Year/Month | QUA_1 | Frequency |
|---|---|------------|-------|-----------|
|  |  | 2006/08 | -0.99 | 155 |
|  |  | 2006/09 | -0.97 | 161 |
|  |  | 2006/10 | -0.98 | 64 |
|  |  | 2006/11 | -0.96 | 94 |

(a) Aggregate Results

| | | Year/Month | ?Party | QUA_1 | Frequency |
|---|---|------------|--------|-------|-----------|
|  |  | 2006/09 | SP | -1.00 | 1 |
| | | | PvdA | -0.95 | 37 |
| | | | CDA | -0.96 | 51 |
| | | | VVD | -0.99 | 64 |

(b) Instantiated Results for September 2006

| Article | |
|--|--|
| Source | Algemeen Dagblad |
| Date | 2006-09-07 |
| Headline | Rutte zet Kamerlid Van Schijndel uit VVD-fractie |
| DEN HAAG VVD-leider Mark Rutte heeft Kamerlid Anton van Schijndel gisteren uit de fractie gezet. Van Schijn- | |



(c) Zooming in on a single article[†]

Figure 9.2: Analysing Internal Conflict: Results

[†]Rutte expels parliamentary Van Schijndel from VVD Fraction
THE HAGUE VVD leader Mark Rutte expelled parliamentary Anton van Schijndel from the fraction yesterday. Van Schijn-...

pattern are automatically highlighted in this visualisation. This screen is very useful for gaining a close-up view of what is happening and for refining the patterns by making sure the matched edges correspond to what the researcher was looking for.

Finally, the user can also visualise the whole network of instantiated patterns per time period, by either clicking the network icon next to a time period or by filling in the ‘visualise’ form field at the top. An example of this is given in figure 9.3 in the next section.

9.4 Using the system: Parties in the news

To test whether the system gives meaningful results for actual queries and that it is usable by non-expert users, the second author of the paper this chapter is based on (a communication scientist with no background in Artificial Intelligence) was asked to use the system to replicate two of the analyses from the original study (Kleinnijenhuis et al., 2007a). Although this is a very informal estimate of the usability, it can indicate whether the system could be used by users from outside Artificial Intelligence.

9.4.1 Use Case 1: Issue ownership

Issue ownership theory states that a party benefits from news about *owned issues*, the issues for which that party is generally seen as the best actor to solve problems (Petrocik, 1996). For example, social democratic parties are often seen as owners of social security while right wing parties are seen to own security issues. A hypothesis is that parties will relatively often make statements about their own issues, in order to increase the visibility of those issues. This hypothesis was tested with the following query:

```

1  ?X _ ?I
2  ?X ont:memberOf ?Party
3  ?Party isa ont:party
4  ?I ont:subIssueOf ?Issue
5  ?Issue isa ont:Issue

```

The frequency of each ?Party – ?Issue instantiation was downloaded as CSV and combined with pure party visibility (obtained using a simple query [`?X = ?Y, ?X partof ?P, ?P isa Party`]) to create table 9.1. This table lists the attention as a column percentage for a selection of parties: the total visibility (the second column) and the visibility of statements by that party on four issue groups (columns 3–6).

The table confirms the hypothesis: the visibility of statements of parties on their own issues are overrepresented compared to their total visibility. The SP (socialist) and especially the PvdA (social democrats) are more visible with Social Security issues than their average visibility.² The CDA (christian democrats) are more visible with norms and values and, together with the VVD, with financial issues. The ‘shared’ issue ownerships show how these parties were fighting each other over control of the issue.

Table 9.1: Visibility of Parties and Issue Statements (column percentages)

| Party | Overall | Issue statements | | | |
|-------|---------|------------------|----------------|-----------|-------------|
| | | Social Security | Norms & Values | Financial | Immigration |
| SP | 5.2 | 5.9 | 6.8 | 3.7 | 0.6 |
| PvdA | 21.8 | 32.5 | 14.1 | 19.0 | 16.6 |
| CDA | 39.9 | 33.1 | 50.5 | 44.0 | 32.1 |
| VVD | 31.1 | 28.0 | 23.3 | 32.7 | 42.8 |

9.4.2 Use Case 2: Parties and conflict

Conflict news is an important factor in the success or failure of parties during elections. Conflict can work in two ways, as an opportunity and as a threat: criticism of opponents provides the party and the members an opportunity to create a distinct profile, but internal criticism can be fatal for a party (Kleinnijenhuis et al., 2007a). The following query shows us the internal and external conflict among the different parties:

```
1 ?X _ ?Y
2 ?X ont:partof ?A
3 ?Y ont:partof ?B
4 ?A isa ont:party
5 ?B isa ont:party
```

The results of this query were visualised for the whole period, which yielded the network in figure 9.3. The dashed lines indicate positive relations and the solid lines negative ones; thicker lines indicate a higher frequency. As can be seen from the figure, the news was dominated by the negative relations between PvdA and CDA, the main contenders for the position of Prime Minister. It is interesting to note that the positive

²See section 2.5 on page 33 for a description of the 2006 elections and the main political parties.

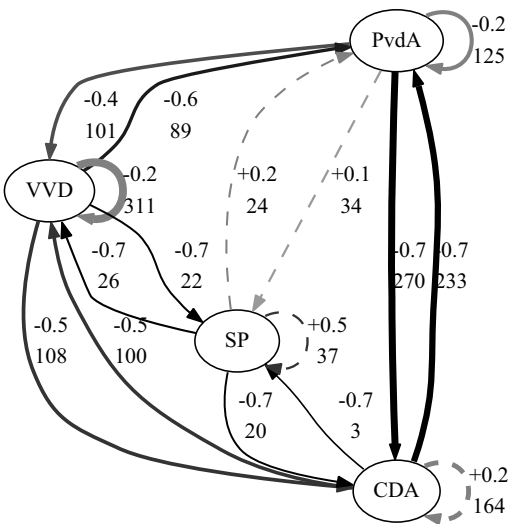


Figure 9.3: Conflict between parties

mutual relation between the two competitors on the left wing, PvdA and SP. In the PvdA and VVD there was a preponderance of internal conflict, which is seen as contributing to their electoral loss.

9.4.3 Use Case 3: Internal conflict

Internal conflict can be fatal for political parties. A party with closed ranks is regarded more stable and more able to manage the country than a party in which party members are fighting for power. In this section we will look at the internal praise and conflict in the parties. We used the following queries:

- | | |
|--------------------|--------------------|
| 1 ?X - ?Y | 1 ?X + ?Y |
| 2 ?X partof ?A | 2 ?X partof ?A |
| 3 ?Y partof ?A | 3 ?Y partof ?A |
| 4 ?A isa ont:party | 4 ?A isa ont:party |

The results of this query, imported into Excel to create a graph, are shown in figure 9.4. The extensive news coverage about the internal conflict within the VVD is clearly visible. Although the party members tried to close ranks and show public support for each other, the attention for this kind of news lags far behind the coverage focusing on the internal fights. The same picture, in somewhat milder form, is seen for the PvdA, where internal criticism of the leadership exceeds the internal praise in

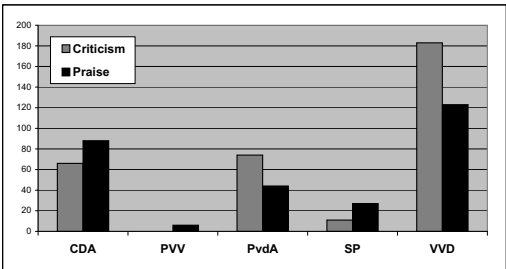


Figure 9.4: Internal criticism and praise for the major parties

the news. The opposite tendency is seen for the SP and the right-wing PVV.

9.4.4 Use Case 4: Dynamic Roles and Criticism

If we investigate the source of the criticism of the CDA, we encounter the interesting case of Minister of Justice Donner: After a critical report about a fire in a prison facility at Amsterdam Airport Schiphol, he stepped down as minister, and he was replaced by Hirsch-Ballin. We want to investigate the praise and criticism directed at the Minister of Justice. To check that the system correctly identified the fact that Hirsch-Ballin replaced Donner as minister, we create three queries: one for the role of minister, and one each for the politicians playing the role:

```
1 ?X _ ont:Donner      1 ?X _ ont:HBallin      1 ?X _ ?Y
2 ?X isa ont:PActor    2 ?X isa ont:PActor      2 ?X isa ont:PActor
                                     3 ?Y ont:leads ont:MoJ
```

The results of each query were exported and combined into the bar graph in figure 9.5, showing the total frequency times quality (valence) of the found statements. The grey bar is Donner, the black bar Hirsch-Ballin, and the white bar the Minister. As can be seen in the chart, the white and grey bars are identical until week 37. From week 39, the values for Minister and Hirsch-Ballin are identical. In week 38, the week of the change, the bar for Minister is lower than both other bars: apparently, both politicians were criticised as minister, but praised after stepping down (Donner) and before being sworn in (Hirsch-Ballin), which is confirmed by a per-day analysis (not shown). This shows that the system correctly handles changing political functions.

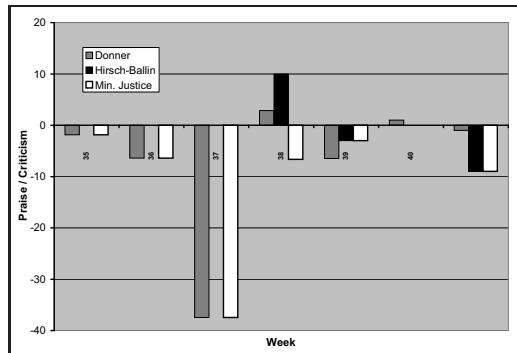


Figure 9.5: Criticism of consecutive Ministers of Justice: Donner and Hirsh-Ballin

9.4.5 Validation Results

It is difficult to provide a quantitative evaluation of a system as presented here, but the above use cases provide a qualitative evaluation by a member of the target audience. Overall, the user found the system useful and relatively easy to work with. Especially appreciated was the ability to 'dig down' from the aggregate views into the instantiations and all the way to the underlying articles, which really helps to understand the data and refine the patterns. The main drawback was the fact that queries had to be input manually, requiring some training and a thorough knowledge of the underlying ontology. A future version of the application will provide a graphical user interface for creating patterns, by dragging and dropping items from the displayed ontology and specifying the required relations.

9.5 Conclusion

The variables used in communication research are generally abstract (using general categories or political functions rather than concrete actors or issues) and complex (representing relations between multiple concepts). This chapter presented a system in which such variables can be defined as patterns on the Semantic Network of media data and background knowledge.

Building on the RDF representation of the Semantic Network as presented in the previous chapter, we defined a high-level query language that is automatically translated into RDF queries. A web interface allows

users to specify the patterns to identify in the news, and view the results on an aggregate level, and zoom in to view the individual articles, which is important in order to get a feeling for the data and to refine the patterns. A network-based visualisation gives the user further insight into the way the articles are coded and which part of the network is matched by the pattern.

In our perspective, the main limitation of this work lies in the complexity of the representation. Even with our simplified query language, formulating queries is not trivial and the user needs a good understanding of the ontology to define meaningful queries. Although this is partly due to the expressivity of the query language and the complex range of queries that can be defined using it, the complexity can be alleviated by offering a graphical query interface and allowing users to drag and drop concepts from the ontology. Also, we can create a 'menu' of frequent query patterns to facilitate getting started with the language.

Together, this and the previous chapter demonstrate how RDF technology can be used to solve relatively complex representational problems in an elegant way, and how an application can be built using this representation that offers non-expert users a way to query this data and browse through the archive. This provides the tools and techniques for creating large-scale Semantic Network repositories, making it easier to share and combine data from different domains, time periods, and research groups. Moreover, it was shown how applications such as the web application presented in this chapter can be created to make these repositories accessible, allowing communication scientists and other interested users to query for patterns in the repositories and analyse and visualise the results.

Part IV

System Description

CHAPTER 10

The AmCAT Infrastructure

AmCAT — the Amsterdam Content Analysis Toolkit — is the name for a set of tools that have been implemented as part of the research described in the previous chapters. This system was developed for storing, analysing, and querying newspaper articles and other media messages. This chapter gives a description of this infrastructure from the functional and technical perspective.

10.1 Introduction

To place the main components of the AmCAT system, the AmCAT navigator and iNet, in perspective, consider the simple workflow illustrated in figure 10.1. A research project usually starts in the **AmCAT Navigator**, in which the researcher can create the project, upload and view documents such as newspaper articles, and conduct exploratory word-count analyses. From there, the researcher can assign all or some articles for manual annotation. These articles will then be annotated using **iNet**, manually extracting the Semantic Network using a specified ontology of actors and issues. Alternatively, the researcher can apply the **automatic annotation** techniques described in part II to extract the Network automatically. In either case, the result is a Semantic Network containing the relations extracted from the documents. This combined Network can then be used for **Semantic Querying** using the techniques described in part III, identifying the patterns that are indicators of the theoretical constructs needed to answer the research question.

The Automatic Annotation and Semantic Querying are described in parts II and III, respectively. This chapter will describe the AmCAT Navigator in section 10.2, and iNet in section 10.3. Both these sections are divided into three parts: the first part focuses on the functional description and requirements of the component; the second part gives an example use case; and the third part gives a more technical description, focusing on architecture details and implementation.

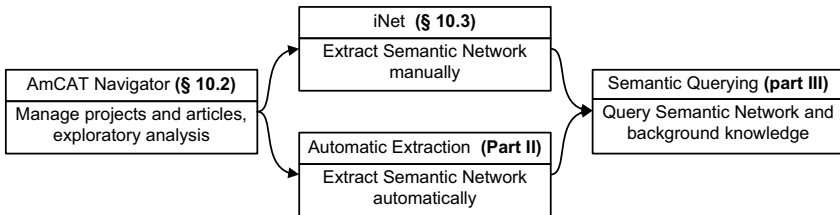


Figure 10.1: Simple Workflow using AmCAT

The AmCAT system described in this chapter is different to the kind of techniques presented in the previous parts. Parts II and III presented experimental systems using state-of-the-art techniques from Natural Language Processing and Knowledge Representation, respectively, which were evaluated for performance and usefulness for answering social science questions. In contrast, AmCAT is a collection of mature programs designed for making content analysis easy and efficient. As the systems presented in the earlier chapters mature, they will also be included in

AmCAT in order to make Semantic Network Analysis as accessible and productive as possible.

The author of this thesis designed and implemented the AmCAT database and iNet annotation program. Jouke Jacobi implemented the majority of the AmCAT navigator.

10.2 The AmCAT Navigator and Database

The heart of the AmCAT system is a database containing the documents, codings, and other data. To manipulate these data, AmCAT contains a number of ‘scripts’ or small programs to create projects, conduct frequency analyses, and perform Natural Language (pre-)Processing. In order to make the data and these scripts more accessible, an Intranet website was created: the AmCAT Navigator.

10.2.1 *Functional Design*

The AmCAT system was designed to make it easier to manage (large) content analysis projects. Documents being one of the essential ingredients, an essential aspect of AmCAT is the storing, viewing, and managing of documents from various sources. AmCAT can import multiple file formats, such as XML and Lexis Nexis text files, and automatically extracts the metadata such as publishing date and source. In the Navigator, it is possible to view the imported documents and browse by project, source, and date.

To enable simple frequency analyses in an efficient manner, AmCAT includes an index of the imported documents, and allows analysts to search for keywords and combinations of keywords, and display the search results graphically and textually. In all cases, it is possible to go from the results back to the original documents; this is required to check the results and make sense of them qualitatively. Also, the results are usable as a document sample for more thorough analysis.

A function of the AmCAT system is to serve as a launching point for other analyses, such as manual coding or linguistic analysis. Since not all of these analyses are known beforehand, the system is flexible in allowing new analyses to be run on selections of documents.

Since AmCAT serves as the infrastructure for Semantic Network Analysis, it is possible to view and query the extracted Semantic Networks using the background ontology, regardless of whether the Semantic Network was extracted manually, automatically, or obtained from another source.

10.2.2 Using the AmCAT Navigator

In order to demonstrate the usage of the AmCAT system, let us walk through an example using the Navigator to start a project and conduct a frequency analysis and assign some articles for manual coding. In this example, assume that we are investigating the framing of Islam before and after the 9/11 attacks, using a subset of the corpus described in chapter 5. Figure 10.2 shows the steps we shall follow in this example.

- (1) **Create Project** To start the investigation, we create a new project called 'Terror.' This takes us to the Project Details page, which shows that the project is empty. On this page, we select *Add Documents*, and upload our articles in a zip file. These articles are added to the database, and the metadata such as publishing date and source are automatically extracted. Currently, the system supports plain text, an XML format, and Lexis Nexis raw text files. Figure 10.3 shows the Project Details screen with the articles loaded. Notice that the system created a number of *batches*. A batch is a group of articles retrieved using the same criteria, for example a search query. This query is stored together with the batch to ensure that it remains clear what the selection criteria of the project were. These batches can be created manually to perform as a form of subdirectories within a project, and the system will generate them automatically if the search query is included in the input files.
- (2) **Select Articles** After creating the project, we use the article selection screen to select all or some of the articles. We can then tabulate the article frequencies per time interval per medium to make sure

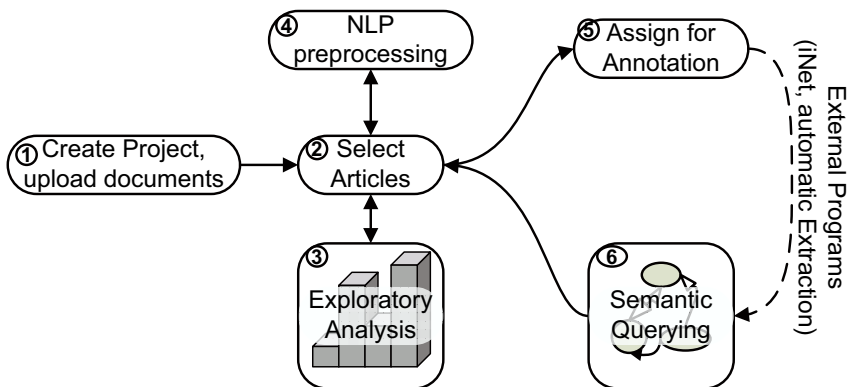


Figure 10.2: Using the AmCAT Navigator

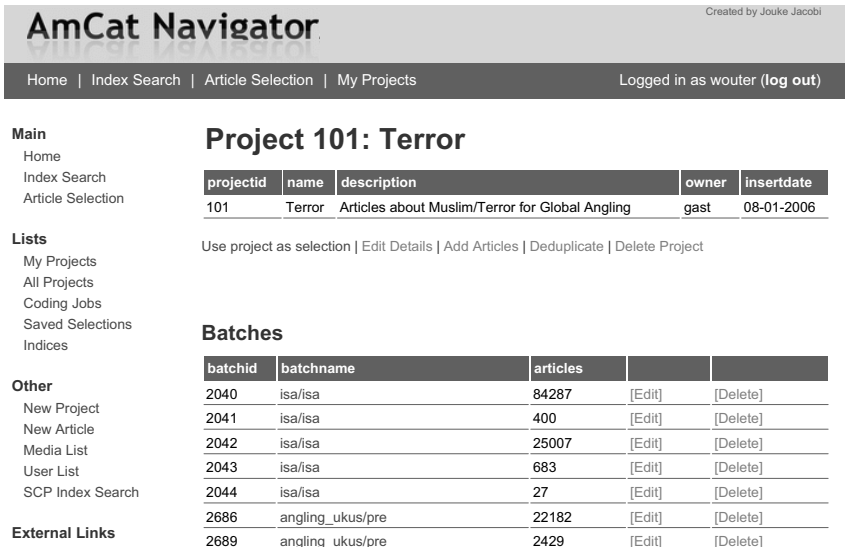


Figure 10.3: AmCAT Screenshot: Project Overview

that there are no gaps in the corpus. Article selections can also be the starting point of other analyses, by selecting the *Run Script* option. Via this option, external scripts can be launched that perform an analysis or preprocessing on documents, such as word counts, grammatical analysis, or deduplicating sets of articles. Figure 10.4 shows the article selection screen used to index all the articles in the ‘Terror’ project. Such an index allows for quick keyword-based searching.

(3) **Exploratory Analysis** In order to get a good feeling for our data, we will first conduct a series of exploratory analyses. As a first analysis, we plot the occurrence of the concepts Islam, Muslim, Terror, and Crime around 9/11 using simple wildcard searches. To understand the association patterns between concepts, we have two options: we can create a proximity query, requiring for example Islam and Terror to occur within 10 words, as illustrated in figure 10.5. In the form shown in the top part of the screenshot, the user can enter the queries to search for, in this case *terror** (any word starting with terror) within 5 words of *musli** or *isla** and a similar query for crime near Islam. Selecting a line graph of search term per year in the Output options results in the graph shown in the bottom of the screenshot. This graph shows how the co-occurrence of Islam with

Article Selection

Select

☒Project☐Batch☐Saved Selection☐Coding Job☐Article Ids☐SQL

Project

101 - Terror

Medium

220 - persberichten

301 - sunday times

302 - the washington post

303 - the times

304 - the independent

Date

Between

1

1

2001

and

1

1

2009

Headline

Action

☐Show Articles☐Show as table☒Run script☐Save as Selection

Run Script

New Lucene Index

Creates a new Lucene search index

name

Terror

splitParagraphs

No

Submit Query

Reset

Figure 10.4: AmCAT Screenshot: Article Selection

both terror and crime increases after 9/11, but the former association increases almost tenfold, while the latter approximately doubles. By clicking on any point of the line, a list such as that shown in the top right hand corner appears, showing the articles comprising that point, with the option of clicking an article to see the full text with the search terms highlighted. This ‘going back to the text’ is an essential feature to make sure that we actually found what we were searching for, and to find quotes and examples to make sense of the quantitative findings.

An alternative to proximity searching is creating a cluster map. This is accomplished by selecting the ‘Cluster Map’ in the Output options. Figure 10.6 shows such a cluster map for three search terms: terror*, crim*, and ‘musli* or isla*’. A cluster map shows one dot for each document that matches the keywords, clustered in circles each representing the set of keywords that matched the documents inside it. For example, the large circle in the top left corner contains 125 dots, each representing one document that contained ‘terror’ but none of the other keywords. The smaller circle to its right is at the intersection of ‘terror’ and ‘Islam’, and contains the 38 documents that matched both ‘terror’ and ‘Islam,’ but not ‘crime.’ The circle at the centre contains the 13 documents that matched all keywords. By clicking on one of the dots, the text of that documents opens on the

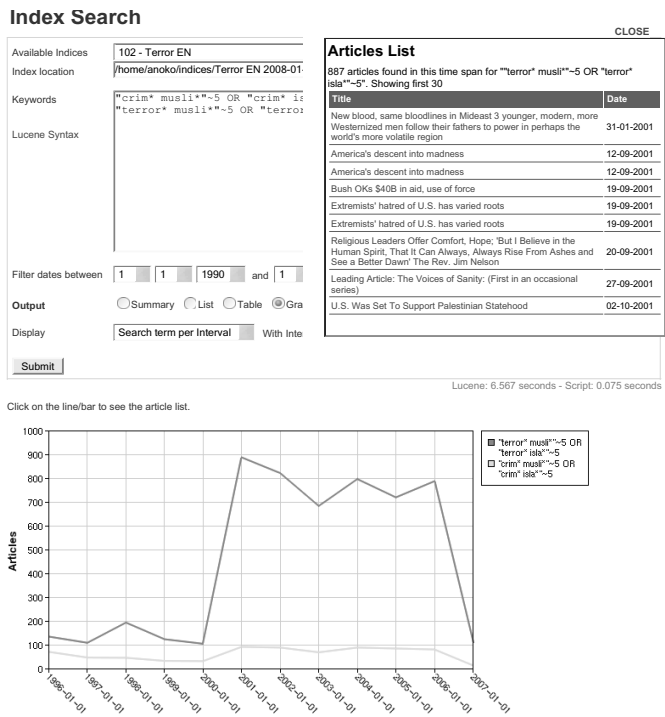


Figure 10.5: AmCAT Screenshot: Index Search results over time

side, again with the matched keywords highlighted.

Line graphs with proximity searches are a great way to quantitatively analyse a fixed set of co-occurrences in large data sets, and cluster maps are useful for exploring all co-occurrences between search terms in a smaller sample of documents. For both search mechanisms, if we have identified an interesting set of articles, for example the articles containing both Islam and Crime in the week after 9/11, we can use that article list as a selection in the Article Selection screen discussed above, focusing further analysis on that interesting subset. Additionally, we can save the results as a Comma Separated Values file to enable easy importing into programs like SPSS or Excel.

- (4) **NLP Preprocessing** To prepare our data for more sophisticated analyses, we can launch linguistic preprocessing scripts from the Article Selection screen. For example, POS tagging enables us to create lists of frequent adjectives in the corpus, or limit our frequency analy-

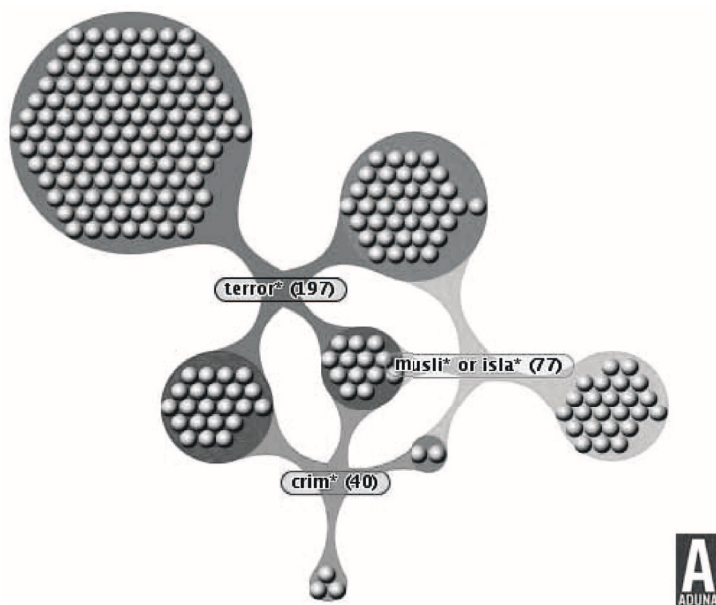


Figure 10.6: AmCAT Screenshot: Co-occurrences visualized as a Cluster Map

ses to only verbs or proper names.¹ Especially the latter is very useful given the frequent use of nouns as last names, such as Bush in English or Bos in Dutch. Additionally, by lemmatizing the corpus we can search on lemmas rather than conjugated word forms, thereby increasing the coverage of our search patterns without needing wildcard queries.² Finally, we can conduct a full syntactic parse of the corpus, allowing us to search for syntactic patterns rather than surface proximity.³ Each of these steps is launched by running the appropriate script in the Article Selection screen.

- (5) **Assign for Coding** The graphs produced by the exploratory analyses in step (3) were interesting but did not tell us how the Islam was framed. For this, we need to conduct a Semantic Network analysis. The first step of such an analysis is to extract the network, either manually or automatically. In both cases, we launch a script from the Article Selection screen based on the selection we made in step

¹See section 3.2.2 on page 44

²See section 3.2.3 on page 45

³See section 3.2.4 on page 46

| codingjobid | name | username | insertdate |
|-------------|--------|----------|------------------------|
| 272 | Terror | wva | 2006-09-29 02:32:04.83 |

Use Coding Job as selection | Change Job Owner | Delete Job (delete sets first)

Sets

| setnr | coder | articles | irrelevant | with arrows | done | |
|-------|-----------------|----------|------------|-------------|------|--------|
| Set 1 | wva (change) | 40 | 3 | 3 | 5 | Delete |
| Set 2 | nel (change) | 40 | 2 | 23 | 26 | Delete |
| Set 3 | jan (change) | 40 | 0 | 1 | 1 | Delete |
| Set 4 | angela (change) | 56 | 22 | 34 | 56 | Delete |
| Set 5 | bram (change) | 56 | 26 | 30 | 56 | Delete |
| Set 6 | ester (change) | 56 | 26 | 31 | 56 | Delete |
| Set 7 | erik (change) | 56 | 21 | 35 | 56 | Delete |

Figure 10.7: AmCAT Screenshot: Managing coding jobs

- (3). In our example, we decided to assign the selected articles for manual coding, creating a *coding job*: we specify the coding schema and ontology to use, and assign it to our coders in sets of 25 articles. In the coding jobs screen, we can now see these jobs and keep track of the progress made by the coders, as shown in figure 10.7. Section 10.3 describes how the coders execute this job using iNet.
- (6) **Semantic Querying** iNet and the automatic extraction programs both store the codings in the database using RDF, as described in chapter 8. This RDF repository contains both the extracted network and the background knowledge in the ontology. Using the semantic querying techniques described in chapter 9, we can now look for very specific patterns in this network. These patterns can be tabulated or shown as a graph, as illustrated in figure 9.2 on page 170. From these queries, we can go back to the Article Selection screen, for example to zoom in on documents showing interesting patterns and refine the analysis of these documents.

10.2.3 Technical Design and Implementation

Figure 10.8 shows the general architectural design of the AmCAT system. The system consists of two servers, represented by grey rectangles: the database server hosting the relational and RDF data stores, and the script server hosting the Navigator and analysis scripts. The right-hand side grey box represents a client, using either a web browser to access the AmCAT navigator or iNet (see section 10.3). Additionally, (power) users can directly access the database or RDF repository to launch cus-

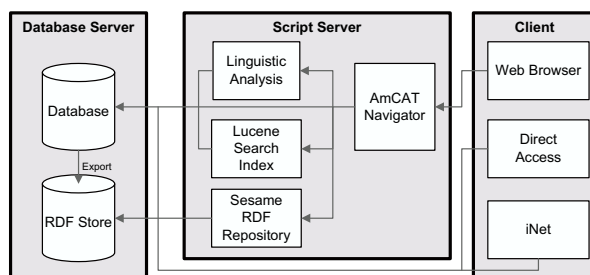


Figure 10.8: AmCAT System Architecture

The large grey boxes represent computers, the smaller boxes components, and the cylindrical shapes data stores. The arrows indicate dependencies.

tom queries. For example, SPSS and Excel both have built-in database-querying facilities, making it possible to extract certain data from SPSS directly without using the AmCAT navigator.

The Script Server

AmCAT Navigator The AmCAT navigator is a website hosted on an Apache web server. The functionality is provided using Python scripts, which access the database directly or by calling other scripts.

Linguistic Analysis The linguistic analysis scripts all obtain their input data from the database and store their results back into the database. Currently, tokenisation and Sentence Boundary Detection are performed using regular expressions.⁴ For POS tagging, both the Tilburg Memory Based Tagger (MBT Daelemans et al., 2007) and a Brill Tagger (Drenth, 1997) are installed.⁵ Lemmatisation is performed by taking the most frequent lemma per tag from the CELEX lexicon (Burnage, 1990), using a Porter Stemmer to guess unknown words (Gaustad and Bouma, 2001).⁶ Additionally, the Alpino syntactic parser is installed, which creates a dependency graph for each sentence and also performs POS tagging and lemmatisation.⁷

Lucene Search Index To facilitate rapid full-text searching, Lucene (an open source search engine toolkit) is used to create and search through

⁴See section 3.2.1 on page 43

⁵See section 3.2.2 on page 44

⁶See section 3.2.3 on page 45

⁷See section 3.2.4 on page 46

indices⁸. Lucene gets its input directly from the database and stores the indices on the script server. The Lucene source was modified to return raw frequencies rather than relevance scores, and to allow combining wildcard and proximity searches.

Sesame RDF Repository As described in part III, RDF is used to store the codings and background knowledge. For this purpose, the open source Sesame RDF Repository (Broekstra, 2005) is installed as part of the AmCAT system. This repository consists of two components: the actual Sesame server runs on the script server, and receives queries through a web interface. For data storage, it uses the relational database on the Database Server. This has the advantage that all the important data in the system resides in one place, making for an easier backup strategy.

The Database Server

The AmCAT database is a *Relational Database*, meaning that the data are stored in tables with a fixed number of columns. These tables are related by references. For example, if an article consists of a number of sentences, the article table would be related to the sentences table with a one-to-many reference. The database maintains *referential integrity*, meaning that it is impossible to create a sentence without an article, or to delete an article without also deleting the sentences contained in it. *SQL*, or Serialised Query Language, is a standardised language for querying relational databases. At the basic level, it allows a user to specify a set of (related) input tables and select specific rows and columns for output.

The tables in the AmCAT database can be categorised into a number of distinct but related components. Figure 10.9 gives an overview of these components, shown as large rounded rectangles. The smaller rectangles are the important tables from each component, and the arrows are the relations between the tables, pointing from the ‘child’ table (such as sentences) to the ‘parent’ table (such as articles).

Documents A central task of the database is to store the newspaper articles or other documents that are to be analysed. This part of the database is fairly straightforward: the *articles* table contains the metadata about each article, and the *texts* table contains the text itself. For each article, multiple copies of the text might be present; currently the system stores the raw text and optionally a lemmatised version. Articles are categorised into batches, which contain articles based on the same search string or other retrieval criteria, and batches are grouped into projects.

⁸<http://lucene.apache.org>

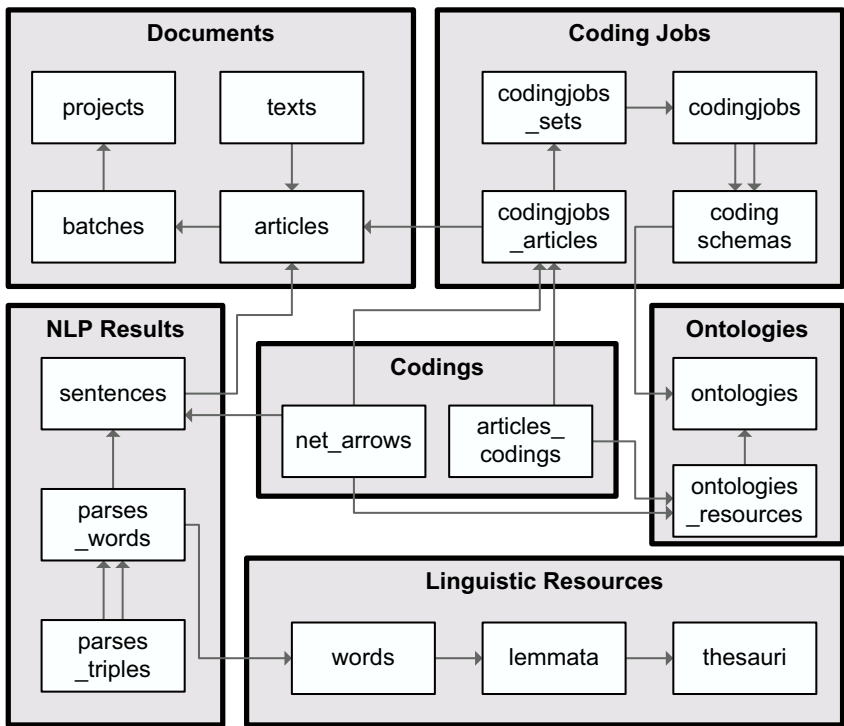


Figure 10.9: AmCAT Database Structure

In the example in the previous section, when we created a new project in the AmCAT navigator, an entry was added to the *projects* table. On adding articles, one entry is created in the *batches* table for each unique keyword query, all related to the project entry. Each article added an entry in the *articles* table related to the correct batch and an entry in the *texts* table related to the article.

Coding Jobs The *codingjobs* table contains all jobs that have been assigned to users. For each codingjob, the coding schema to be used is specified by referencing the *codingschemas* table. For each field in a coding schema, there will be a related entry in the *codingschemas_fields* table. One coding job can consist of multiple sets, each of which has an entry in the *codingjobs_sets* table, and related to a user in the *users* table. The *codingjobs_articles* table contains an entry for each article assigned to a coder. If the same article has been assigned to multiple coders, it will have multiple entries in this table, each related to a different *codingjob_set* and to

the same *article*. In the example scenario, we created one entry in the *codingjobs* table. For each set of 25 articles, we also created one entry in the *codingjobs_sets* table related to that codingjob, and 25 entries in the *codingjobs_articles* table related to that codingjob_set.

Codings For reasons of performance and referential integrity, codings are stored in the relational database and exported to RDF rather than stored in RDF directly. Although the coding schema allows the specification of the table to store the codings in, the default is to use the tables *articles_codings* and *net_arrows* for article- and sentence-level codings, respectively. For each article, one entry is created in the *articles_codings* table related to the *codingjobs_articles* containing the codings for that article. Referencing the *codingjobs_articles* rather than the *articles* table directly allows the same article to be coded multiple times. For the sentence-level codings, an entry in *net_arrows* is created for each NET arrow, related to the appropriate entry in the *sentences* table. In both cases, references to *ontologies_resources* (and other reference tables) ensures the integrity of entered codings.

Ontologies The ontologies used by the NET coding are stored in RDF. However, for each ontology there is a ‘liaison’ entry in the *ontologies* table, and each resource has an entry in the *ontologies_resources* table containing URL and label.

Linguistic Resources The database contains a *lemmata* table filled with all Dutch lemmata from the CELEX database. Additionally, for each lemma, one or more *words* are stored along with the frequency of that word. The *thesauri* table contains the lemma categories according to Brouwers’ thesaurus.

Linguistic Preprocessing Results The *texts* table mentioned above contains the raw text of each article. As will be described below, the AmCAT system contains a number of preprocessing tools. The *sentences* table contains the text split into sentences by sentence boundary detection. *Parses_words* contains one entry for each word in the sentence, referencing the correct entry in the *words* table. Thus, while the *sentences* table contains the actual text, *parses_words* contains only the word IDs. Finally, *parses_triples* contains the triples output by dependency parsing as a dependency relation between two entries from *parses_words*. Note that this setup contains redundancy: the *sentences*, *parses_words*, and *texts* tables all contain the raw text of the article. The reason for using this setup is

that it does not force all text to be analysed in the same way, maintaining flexibility at the price of increased storage size.

10.2.4 *AmCAT Use and Performance*

The AmCAT system is in use for research by the Vrije Universiteit, the Netherlands News Monitor, and the University of Amsterdam. The system has been used to analyse 13 million newspaper articles in 7 languages, around 100,000 Dutch, French, and British parliamentary debates, 1,500 Dutch television news transcripts, 2,191 surveys with open questions, and 12 million forum and blog postings. The system has been actively used by 13 users since March, 2004. The exploratory analysis functionality has been used for practical assignments in content and media analysis courses at the Vrije Universiteit, the University of Amsterdam, and the University of Friedrichshafen (Germany).

Currently, the database server runs Microsoft SQL Server on two 2.7 GHz processors, 4 megabytes internal memory and 3 RAID 10 15,000 rpm SCSI hard disks; the script server runs Debian Linux on a computer with 2.8 GHz processor and 1 GB internal memory. On this hardware, creating a frequency per month per source table for the Terror project described above (167,386 articles) took 2.9 seconds. Creating the index for this project took around an hour but after the indexing is completed the keyword query to create the graph took 7 seconds. Lemmatising and POS tagging speed is approximately 2,000 articles per minute.

10.3 The iNet Coding Program

Ideally, automatic coding would supplant manual newspaper coding completely. However, performance considerations and the fact that Machine Learning approaches need training data mean that manual coding will remain necessary in the foreseeable future. Since the codings needed for Network Text Analysis are more complex than for most other Content Analysis methods, the input program iNet was created to allow easy and efficient coding. This program is a replacement for the earlier DOS-based CETA program described in De Ridder (1994).

10.3.1 *iNet: functional design*

iNet was created for the purpose of assisting manual coding using Semantic Network methods. The main focus is on making coding easier and more reliable. Coders are able to log in, and select documents from a list of documents they are assigned to code. During the coding, they are

assisted by automatic completion of fields from the ontology and by the ability to code quickly using only the keyboard. Errors, such as omitting a required field, are signalled as such but do not block further coding, thereby preventing a break in the concentration of the coder (the coder has to correct the error before closing a document to prevent malformed data from entering the system). A visualisation of the codings is provided to help the coder ascertain that the graph he is creating matches the meaning of the document.

A second goal is workflow support for the analyst by means of a tight integration with the AmCAT database. Documents can be assigned to coders using the AmCAT Navigator, and codings are immediately stored in the AmCAT database so that the analyst can check progress and view results immediately, without having to send around documents and collect and read result files. The results can be exported to CSV in the AmCAT navigator or directly retrieved from the database using SPSS or Excel.

The third goal in designing iNet is flexibility with regard to the coding schema. For each coding job, a coding schema can be defined that specifies which fields have to be coded at the article level and at the sentence level, and what values are valid for each field. Hence, iNet can easily be used for content analysis methods other than Semantic Network Analysis: it can be used for any method that requires certain variables to be filled in for certain units, in other words, practically every thematic content analysis method. When a coder selects a job, the schema is automatically determined from the job definition and the right fields are displayed, minimising the possibility for confusion by the coder.

A fourth design goal is extendability of the application. Although the second goal requires the application to integrate tightly with the AmCAT system, iNet is designed to make it possible to extend this to other systems or to standalone operation by creating new *plug-ins*⁹ to handle article input and coding output. Moreover, it is possible to add extensions for including non-textual material such as video or pictures, or defining new kinds of coding fields.

iNet was explicitly not created to perform automatic content analysis or to perform project or document management or analysis; in that sense it is more narrow than many other software packages.¹⁰ It is assumed that iNet is used in conjunction with AmCAT or an equivalent infrastructure, although iNet could be used on its own on smaller projects.

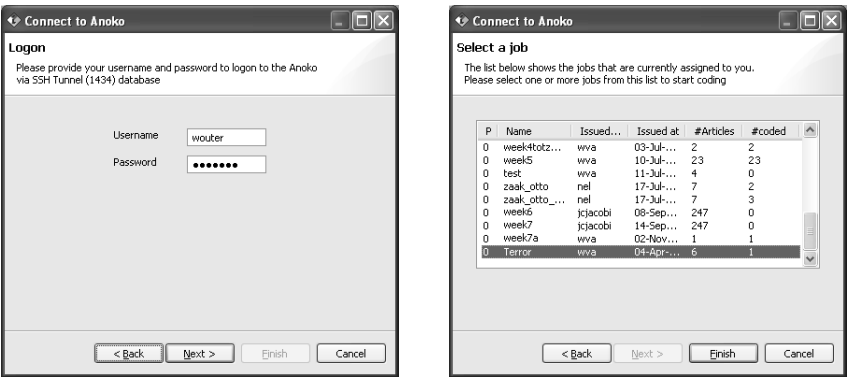
⁹See page 200

¹⁰See section 2.6 on page 35

10.3.2 Using iNet

Section 10.2.2 showed how the AmCAT Navigator can be used to search through documents and select interesting sets of documents for further analysis, such as manual analysis using iNet. Suppose that we have decided to manually code a selection of articles, for instance the articles on the day after 9/11 containing muslim* in the Washington Post. In the *Article Selection* screen, we can create a *Coding Job* in a similar way to creating a new index. For a coding job, we select the relevant articles, specify the coder(s) to receive the job. The current job only has 6 articles. For larger jobs, the script can break the job into smaller sets, for example 25 articles per set. This creates coding sets of a manageable size, and makes it easier to (re)distribute articles among coders. We also specify the coding schema to be used, in our case a normal NET schema with an appropriate ontology for the coding. Since we are interested in looking at the way Muslims are framed post-9/11, the ontology contains relevant societal and political actors and a number of labels, such as citizen, criminal, and terrorist, that can be applied to actors. This ontology has to be created in RDF externally, for example in an ontology editing tool such as Protege (Knublauch et al., 2004). Once the job has been set up, a coder would roughly follow these steps in iNet:

- (1) **Login Screen** The selected coder logs on using iNet (figure 10.10a)
- (2) **Job List** The coder sees the jobs that have been assigned to him or her, including the job created above. As shown in figure 10.10b, each job



(a) Login Dialog

(b) Job List Dialog

Figure 10.10: iNet Screenshot: Login and Job List dialogs

is listed together with the total number of articles and the number of articles that have already been coded.

- (3) **Articles** When the coder selects a job to code, the main iNet screen is loaded. Figure 10.11 shows the top half of this screen. The Articles *view* on the left-hand side shows the articles contained in this job, and whether those articles have already been coded. The right-hand side shows the text of the selected article, split (unitised) into sentences for coding.
- (4) **Coding Editor** Double-clicking one of the articles in the Articles view opens a new *coding editor* tab for that article, shown in the left-hand side of figure 10.12. The top half of the editor contains sentence-level codings, in this case only relevance and a comments field. The bottom half contains the sentence-level codings, i.e. the source, subject, quality, type, and object of the NET triples. Each triple is connected to a specific unit (sentence). During coding, that sentence is highlighted in the Text view to avoid mistakes. A network visualisation of the codings is displayed in the Visualiser view, as shown in the right-hand side of the figure. This view is updated during coding to allow the coder to check whether the network matches his reading of the article.
- (5) **Coding** The coder can code the selected article by filling in the relevant values in the bottom half of the iNet screen. When a coder starts typing, iNet suggests possible completions from the ontology, as shown in figure 10.13. If an unknown object is used, iNet signals an error by colouring the relevant cell dark red and the whole row light red, such as the 'sdfg' value in unit 2.1 in the screenshot. In the right-hand side of the screenshot, the Ontology view is shown. This view contains a searchable tree representation of the ontology. In addition to typing in a name, the coder can also click the ellipsis (...) button to open a dialog containing the same representation as

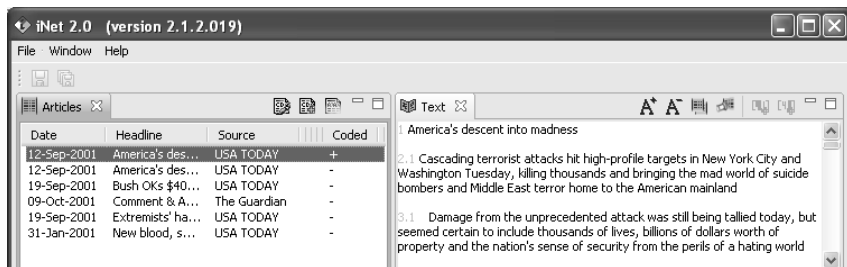


Figure 10.11: iNet Screenshot (top half): Articles and Text Views

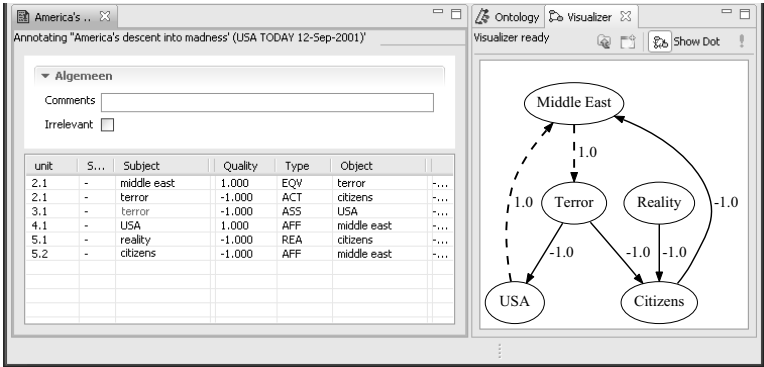


Figure 10.12: iNet Screenshot (bottom half): Coding Screen and Visualizer

the Ontology view, whereby the coder can go through the tree and select the appropriate object. All coding can be done either with the mouse or the keyboard.

(6) **Save** After finishing coding, the user can save the codings to the database by clicking the Disk icon shown in the top of figure 10.11. To prevent invalid codings from entering the database, codings can only be saved if there are no errors in the coding.

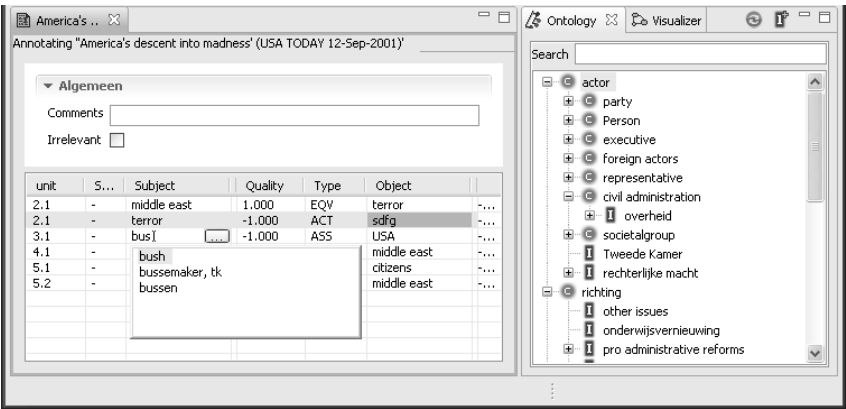


Figure 10.13: iNet Screenshot (bottom half): Autocomplete and Ontology

10.3.3 iNet technical implementation

The following sections provide some details on the technical implementation of iNet, especially the integration with the AmCAT database structure, the internal *plug-in* structure of iNet, and the possibilities for extending iNet.

iNet and the AmCAT Database

Figure 10.14 visualises the iNet workflow and its integration with the AmCAT database, following the same steps as in the walk-through above. *Net_arrows* is highlighted as the central table containing all sentence-level codings and to avoid confusingly crossing lines in the figure. The article-level codings contained in *articles_codings* are used identically to the *net_arrows*, and have been left out of the picture for clarity.

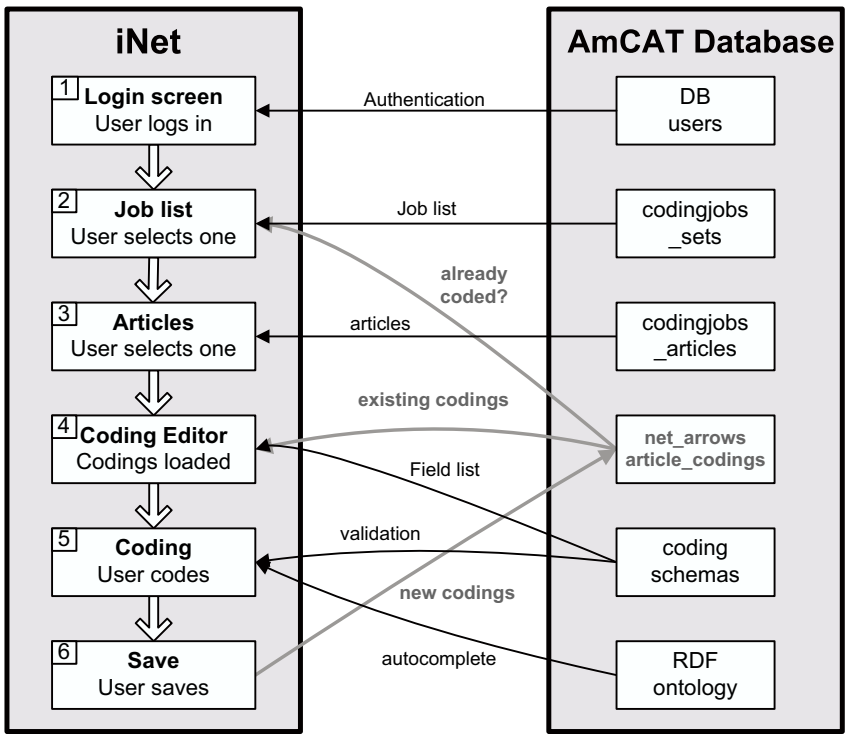


Figure 10.14: iNet Workflow and Integration with AmCAT

- (1) **Login Screen** The session begins with a user connecting to the AmCAT database and logging onto the system. The user is authenticated using the authentication mechanism of the database.
- (2) **Job List** After the user logs on, the table *codingjobs_sets* is used to show the list of coding jobs that are assigned to him or her. This information is joined with *net_arrows* to show whether the jobs have already been coded.
- (3) **Articles** After selecting a job, the articles are loaded from the *articles* table. The sentences comprising each article are also loaded from the *sentences* table (not shown).
- (4) **Coding Editor** When the editor is opened for a specific article, the codings are loaded from the *net_arrows* table. Moreover, the *coding_schemas* table is consulted to determine which fields to display for the article- and sentence-level codings.
- (5) **Coding** During coding, the input is validated according to the information from *coding_schemas*. For the fields linked to the ontology, such as NET subject and object, the RDF repository containing the ontology is used to display the tree representation, validate input, and suggest completions.
- (6) **Save** When the user saves codings, the input is validated as above, and the new codings are stored in the *net_arrows* table.

iNet Architecture and Implementation

iNet is written in Java using the Eclipse Rich Client Platform (RCP). In Eclipse RCP, code is organised in units called *plug-ins*, which can contain one or more Java packages and other files such as graphical resources. Plug-ins can be coupled together in two ways: one plug-in can import the other plug-in, creating a tight coupling where the importing plug-in is dependent on the imported plug-in and can use all of its public classes and interfaces. Alternatively, a plug-in can define an extension point, essentially an interface declaration. Another plug-in can then extend the previous plug-in by implementing that interface. This creates a loose coupling, where the plug-in defining the extension point can interact with the extending plug-in without knowing which plug-ins actually do the extending. For example, by copying a new plug-in into the iNet directory that extends the Eclipse Views extension point, a new view will show up in iNet the next time it is launched, without modifying the rest of the program.

Figure 10.15 shows the main plug-ins that iNet consists of, the black arrows indicating imports (tight coupling) and the grey lines representing extensions (loose coupling). In the bottom left corner is the *org*.

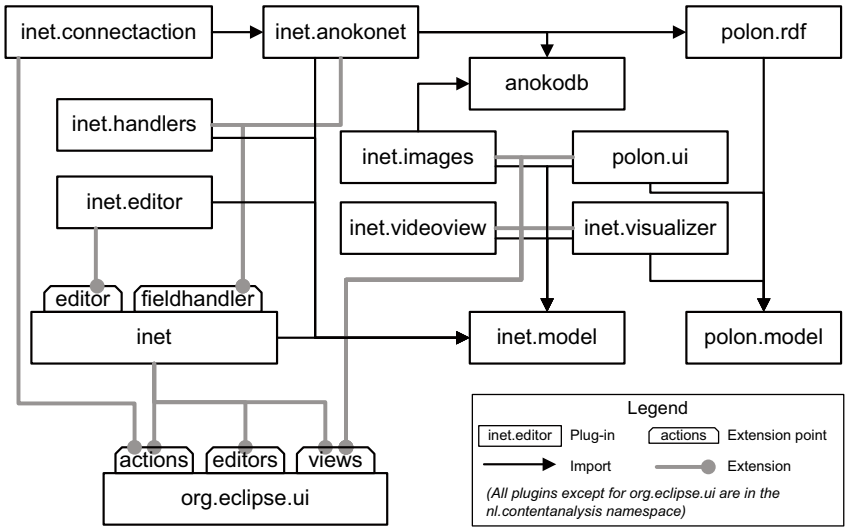


Figure 10.15: iNet Plug-in Structure and Dependencies

`eclipse.ui` plug-in, which is a built-in Eclipse plug-in that takes care of a large part of the User Interface. In particular, it provides three relevant extension points, *Actions*, *Editors*, and *Views*. *Actions* represent user-initiated actions such as menu options and toolbar buttons. *Editors* and *Views* are the main components of the Eclipse RCP interface, where editors such as the main coding editing screen in iNet are meant for editing a unit of data, and views give additional information on the data currently being edited, such as the view that shows the text being coded. Directly above `org.eclipse.ui` is the `inet`¹¹ plug-in, which is the main application plug-in. `inet` is responsible for launching the applications and communicating with the other plug-ins. For this, it imports the `inet.model` plug-in, which is an interface description of the data model behind iNet, defining concepts such as an Article and a Coding. As can be seen from the many arrows directed at `inet.model`, this plug-in is imported by almost all plug-ins and provides the shared vocabulary needed for interaction. Apart from a tight coupling with the model, the main `inet` plug-in also defines two extension points, `editor` and `fieldhandler`. By extending the `editor` extension point, a plug-in can register itself as an editor for certain document types. The `inet.editor` plug-in, shown above the `inet` plug-in, provides the default editor implementa-

¹¹This is actually the `nl.contentanalysis.inet` plug-in, but for the sake of readability the `nl.contentanalysis` is omitted for all plug-ins

tion, which shows the article codings as fields and the sentence codings as a table. *Fieldhandler* extensions define how to handle specific fields within an editor. Actually, this is a set of three extension points: *codingschemahandler* can provide functionality at the schema level; *coding-fieldhandler* extensions provide the (de)serialising functionality needed to turn the persisted code value into a useful object; *fieldeditors* provide editors to show and manipulate these objects. For example, for the arrow type choice field, the schema handler implements the logic that it has to be evaluative if the coded object is Ideal, which is an inter-field dependency. The field handler converts the integer value stored in the database into an object representing the correct value. Finally, the field editor shows the label in the editor, and provides a drop-down menu if the user edits the field. The handlers for the default field types, such as texts, numbers, and choice fields, are defined in the `inet.handlers` plug-in, which extends these extension points. Due to the interaction between these extension points, they are generally extended together and shown as one point in the Figure.

The other plug-in extending the *fieldhandler* extension points is the *anokonet* plug-in. This plug-in handles the interaction between the AmCAT system and iNet, and defines the ontology field used for the subject and object fields. Moreover, using the generic plug-in *anokodb*, it takes care of retrieving data from the database and saving new codings. In the upper left corner, `inet.connectaction` provides the 'Connect to AmCAT' action, and calls the `inet.anokonet` plug-in to retrieve the selected job and start the coding process. On the right-hand side of the figure, there are two plug-ins that provide the ontology functionality. `polon.model` is a general ontology model, and `polon.rdf` provides an implementation of this model based on `rdf`.

Finally, there are four plug-ins in the centre of the figure which provide additional iNet views. `inet.images` is used for displaying images associated with documents, such as newspaper photographs. It imports the *anokodb* to obtain the images from the database. *videoview* is used to display video fragments stored on a local medium, and allows timestamps to be inserted into the coding editor. `polon.ui` provides the tree-view ontology display shown in the bottom right-hand corner of figure 10.13. `inet.visualizer` shows the graph visualisation shown in the bottom of the screenshot, importing the ontology model for the ontology-based aggregation.

Customisability

iNet was developed with the NET method in mind. However, it was also designed to be easily customisable. At the simplest level, all views

and coding editors can be moved around, resized, and maximised by the user. This makes it possible to compare the codings of different coders of the same article by the researcher or supervisor by loading different jobs at the same moment.

A more substantial area of customisability is the coding schema. In that schema, the researcher can specify the fields to be coded for each job, including field type and possible values. iNet will automatically display the right fields for each job when the user connects to the database. This makes it easy to manage multiple projects with different coding requirements, such as extra fields or article-level codings, without needing different versions or configurations of the client program.

Another method for customising iNet is writing plug-ins. In fact, a large part of the basic functionality comes from a set of existing plug-ins. The program contains three main hooks for plug-ins:

Data Connection The current version of iNet is tightly integrated with the AmCAT system. However, this integration is mainly located in one plug-in that handles connecting to the database and loading and saving the internal data structure. It is possible to write a plug-in to load and save text and codings from other formats, such as XML, plain text, or MS Access databases, allowing for standalone use of iNet. These plug-ins are activated by menu commands, currently File→Connect→Connect to AmCAT.

Additional Views The screenshot shown above contains views that show the article list, the text of the current article, the ontology, and a network visualisation. A plug-in can contain a new view, and link it to the article currently being coded. For example, extra views have been made that show pictures and videos linked to the articles and, for the videos, allow the current timestamp to be inserted in the current code.

Annotation Plug-ins The final type of plug-ins deal with the way codings are displayed and edited. In fact, each of the data types in a coding schema describe which plug-in to use for displaying and editing a field. Thus, it is possible to create a new field type and write a custom editor by writing one plug-in and modifying the coding schema. Additionally, the coding schema can specify a plug-in that can contain project-specific logic at the level of codings rather than fields. For example, the NET coding plug-in has special rules for interaction between the subject, type, and object fields.

10.3.4 iNet Use and Performance

The first version of iNet dates from 2002. iNet 2.1, the version described in this chapter, was written in 2006 and has been used for the 2006 election study, the analyses by the Netherlands News Monitor, and for various small studies. In total, over 150 thousand NET triples have been coded by 37 different coders. During the 2006 election project, more than 10 coders would regularly be coding simultaneously. The query to directly extract the almost 40,000 triples from the 2006 election study and the associated metadata and issue and actor categorisation executes in 3.5 seconds.

10.4 Conclusion

This chapter described the AmCAT system, consisting of the AmCAT navigator and database and the iNet coding program. The AmCAT navigator is an easy-to-use web-based interface to the AmCAT database containing documents with associated metadata and codings. The navigator is also used to conduct exploratory analysis, linguistic preprocessing, and to assign documents to be coded manually. iNet is a program for manual coding specially designed for Semantic Network Analysis. Taken together, AmCAT is a mature and user-friendly system for managing and conducting content analysis.

Discussion and Conclusion

In the social sciences, Content Analysis is used to analyse messages such as newspaper articles as part of studies into the interaction of the message with its social context: Why was it sent? How is it interpreted? How will that affect the receiver? Semantic Network Analysis is a Content Analysis technique that works by extracting a network of relations between objects, such as political actors and issues, and deriving the answer to the research question from that network. The goal of this thesis was to investigate whether it is possible to leverage recent advances in Natural Language Processing to facilitate the automatic extraction of Semantic Networks from textual messages, and to use Knowledge Representation to represent these networks and combine them with background knowledge about the actors and issues involved in order to answer relevant research questions by querying the network. In particular, the thesis tried to answer two Research Questions:

- RQ1 Can we automate the extraction of Semantic Networks from text in a way that is useful for Social Science?*
- RQ2 Can we represent Semantic Network data in a formal manner and query that representation to obtain the information needed for answering Social Science research questions?*

As summarised below, this thesis gives a positive answer to both research questions. The Natural Language Processing techniques described and implemented in part II allow the automatic extraction of Semantic

Networks with sufficient accuracy to provide valid answers to Communication Research questions. The RDF-based Knowledge Representation described in part III allows both the extracted Semantic Network and the background knowledge to be formally represented, and research questions can be expressed as queries or patterns over this representation and can be automatically identified.

This represents an important contribution to the field of Content Analysis. The automatic extraction of Semantic Networks takes automatic content analysis to a new level, showing that automatic analysis is feasible for more than simple word counting. This makes it possible to answer research questions automatically that previously required costly manual analysis. The representation and querying of the extracted networks show how a single extracted network can be used to answer different research questions by combining the network with the appropriate background knowledge. Moreover, the clear semantics of the representation and the queries make it easier to share, combine, and improve the extracted networks and the queries used to answer research questions.

Additionally, the techniques described in this thesis can play an important role in bringing quantitative communication science closer to the public. Automatic analysis makes it possible to quickly analyse newspaper coverage of important (media) events, creating timely empirical data to contribute to the often quick and intense public debates such events evoke. Moreover, the representation and querying techniques make these empirical data more accessible, allowing interested parties to search the news semantically rather than by using keywords, making it easier for empirical evidence to be used in public debate.

Summary

In part I, chapters 2 – 4 provided some background material on the fields that this thesis draws on: Content Analysis, Natural Language Processing, and Knowledge Representation. Chapter 2 also defined Semantic Network Analysis as a Content Analysis technique. Figure 11.1, replicated from figure 1.1, summarises the process of Semantic Network Analysis: From the source material, a network representation is extracted either manually or automatically. This network representation contains the relations between objects such as actors and issues and represents the text as closely as possible. Subsequently, this representation is queried to answer the social science research question. This querying uses background knowledge to aggregate the actors and issues in the network to more abstract concepts, and operationalises the research question as a pattern on this abstract network. In this way, the extraction step is in-

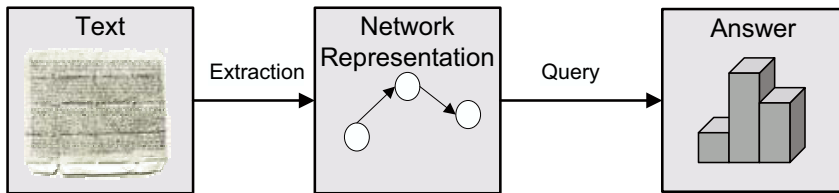


Figure 11.1: Semantic Network Analysis (replica of figure 1.1)

dependent of the research question and separate from a querying step, which is in turn independent of the extraction. This makes it easier to combine, share, and reuse both Semantic Network data and the queries and background knowledge used to operationalise the research question.

Part II investigated the first research question: can Semantic Networks be automatically extracted from text? Chapter 5 showed how co-occurrences between actors and events based on conditional reading probability can be interpreted as a simple association-based Semantic Network. Additionally, it showed how parts of this network can be interpreted as Associative Frames, embedding automatic extraction in the theories of Framing and second-level Agenda Setting. In Associative Framing, association is operationalised as conditional reading chance: The association between concepts A and B is the chance of a random message containing B given that it contains A. This operationalisation is a simple, asymmetric measure that is easy to interpret. Finally, a case study of globalisation and localisation of the news coverage of terrorist events showed how these associative networks can be used to answer relevant social science research questions.

Chapter 6 enriched this Semantic Network by using patterns over grammatical relations to distinguish between the source, agent (semantic subject), and patient (semantic object) roles. This creates a directed Semantic Network, which can be used for answering research questions that require distinguishing the acting and receiving objects in an association. To test the accuracy of this extraction, the output of the system was compared to a manual Semantic Network analysis of the Dutch 2006 parliamentary election coverage (Kleinnijenhuis et al., 2007a). At the sentence level, the similarity was acceptable but left room for improvement. Additionally, the automatically and manually extracted networks at the level of analysis were compared for a number of substantive hypotheses concerning media authority. At this level of aggregation, performance was good. Moreover, the results of modelling media authority using a number of independent 'real-world' variables were compared, and the

model outcome was highly similar for the manually and automatically extracted networks. This means that extracting the directed network using syntactic patterns is feasible and immediately useful for communication research, provided the level of aggregation is high enough.

The last enrichment to this network was done in chapter 7, where Machine Learning techniques from Sentiment Analysis were used to determine whether a relation is positive or negative. This creates a signed network that can be used for various communication research questions dealing with issue positions or support and criticism. In order to test these techniques, the automatically extracted signed network was again compared to the manual Semantic Network Analysis of the Dutch 2006 elections. Similar to the directed network, performance at the sentence level was acceptable but left room for improvement, and performance was significantly better than chance or a simple baseline system. Performance at the level of analysis was tested by reproducing a number of analyses from Kleinnijenhuis et al. (2007a), and performance was good in the majority of cases. As above, this showed that signed networks can be automatically extracted from text sufficiently well to answer interesting communication research questions, as long as the unit of analysis is large enough.

The second research question, whether the extracted networks can be formally represented and queried to answer relevant social science research questions using Knowledge Representation techniques, was investigated in part III. Chapter 8 examined the representation problem: How can Semantic Networks be represented in RDF, especially regarding meta statements (statements about statements) and temporal roles? Meta statements are necessary since the primitive of Semantic Networks is a relation, about which additional information needs to be expressed, such as connection strength, quoted source, and metadata such as publication date. Temporal roles are needed in the background knowledge since the roles played by actors, such as political functions or party membership, are very important in interpreting the network, and these roles are not static, especially not in longitudinal research. Although RDF does not directly allow meta statements and temporal roles to be represented, both can be expressed using a variety of more complicated structures. It was decided to use RDFS reification for meta statements and adjunct instances for temporal roles. Custom RDF reasoning was used to activate the correct roles depending on the publishing date of the coded text. Finally, this chapter described the background ontology used in the 2006 election study, showing how the temporal roles and a multiple issue hierarchy can be used to represent the necessary political background knowledge.

Chapter 9 addressed querying the representation of the Semantic Network described in chapter 8 to answer social science research questions,

for example on patterns of conflict or cooperation between actors or issue position made by actors. To hide the complexity of the reification and adjunct instances from the user, a custom querying language was developed on the Semantic Network relations and background knowledge, and queries in this language were automatically translated to standard RDF queries. A prototype web application was developed to execute such queries on the 2006 election study data. The results can be seen on the aggregate level or on the level of individual objects and articles. Moreover, the matched relations can be visualised as part of the overall network of individual articles or as a standalone network. By duplicating a number of analyses from the original election study, it was shown that the queries can be used to obtain the answer to the research questions posed in an actual study.

Finally, chapter 10 in part IV presented the AmCAT and iNet programs. AmCAT (Amsterdam Content Analysis Toolkit) is a document database and management system that can be used to manage content analysis projects, conduct keyword-based (exploratory) analyses, assign documents to coders for manual coding, and launch programs to conduct linguistic preprocessing and automatic coding. iNet is a manual coding program that aims to make coding accurate and efficient by integrating tightly with the AmCAT system, allowing autocompletion and schema-based validation, and immediate visualisation of the resulting network for checking whether the coding matches the meaning of the document. In contrast to the experimental or prototypical programs created in parts II and III, these programs are relatively mature and already in use for research and education by a number of scholars at different universities.

Discussion

This thesis has shown how existing techniques in Artificial Intelligence can be used and adapted to aid the extraction, representation, and querying of Semantic Networks. This is an important step forward for Semantic Network Analysis, and represents a contribution to the study of communication in the social sciences. This section will discuss some of the decisions and consequences regarding automating the extraction of Semantic Networks and representing and querying them.

Evaluating Performance

In chapters 6 and 7, performance was measured by comparing the automatically extracted Semantic Networks to existing manually extracted networks. In such a comparison, the manual analysis functions as a Gold

Standard. This raises the questions of what metric should be used to calculate performance. In manual Content Analysis, reliability is generally assessed by comparing the output of two different coders on the same material. Two popular metrics for reliability are Krippendorff's alpha and Cohen's kappa. These metrics, however, assume a situation of comparing two (fallible) coders, and no source of 'true' codings. Since it is assumed that the Gold Standard is correct, these metrics are not necessarily appropriate. In Machine Learning or Statistical Natural Language Processing, performance is often measured as an F1 Score.¹ This score is independent of true negatives, which are often not informative, and does not correct for chance agreement. The scores attained by a system are often compared to existing systems, or to baselines established by purely guessing the answers or by very simple systems.

A second question on evaluating performance is: How good is good enough? Authors such as Landis and Koch (1977, p.165) and Krippendorff (2004, p.241) give a tentative indication for how high the reliability should be to consider the coding to be sufficiently reliable. In Machine Learning, the goal is generally to improve on existing systems rather than judging whether a system is 'good enough', so the performance increase is more important than the absolute performance of a system. Most of the F1 scores calculated in the referenced chapters are around the minimal reliability considered acceptable, so the guidelines cited above do not give a conclusive answer, even if it is assumed that F1 scores can be directly compared to alpha or kappa scores. The performance is better than chance, possibly acceptable, and certainly not perfect.

A more informative way of judging whether the performance is good enough involves thinking about *what* it is supposed to be good enough for. This brings us back to the criterion defined in the introduction: The extraction techniques need to perform sufficiently to give a valid answer to relevant social scientific research questions. This can be translated into requiring that the answer to the research question based on the automatically extracted network should be highly similar to the answer derived from the manually extracted network: Performance should be measured at the desired level of analysis (cf. Krippendorff, 2004, p.292). For general-purpose techniques such as the ones presented in this thesis, this means that performance needs to be tested on a collection of questions that are representative of the questions the techniques will be used to answer. Both chapter 6 and chapter 7 present a variety of research questions and determine the correlation between the answers from the automatically and manually extracted networks. In most cases the results were very promising, but in some cases the automatic analysis deviated substantially from manual analysis. Three factors are most

¹See section 3.4 on page 49

important in determining the performance at the level of analysis: The performance of the underlying system, the degree to which the errors are systematic, and the granularity (i.e. number of units of measurement that were aggregated in each unit of analysis). In general, the better the system performs at the unit of measurement, the more the errors made by the system resemble white noise, and the more units of measurement are aggregated to form one unit of analysis, the higher the correlation will be.

A final question regarding evaluating performance is whether the manual coding is a good Gold Standard. Inherent in Content Analysis is that coders need to do some interpreting of the texts, and texts are often ambiguous to begin with. Thus, even if the system were perfect, it would not necessarily be in perfect agreement with coders. Standard reliability tests of the human coding give an indication of their agreement, but they could agree on the wrong decision by convergence during training or due to mistakes in the codebook. That texts are ambiguous means that sometimes two different answers can both be acceptable. Two approaches to overcome these problems are expert judgements and measuring predictive validity. By asking domain experts whether the automatic codings are acceptable, it is possible to circumvent the problem of different correct codings for the same text. Predictive validity requires testing whether the measured quantity can be used to successfully predict another quantity that is measured independently, such as opinion polls. Chapter 6 conducted two relatively simple tests to measure expert judgement score and predictive validity, and attained good scores on both measures.

Improving Performance

Chapter 6 presented a rule-based system to derive semantic roles from the syntactic structure of sentences. The performance of such systems can be improved by spending more time developing and testing the rules. The literature in Natural Language Processing, however, suggests that there is a limit to the efficacy of rule-based systems (e.g. Manning and Schütze, 2002, p.3,15–19): Language is riddled with ambiguity and complexity, and rule-based systems soon become so complex that it is very difficult to improve one part without causing unintended degradation in another part. A possible solution to this problem is using Machine Learning to learn or improve the rules, and since there is a good amount of training data available, this is probably the most promising direction to take. One possibility is to use a transformational approach as used in the Brill Part-Of-Speech tagger (Brill, 1995; Manning and Schütze, 2002, p.362), where error correction transformations are learnt on top of the

output of the rule-based system. A more pure machine-learning approach could be similar to the opinion detection and attribution by Choi et al. (2006)², or viewing the problem as a graph transformation task (from syntax graph to semantic role graph), and using techniques such as described by Jijkoun (2007).

Chapter 7 presented a system to determine the polarity of relations based on Machine Learning techniques. For improving the performance of Machine Learning systems, two aspects are important: The features of the texts that are used by the learner, and the size and quality of the training corpus. Chapter 7 listed a number of possibilities for improving the feature set, which should lead to improved performance. The largest gain, however, is probably in improving the training corpus. The work presented in chapter 7 was based on a selection of statements from the Dutch 2006 election study. Previous research at the Vrije Universiteit and University of Amsterdam resulted in a number of other Semantic Network data sets that could be combined with the 2006 elections: the election studies of 1994 – 2003 as described in Kleinnijenhuis et al. (2003, 1998, 1995), the Dutch campaigns for the referendum on the EU constitution (Kleinnijenhuis et al., 2005), a longitudinal investigation of economic news (Kleinnijenhuis et al., 1997), the Bosnian war coverage described in Ruigrok (2005), a content analysis conducted for the Dutch government (Kleinnijenhuis and Van Hoof, 2008), content analyses on the coverage of the affairs around Princess Mabel Wisse Smit, Ayaan Hirsi Ali, and Geert Wilders in the Netherlands (Scholten et al., 2007). If these data sets can be combined, it would result in a data set of over 250,000 triples spanning more than a decade and on various (mostly politics-related) topics. Creating this combined data set is not a trivial task, however. This is partly because the vocabulary used in those data sets have to be aligned with the ontology described in section 8.4. A second problem concerns matching the statements to the text. In all cases, the nodes and relations in the Semantic Networks are not matched to the words in the text³. For the earlier data, the statements are not even matched at the sentence level: The data contains the metadata such as headline and date of the newspaper article it is derived from, but it does not contain a sentence identifier. Moreover, the articles have to be obtained and matched using the metadata, which is not always possible, either due to errors or omissions in the metadata, or simply because the articles are not available digitally, which is especially problematic for the older articles. If this combined data set can be created, it can be expected to greatly enhance performance of the Machine Learning methods, as it would represent a forty-fold increase in training data; even if only the more recent studies

²See section 7.4 on page 122

³See section 7.5.3 on page 129

can be combined and matched with the original texts, the resulting data set will still mean a tenfold increase in training data.

Generalisability

The question of generalisability is whether the techniques presented in this thesis will also work for other domains, languages, and research questions. The choice of domain and language have already been discussed in the introduction: The systems presented in chapters 6 and 7 cannot be directly used for languages other than Dutch, and might not perform as well on texts other than political newspaper articles. However, since not many assumptions are made that are specific to newspaper styles, the system can still be expected to perform reasonably well on other texts, such as parliamentary debates. Moreover, part of the domain-specific information, such as the names of relevant actors, functions, and issues, is derived from the ontology. Hence, if a good ontology is made for other domains, the system should also perform well on texts from those domains. Whether these expectations hold true should ultimately be determined empirically. In terms of language, even though the currently developed system will certainly not work for other languages, the techniques presented in those chapters can be used to create new systems for other languages, provided the preprocessing tools and training data are available. Moreover, showing that systems with acceptable performance can be developed for Dutch makes it quite likely that such systems can be developed for other languages, as the amount and quality of resources for a number of major languages is at least on the same level as those for Dutch. Finally, the performance at the level of analysis reported in those chapters is determined on a limited number of use cases and research questions. Although those questions were not selected to be easy to answer, it is an open question whether these performance results are generalisable to other research questions.

Representing Semantic Networks

In addition to automatically extracting Semantic Networks, a second contribution of this thesis was showing how Knowledge Representation techniques can be used for formally representing Semantic Networks and associated background knowledge. In particular, the Resource Description Framework (RDF) and RDF Schema were used, with reification to describe the NET triples, adjunct instances to represent temporal roles, and query translation and custom reasoning to allow the analysis of the Semantic Network with simple but powerful queries. This shows how RDF and RDF tools can be used successfully for representing complex

information, and using that representation for solving the problem of constructing theoretic variables from Semantic Networks for social science research.

However, as already indicated by the need for custom reasoning and query translation, non-standard constructs had to be used in RDF to represent the desired information. By using these non-standard constructs, the semantics of the information is no longer explicit: A third party interpreter will not be able to understand the temporal roles used in our ontology, and, due to the problems with RDFS reification, most interpreters will not be able to leverage the NET graph described using reification. This implies that standard RDF tools for editing the temporal roles in our ontology or for visualising the Semantic Network cannot be easily used: As the tool does not understand our custom constructs, it can only show or edit the low-level RDF structure, rather than the implied semantics of the higher-level Semantic Networks.

These problems are caused by needing to represent statements about statements or metastatements: RDF triples are used to represent NET triples, and extra information needs to be added to those triples, such as the temporal validity of roles and quantitative aspects of NET triples⁴. This is a general problem in using RDF to describe RDF graphs, and ideally, it should have a general solution. This requires a good standard for RDF metastatements, allowing for representing the different semantics of different use cases, such as describing graphs and adding information to triples. If the RDF community can devise a standard for representing this information, either by extending the RDF standard or by using distinguished vocabulary, more aspects of Semantic Networks can be expressed with a standardised meaning. This makes it easier to combine data with third parties, and the tools for tasks such as editing an ontology with temporal roles, or visualising graphs described in RDF, can be shared and reused.

The road ahead: opportunities for communication analysis

In Semantic Network Analysis, networks of objects (such as actors and issues) are extracted from text either manually or automatically. By formally representing these networks and combining them with appropriate background knowledge, they can be queried to answer various research questions using different queries or object categorisations. This allows for the reuse and combination of data from different studies, and

⁴See section 8.2 on page 147

the sharing of data between different researchers or research groups even if the questions these researchers are interested in vary. Because the data sets from different studies can be combined, in effect, this creates a large, shared, distributed Semantic Network data repository. Creating such a repository will have great benefits for analysing communication.

To convince communication scholars to use Semantic Network Analysis and make their data available, this has to be made as easy and rewarding as possible. As described above, both iNet and the AmCAT navigator are open source and available free of charge, and mature enough to be immediately useful for conducting content analyses. The querying prototype presented earlier has to be merged into the AmCAT navigator to make the coding results easily accessible. This creates the infrastructure needed for easy and productive Semantic Network Analysis.

The first step towards creating a shared repository is already being taken at the Vrije Universiteit Amsterdam in cooperation with the University of Amsterdam. As described above, Semantic Network Analysis studies conducted previously at these universities have yielded a number of Semantic Network data sets. By merging these data sets into a single repository and by opening it up to the public, we can show the feasibility and merits of working towards such repositories. Additionally, the AmCAT implementation at the Vrije Universiteit needs to be converted into an open repository, where interested scholars can create new data sets and re-analyse existing data. This makes it very easy for an interested scholar or research group to start working with Semantic Network Analysis, as AmCAT provides a sophisticated infrastructure for coding and analysing data without needing initial investments in setup and hardware.

Creating large, shared Semantic Network data repositories produce a number of advantages. Substantively, such repositories are immediately interesting, because the size allows for testing models that are too complex to be tested on the individual data sets. Moreover, because the data will be from different time periods or events, the conclusions are easier to generalise.

If the repositories contain the original texts from which the networks were extracted, it can also be a great boost to the automatic extraction techniques described earlier, as these techniques can be improved by using more training data. Moreover, the fact that these repositories contain the networks from different time periods, domains, or even languages, makes it possible to determine the generalisability of the methods used for extraction.

The biggest advantage offered by large, shared repositories, however, is that it enables a new way of analysing communication. Concepts from research questions of theories can be operationalised as queries which

have clear semantics. This makes it easy to adopt, refine, and compare operationalisations, and duplicate research. Refined operationalisations can be immediately tested on the existing data, and existing operationalisations can be used easily on new data. Additionally, the existence of large, shared repositories allows new or refined theories to be tested immediately, without having to redo coding with an updated codebook. This makes it easier to conduct research by utilising scarce resources more efficiently.

Finally, because Semantic Network Analysis data allow different research questions to be answered on the same data, it allows for different theories or models to be tested simultaneously, in the spirit of the Multi-Level Multi-Theoretical framework for network communication studies (Monge and Contractor, 2003). Testing competing hypotheses can shed light on which hypothesis is better able to explain the data, and testing complementary hypotheses can give a better overall explanation of the data; both increase our knowledge of communication processes and effects in ways that studies that stay within a single theoretical framework cannot provide.

If the automatic extraction of Semantic Networks can be improved to make sophisticated analyses possible for a wider variety of research questions, domains, and languages, it will become easier to analyse large bodies of communication, making it possible to test more complex models of the interaction between the message and its social context. Additionally, it makes it possible to investigate current events in time to provide a useful contribution to the public debate. Moreover, the formal representation and querying of Semantic Networks makes it possible to reuse data sets for different research questions and to share and combine data sets between research groups. This creates the opportunity for communication analysts to create large, open repositories with transparent definitions of the data and the queries used for analysis. This can be a large benefit to the communication analysis community, by making it easier to compare, test, and refine theories and hypotheses. This thesis has brought these goals a step closer.

Bibliography

- Abney, S. (1997). Stochastic attribute-value grammars. *Computational Linguistics*, 23:597–618.
- Alexa, M. and Züll, C. (1999). A review of software for text analysis. ZUMA-Nachrichten Spezial Band 5. Mannheim: ZUMA.
- Althaus, S. L., Edy, J. A., and Phalen, P. F. (2001). Using substitutes for full-text news stories in content analysis. *American Journal of Political Science*, 45(3):707–724.
- Ansolabehere, S., Iyengar, S., Simon, A., and Valentino, N. (1994). Does attack advertising demobilize the electorate? *The American Political Science Review*, 88(4):829–838.
- Antoniou, G. and Van Harmelen, F. (2004). *A Semantic Web Primer*. MIT Press, Cambridge, Ma.
- Areces, C., Blackburn, P., and Marx, M. (1999). A road-map on complexity for hybrid logics. In *Proceedings of the Annual Conference of the European Association for Computer Science Logic (CSL-99)*, number 1683 in LNCS, pages 307–321. Springer, Berlin.
- Axelrod, R., editor (1976). *Structure of decision: The Cognitive Maps of political Elites*. Princeton University Press.
- Azar, E. E. (1980). The conflict and peace data bank (COPDAB) project. *The Journal of Conflict Resolution*, 24(1):143–152.
- Azar, E. E. (1982). The codebook of the conflict and peace data bank (COPDAB). College Park, MD: Center for International Development, University of Maryland.
- Baroni, M. and Vegnaduzzo, S. (2004). Identifying subjective adjectives through Web-based mutual information. In Buchberger, I. E., editor, *Proceedings of KONVENS 2004*, pages 17–24, Vienna. ÖGAI.
- Bartels, L. M. (1988). *Presidential primaries and the dynamics of public choice*. Princeton University Press, Princeton, NJ.
- Bennett, L. (1990). Toward a theory of press-state relations in the United States. *Journal of Communication*, 40(2):103–125.
- Bennett, L. and Paletz, D. L., editors (1994). *Taken by Storm : The Media, Public Opinion, and U.S. Foreign Policy in the Gulf War*. University of Chicago Press, Chicago.
- Benoit, W. L. (2007). *Communication in Political Campaigns*. Peter Lang, New York.
- Berelson, B. R. (1952). *Content analysis in Communication Research*. Free Press, Glencoe, IL.
- Berger, A. L., Della Pietra, S. A., and Della Pietra, V. J. (1996). A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

- Berghmans, J. (1994). Wotan, een automatische grammatikale tagger voor het Nederlands. Master's thesis, Dept. of Language and Speech, University of Nijmegen.
- Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The Semantic Web. *Scientific American*, 284(5):34–43.
- Bouma, G. (2003). Doing Dutch pronouns automatically in optimality theory. In *Proceedings of the EACL 2003 Workshop on The Computational Treatment of Anaphora*.
- Bouma, G., Mur, J., and van Noord, G. (2003). Reasoning over dependency relations for QA. In *IJCAI'05 workshop on Knowledge and Reasoning for Answering Questions*, Edinburgh, Scotland.
- Breck, E., Choi, Y., and Cardie, C. (2007). Identifying expressions of opinion in context. In Veloso, M. M., editor, *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI-07)*, Hyderabad, India.
- Brickley, D. and Guha, R. V. (2004). RDF vocabulary description language 1.0: RDF Schema. W3C Recommendation.
- Brill, E. (1995). Transformation-based error-driven learning and natural language processing: A case study in part of speech tagging. *Computational Linguistics*, 21:543–565.
- Broekstra, J. (2005). *Storage, Querying and Inferencing for Semantic Web Languages*. PhD thesis, Free University Amsterdam.
- Brosius, H.-B. and Eps, P. (1995). Prototyping through key events: News selection in the case of violence against aliens and asylum seekers in Germany. *European Journal of Communication*, 10(3):391–412.
- Brouwers, L. (1989). *Het juiste woord: Standaard betekeniswoordenboek der Nederlandse taal (7th edition; ed. F. Claes)*. Standaard Uitgeverij, Antwerpen, Belgium.
- Budge, I. and Farlie, D. J. (1983). *Explaining and Predicting elections: Issue-Effects and Party Strategies in Twenty-three democracies*. Allen and Unwin, London.
- Budge, I., Klingemann, H.-D., Volkens, A., Bara, J., and Tanenbaum, E. (2001). *Mapping Policy Preferences. Estimates for Parties, Electors, and Governments 1945-1998*. Oxford University Press, Oxford.
- Budge, I. and Laver, M. (1986). Office seeking and policy pursuit in coalition theory. *Legislative Studies Quarterly*, 11(4):485–506.
- Burnage, G. (1990). CELEX - a guide for users. Centre for Lexical Information, University of Nijmegen.
- Capella, J. N. and Jamieson, K. H. (1997). *Spiral of cynicism*. Oxford University Press, New York.
- Carey, J. W. (1989). *Communication as Culture*, volume I of *David Thorburn(ed.), Media and Popular Culture*. Unwin Hyman, Boston, MA.
- Carley, K. M. (1986). An approach for relating social structure to cognitive structure. *Journal of Mathematical Sociology*, 12(2):137–189.
- Carley, K. M. (1993). Coding choices for textual analysis: A comparison of content analysis and map analysis. *Sociological Methodology*, 23:75–126.
- Carley, K. M. (1997). Network text analysis: The network position of concepts. In Roberts, C., editor, *Text Analysis for the Social Sciences*, pages 79–100. Lawrence Erlbaum Associates, Mahwah, NJ.
- Carroll, J. J., Bizer, C., Hayes, P., and Stickler, P. (2005). Named graphs, provenance and trust. In *Proceedings of the Fourteenth International World Wide Web Conference (WWW2005)*, Chiba, Japan, volume 14, pages 613–622.
- Choi, Y., Breck, E., and Cardie, C. (2006). Joint extraction of entities and relations for opinion recognition. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, Sydney, Australia.
- Clausen, L. (2003). Global news communication strategies: 9.11.2002 around the world. *Nordicom Review*, 24(2):105–116.

- Clausen, L. (2004). Localizing the global: 'domestication' processes in international news production. *Media, Culture & Society*, 26(1):25–44.
- Cohen, B. C. (1963). *The press and foreign policy*. Princeton University Press, Princeton, NJ.
- Collins, A. M. and Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, 8:240–248.
- Cook, T. E. (1994). Domesticating a crisis : Washington newsbeats and network news after the Iraqi invasion of Kuwait. In Bennett, L. and Paletz, D. L., editors, *Taken by Storm : The Media, Public Opinion, and U.S. Foreign Policy in the Gulf War*. University of Chicago Press, Chicago.
- Corman, S. R., Kuhn, T., McPhee, R., and Dooley, K. (2002). Studying complex discursive systems: Centering resonance analysis of communication. *Human Communication Research*, 28(2):157–206.
- Daelemans, W., Zavrel, J., den Bosch, A. V., and der Sloot, K. V. (2007). MBT: Memory-Based Tagger, version 3.1, reference guide. Technical Report 07–08, ILK Technical Report Series.
- D'Angelo, P. (2002). News framing as a multi-paradigmatic research program: A response to Entman. *Journal of Communication*, 52(4):870–888.
- De Ridder, J. A. (1994). *Van tekst naar informatie: ontwikkeling en toetsing van een inhoudsanalyse-instrument*. PhD thesis, University of Amsterdam.
- Dearing, J. W. and Rogers, E. M. (1996). *Agenda setting*. Sage, Thousand Oaks, CA.
- Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. A. (1990). Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407.
- Diehl, P. F. (1992). What are they fighting for? The importance of issues in international conflict research. *Journal of Peace Research*, 29(3):333–344.
- Diesner, J. and Carley, K. M. (2004). Automap1.2 - extract, analyze, represent, and compare mental models from texts. Technical Report CMU-ISRI-04-100, Carnegie Mellon University, School of Computer Science, Institute for Software Research International, Pittsburgh, PA.
- Dille, B. and Young, M. D. (2000). The conceptual complexity of presidents Carter and Clinton: An automated content analysis of temporal stability and source bias. *Political Psychology*, 21(3):587–596.
- Dixon, R. M. W. (1991). *A New Approach to English Grammar, on Semantic Principles*. Oxford University Press, Oxford, England.
- Domke, D., Shah, D. V., and Wackman, D. B. (1998). Media priming effects: Accessibility, association and activation. *International Journal of Public Opinion Research*, 10:51–75.
- Donsbach, W. and Jandura, O. (2003). Chances and effects of authenticity: Candidates of the German federal election in TV news. *The Harvard International Journal of Press/Politics*, 49:49–65.
- Donsbach, W., Mattenklott, A., and Brosius, H.-B. (1993). How unique is the perspective of television? *Political Communication*, 10:41–57.
- Doreian, P., Batagelj, V., and Ferligoj, A. (2004). *Generalized Blockmodeling*. Cambridge University Press, Cambridge.
- Downs, A. (1957). *An Economic Theory of Democracy*. Harper Row, New York.
- Drenth, E. (1997). Using a hybrid approach towards Dutch part-of-speech. Master's thesis, Alfa-Informatica, University of Groningen, Groningen.
- Druckman, J. N. (2001). The implications of framing effects for citizen competence. *Political Behavior*, September 2001:225–256.
- Druckman, J. N. (2004). Political preference formation: Competition, deliberation, and the (ir)relevance of framing effects. *American Political Science Review*, 98(4):671–686.
- Dumbill, E. (2003). Tracking provenance of RDF data. Technical report, ISO/IEC.
- Eagley, A. H. and Chaiken, S. (1998). Attitude structure and function. In Gilbert, D. T.,

- Fiske, S. T., and Lindzey, G., editors, *The Handbook of Social Psychology*. McGraw-Hill, New York.
- Entman, R. M. (1991). Framing U.S. coverage of international news: Contrasts in narratives of the KAL and Iran Air incidents. *Journal of Communication*, 41(4):6–25.
- Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4):51–58.
- Fan, D. P. (1985). Ideodynamics — the kinetics of the evolution of ideas. *Journal of Mathematical Sociology*, 11(1):1–23.
- Fan, D. P. (1988). *Predictions of Public Opinion from the Mass Media*. Greenwood, Westport, CT.
- Fan, D. P. (1996). Predictions of the Bush-Clinton-Perot presidential race from the press. *Political Analysis*, 6:67–105.
- Fan, D. P., Brosius, H.-B., and Esser, F. (2001). Computer and human coding of German text on attacks on foreigners. In West, M. D., editor, *Applications of Computer Content Analysis*, volume 17 of *Progress in Communication Sciences*. Ablex Publishing, New York.
- Farnsworth, S. J. and Lichter, S. R. (2006). The 2004 New Hampshire Democratic primary. *Harvard international journal Press/Politics*, 11(1):53–63.
- Fellbaum, C., editor (1998). *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.
- Festinger, L. (1957). *A Theory of Cognitive Dissonance*. University Press, Stanford, CA.
- Fiske, S. T. and Linville, P. W. (1980). What does the schema concept buy us? *Personality and Social Psychology Bulletin*, 6:543–557.
- Fiske, S. T. and Taylor, S. E. (1991). *Social Cognition*, 2nd Ed. McGraw-Hill, New York.
- Franceschet, M. and de Rijke, M. (2005). Model checking for hybrid logics (with an application to semistructured data). *Journal of Applied Logic*, (in press).
- Galtung, J. and Ruge, M. H. (1965). The structure of foreign news: The presentation of the Congo, Cuba and Cyprus crises in four Norwegian newspapers. *Journal of Peace Research*, 2:64–91.
- Gamson, W. A. and Modigliani, A. (1987). The changing culture of affirmative action. In Braungart, R., editor, *Research in Political Sociology*, volume 3, pages 137–177. Jai Press, Inc, Greenwich, CT.
- Gaustad, T. and Bouma, G. (2001). Accurate stemming and email classification. In *Computational Linguistics in the Netherlands (CLIN)*.
- Goffman, E. (1974). *Frame analysis: an essay on the organization of experience*. Northeastern University Press, Boston.
- Graber, D. A. (1988). *Processing the News: How People Tame the Information Tide*. University Press of America, Lanham, MD.
- Grefenstette, G., Qu, Y., Evans, D. A., and Shanahan, J. G. (2006). Validating the coverage of lexical resources for affect analysis and automatically classifying new words along semantic axes. In Shanahan, J. G., Qu, Y., and Wiebe, J., editors, *Computing Attitude and Affect in Text: Theory and Applications*, pages 93–108. Springer, Dordrecht, the Netherlands.
- Grefenstette, G. and Tapanainen, P. (1994). What is a word, what is a sentence? Problems of tokenization. In *Proceedings of Third Conference on Computational Lexicography and Text Research (COMPLEX'94)*, pages 7–10, Budapest, Hungary.
- Grundmann, R., Smith, D., and Wright, S. (2000). National elites and transnational discourses in the Balkan war. *European Journal of Communication*, 15(3):299–320.
- Guarino, N. (1992). Concepts, attributes and arbitrary relations: Some linguistic and ontological criteria for structuring knowledge bases. *Data and Knowledge Engineering*, 8(2):249–261.
- Guha, R. V., McCool, R., and Fikes, R. (2004). Contexts for the Semantic Web. In *Proceedings of the Third International Conference on the Semantic Web (ISWC'04)*.

- Gurevitch, M., Levy, M. R., and Roeh, I. (1991). The global newsroom: Convergences and diversities in the globalization of television news. In Dahlgren, P. and Sparks, C., editors, *Communication and Citizenship: Journalism and the Public Sphere*. Routledge, London.
- Gutierrez, C., Hurtado, C., and Vaisman, A. (2005). Temporal RDF. In *ESWC 2005*, number 3532 in LNCS, pages 93–107, Berlin, Germany. Springer.
- Hagen, L. M. (1993). Opportune witnesses: An analysis of balance in the selection of sources and arguments. *European Journal of Communication*, 8:317–343.
- Hallin, D. C. (1986). *The Uncensored War: The Media and Vietnam*. University of California Press, Los Angeles.
- Harcup, T. and O'Neill, D. (2001). What is news? Galtung and Ruge revisited. *Journalism Studies*, 2(2):261–280.
- Harmel, R., Tan, A., and Janda, K. (1995). Substance v. packaging: An empirical analysis of parties' issue identity. In *paper delivered at the 1995 Annual Meeting of the American Political Science Association*, Chicago.
- Hart, R. P. (1985). Systematic analysis of political discourse: the developments of DICTION. In Sanders, K. R., Kaid, L. L., and Nimmo, D., editors, *Political Communication Yearbook 1984*. Southern Illinois University Press, Carbondale, IL.
- Hart, R. P. (2001). Redeveloping DICTION: Theoretical considerations. In West, M. D., editor, *Applications of Computer Content Analysis*, volume 17 of *Progress in Communication Sciences*. Ablex Publishing, New York.
- Hartman, K. (2000). Studies of negative political advertising: an annotated bibliography. *Reference Services Review*, 28(3):248–261.
- Hatzivassiloglou, V. and McKeown, K. (1997). Predicting the semantic orientation of adjectives. In Cohen, P. R. and Wahlster, W., editors, *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics (ACL-97)*, pages 174–181, Somerset, New Jersey. Association for Computational Linguistics.
- Hatzivassiloglou, V. and Wiebe, J. (2000). Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING-2000)*, pages 299–305, San Francisco, CA. Morgan Kaufmann.
- Heider, F. (1946). Attitudes and cognitive organization. *Journal of Psychology*, 21:107–112.
- Herbst, S. (2003). Political authority in a mediated age. *Theory and Society*, 32:481–503.
- Hetherington, M. J. (1996). The media's role in forming voters' national economic evaluations in 1992. *American Journal of Political Science*, 40(2):372–395.
- Hjarvard, S. (2001). News media and the globalization of the public sphere. In Hjarvard, S., editor, *News in a Globalized Society*. Nordicom, Goteborg.
- Holsti, O. R. (1964). An adaptation of the 'General Inquirer' for the systematic analysis of political documents. *Behavioral Science*, 9:332–338.
- Holsti, O. R. (1969). *Content Analysis for the Social Sciences and Humanities*. Addison-Wesley, Reading MA.
- Holsti, O. R., Brody, R. A., and North, R. C. (1964a). Theory and measurement of interstate behavior: A research application of automated content analysis. Stanford University: Studies in International Conflict and Integration, February 1964, multilith.
- Holsti, O. R., Brody, R. A., and North, R. C. (1964b). Violence and hostility: The path to World War. Stanford University: Studies in International Conflict and Integration, February 1964, multilith.
- Iker, H. and Harway, N. (1969). A computer system approach to the recognition and analysis of content. In Gerbner, G., Holsti, O. R., Krippendorff, K., Paisley, W., and Stone, P. J., editors, *The analysis of communication content: Developments in scientific theories and computer techniques*, pages 381–405. Wiley, New York.
- Iyengar, S. (1991). *Is anyone responsible? How television frames political issues*. University of

- Chicago Press, Chicago.
- Iyengar, S., Norpoth, H., and Hahn, K. S. (2003). Consumer demand for election news: The horse race sells. *Journal of Politics*, 66:157–175.
- Janis, I. and Fadner, R. (1943). The coefficient of imbalance. *Psychometrika*, 8(2):105–119.
- Jasperson, A. E. and Fan, D. P. (2004). The news as molder of campaign ad effects. *International Journal of Public Opinion Research*, 16(4):417–436.
- Jijkoun, V. (2007). *Graph Transformations for Natural Language Processing*. PhD thesis, University of Amsterdam.
- Jurafsky, D. and Martin, J. H. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall, Upper Saddle River, NJ.
- Kaid, L. L., Harville, B., Ballotti, J., and Wawrzyniak, M. (1993). Telling the Gulf War story: Coverage in five papers. In Greenberg, B. S. and Gantz, W., editors, *Desert Storm and the Mass Media*. Hampton, Cresskill, NJ.
- Kanamaru, T., Murata, M., and Isahara, H. (2007). Japanese opinion extraction system for Japanese newspapers using machine-learning method. In *Proceedings of NTCIR-6 Workshop Meeting*, Tokyo, Japan.
- Kaplan, N., Park, D. K., and Ridout, T. N. (2006). Dialogue in American political campaigns? An examination of issue convergence in candidate television advertising. *American Journal of Political Science*, 50:724–736.
- Karim, K. H. (2002). Making sense of the ‘Islamic peril’: Journalism as cultural practice. In Zelizer, B. and Allan, S., editors, *Journalism After September 11*. Routledge, London.
- Katz, B. and Lin, J. (2003). Selectively using relations to improve precision in question answering. In *Proceedings of the workshop on Natural Language Processing for Question Answering (EACL 2003)*, pages 43–50.
- Kim, S.-M. and Hovy, E. H. (2006). Extracting opinions expressed in online news media text with opinion holders and topics. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text at the joint COLING-ACL 2006 conference*, pages 1–8, Sydney, Australia.
- Kiousis, S., Mitrook, M., Wu, X., and Seltzer, T. (2006). First- and second-level agenda-building and agenda-setting effects. *Journal of Public Relations Research*, 18(3):265–285.
- Kirkpatrick, B. (1998). *Roget’s Thesaurus of English Words and Phrases*. Penguin, Harmondsworth, England.
- Kleinnijenhuis, J., De Ridder, J. A., and Rietberg, E. M. (1997). Reasoning in economic discourse: An application of the network approach to the Dutch press. In Roberts, C. W., editor, *Text Analysis for the Social Sciences; Methods for Drawing Statistical Inferences from Texts and Transcripts*, pages 191–207. Lawrence Erlbaum Associate, Mahwah, New Jersey.
- Kleinnijenhuis, J. and Fan, D. P. (1999). Media coverage and the flow of voters in multiparty systems: The 1994 national elections in Holland and Germany. *International Journal of Public Opinion Research*, 11(3):233–256.
- Kleinnijenhuis, J., Oegema, D., de Ridder, J. A., and Bos, H. (1995). *De democratie op drift: Een evaluatie van de verkiezingscampagne van 1994*. VU University Press, Amsterdam.
- Kleinnijenhuis, J., Oegema, D., de Ridder, J. A., and Ruigrok, N. (1998). *Paarse Polarisatie: De slag om de kiezer in de media*. Samson, Alphen a/d Rijn.
- Kleinnijenhuis, J., Oegema, D., de Ridder, J. A., Van Hoof, A. M. J., and Vliegienthart, R. (2003). *De puinhopen in het nieuws*, volume 22 of *Communicatie Dossier*. Kluwer, Alphen aan de Rijn (Netherlands).
- Kleinnijenhuis, J. and Pennings, P. (2001). Measurement of party positions on the basis of party programmes, media coverage and voter perceptions. In Laver, M., editor, *Estimating the Policy Positions of Political Actors*, pages 162–182. Routledge, London and New York.

- Kleinnijenhuis, J., Scholten, O., Van Atteveldt, W., Van Hoof, A. M. J., Krouwel, A., Oegema, D., De Ridder, J. A., Ruigrok, N., and Takens, J. (2007a). *Nederland vijfstromenland: De rol van media en stemwijzers bij de verkiezingen van 2006*. Bert Bakker, Amsterdam.
- Kleinnijenhuis, J., Takens, J., and Van Atteveldt, W. (2005). Toen Europa de dagbladen ging vullen. In Aarts, K. and van der Kolk, H., editors, *Nederlanders en Europa: het referendum over de Europese grondwet*, chapter 6. Bert Bakker, Amsterdam. [in Dutch].
- Kleinnijenhuis, J. and Van Hoof, A. M. J. (2008). Media coverage of government policies and public satisfaction with information provision and policy results. *To appear in Political Quarterly / Politische Vierteljahresschrift*.
- Kleinnijenhuis, J., Van Hoof, A. M. J., Oegema, D., and De Ridder, J. A. (2007b). A test of rivaling approaches 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.
- Knublauch, H., Fergerson, R., Noy, N. F., and Musen, M. A. (2004). The Protege OWL plugin: An open development environment for Semantic Web applications. In *Third International Conference on the Semantic Web (ISWC-2004)*, Hisroshima, Japan.
- Krackhardt, D. (1987). QAP partialling as a test of spuriousness. *Social Networks*, 9:171–186.
- Krippendorff, K. (2004). *Content Analysis: An Introduction to Its Methodology (second edition)*. Sage Publications, Thousand Oaks, CA.
- Landis, J. and Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174.
- Lappin, S. and Leass, H. (1994). An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–561.
- Lasswell, H. D. (1948). The structure and function of communication in society. In Bryson, L., editor, *The Communication of Ideas*. Harper, New York.
- Laver, M., Benoit, K., and Garry, J. (2003). Extracting policy positions from political texts using words as data. *American Political Science Review*, 97(2):311–331.
- Laver, M. and Garry, J. (2000). Estimating policy positions from political texts. *American Journal of Political Science*, 44(3):619–634.
- Lazarsfeld, P. F., Berelson, B. R., and Gaudet, H. (1944). *The people's choice: How the voter makes up his mind in a presidential campaign (3rd ed.)*. Duell, Sloan, and Pearce, New York.
- Lee, C. C., Chan, J. M., Pan, Z., and So, C. Y. K. (2002). *Global Media Spectacle: News War over Hong Kong*. State University of New York Press, New York.
- Lenat, D. B. and Guha, R. V. (1990). *Building large knowledge-based systems*. Addison-Wesley, Reading, MA.
- Levin, B. (1993). *English Verb Classes and Alternations*. University of Chicago Press, Chicago.
- Lewins, A. and Silver, C. (2007). *Using Software in Qualitative Research : A Step-by-Step Guide*. Sage Publications, London.
- Lewis-Beck, M. S. (2006). Does economics still matter? Econometrics and the vote. *Journal of Politics*, 68:208–212.
- Lijphart, A. (1975). *The politics of accommodation. Pluralism and democracy in the Netherlands (2nd edition, revised)*. University of California Press, Berkeley, CA.
- Lin, D. (1998). Automatic retrieval and clustering of similar words. In *Proceedings of COLING-ACL*, pages 768–774, Montreal, Canada.
- Lippmann, W. (1922). *Public Opinion*. Macmillan, New York, NY.
- MacGregor, R. and Ko, I.-Y. (2003). Representing contextualized data using Semantic Web tools. In *Practical and Scalable Semantic Web Systems (workshop at second ISWC)*.
- Manning, C. and Schütze, H. (2002). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, fifth printing edition.

- Marcus, M., Marcinkiewicz, M. A., and Santorini, B. (2004). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.
- Markoff, J., Shapiro, G., and Weitman, S. R. (1975). Toward the integration of content analysis and general methodology. *Sociological Methodology*, 6:1–58.
- Masolo, C., Vieu, L., Bottazzi, E., Catenacci, C., Ferrario, R., Gangemi, A., and Guarino, N. (2004). Social roles and their descriptions. In Dubois, D., Welty, C., and Williams, M. A., editors, *Proceedings of the Ninth International Conference on the Principles of Knowledge Representation and Reasoning (KR2004)*, pages 267–277, Whistler, Canada.
- Mathieu, Y. (2006). A computational semantic lexicon of French verbs of emotion. In Shanahan, J. G., Qu, Y., and Wiebe, J., editors, *Computing Attitude and Affect in Text: Theory and Applications*, pages 109–123. Springer, Dordrecht, the Netherlands.
- McClelland, C. A. (1976). World Event/Interaction Survey codebook. (ICPSR 5211) Ann Arbor: Inter University Consortium for Political and Social Research.
- McCombs, M. E. (2004). *Setting the Agenda: The Mass Media and Public Opinion*. Polity Press, Cambridge.
- McCombs, M. E. and Estrada, G. (1997). The news media and the pictures in our heads. In Iyengar, S. and Reeves, R., editors, *Do the media govern?*, pages 237–247. Sage, London.
- McCombs, M. E. and Ghanem, S. (2001). The convergence of agenda setting and framing. In Reese, S. D., Gandy, O. H., and Grant, A. E., editors, *Framing public life*, pages 95–106. Lawrence Erlbaum, Mahwah, NJ.
- McCombs, M. E., Lopez-Escobar, E., and Llamas, J. P. (2000). Setting the agenda of attributes in the 1996 Spanish general election. *Journal of Communication*, 50(2):77–92.
- McCombs, M. E. and Shaw, D. L. (1972). The agenda-setting function of mass media. *Public Opinion Quarterly*, 36:176–187.
- McLuhan, M. (1960). *Understanding Media: The Extension of Man*. McGraw-Hill, New York.
- Meijer, M. and Kleinnijenhuis, J. (2006). News and corporate reputation: Empirical findings from the Netherlands. *Public Relations Review*, 32(4):341–348.
- Mihalcea, R., Banea, C., and Wiebe, J. (2007). Learning multilingual subjective language via cross-lingual projections. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL'07)*, pages 968–975, Prague, Czech Republic.
- Mika, P. and Gangemi, A. (2004). Descriptions of Social Relations. In *Proceedings of the 1st Workshop on Friend of a Friend, Social Networking and the (Semantic) Web*.
- Miles, A. and Bechhofer, S. (2008). SKOS Simple Knowledge Organization System reference. W3C Working Draft.
- Miles, M. and Huberman, A. (1994). *Qualitative Data Analysis: an Expanded Sourcebook*. Sage, London.
- Miller, G. (1990). Wordnet: An on-line lexical database. *International Journal of Lexicography (Special Issue)*, 3:235–312.
- Miller, G. (1995). *WordNet: a lexical database for English*. ACM Press, New York.
- Minsky, M. A. (1975). A framework for representing knowledge. In Winston, P. H., editor, *The Psychology of Computer Vision*. McGraw-Hill, New York.
- Mitkov, R. (2002). *Anaphora resolution*. Longman, Harlow, UK.
- Mohler, P. P. and Züll, C. (1998). TEXTPACK User's Guide. ZUMA, Mannheim.
- Monge, P. R. and Contractor, N. S. (2003). *Theories of Communication Networks*. Oxford University Press, Oxford, England.
- Mur, J. and van der Plas, L. (2006). Anaphora resolution for off-line answer extraction using instances. In *Proceedings of the Workshop for Anaphora Resolution (WAR)*.
- Nacos, B. L. and Torres-Reyna, O. (2003). Framing Muslim-Americans before and after 9/11. In Norris, P., Just, M. R., and Kern, M., editors, *Framing Terrorism: Understanding Terrorist Threats and Mass Media*. Routledge, New York.
- Namenwirth, J. Z. and Weber, R. P. (1987). *Dynamics of Culture*. Allen & Unwin, Winchester MA.

- Nelson, T. E., Oxley, Z., and Clawson, R. A. (1997). Toward a psychology of framing effects. *Political Behavior*, 19(3):221–46.
- Neuendorf, K. A. (2002). *The Content Analysis Guidebook*. Sage, Thousand Oaks, CA.
- Neuman, W. R., Just, M. R., and Crigler, A. N. (1992). *Common Knowledge*. University of Chicago Press, Chicago.
- Norris, P. (1995). Rhe restless search: Network news framing of the post-Cold-War world. *Political Communication*, 12:357–370.
- Norris, P., Just, M. R., and Kern, M. (2003). *Framing Terrorism: Understanding Terrorist Threats and Mass Media*. Routledge, New York.
- Noy, N. F. (2005). Representing classes as property values on the Semantic Web. W3C Working Group Note 5 April 2005, <http://www.w3.org/TR/swbp-classes-as-values/>.
- Noy, N. F. and Rector, A. (2005). Defining N-ary relations on the Semantic Web. Working Draft, W3C Semantic Web best practices group.
- NWO (2006). Omstreden democratie, onderzoeksprogramma 2006-2011. Nederlandse Organisatie voor Wetenschappelijk Onderzoek (www.nwo.nl/omstredendemocratie).
- O’Heffernan, P. (1991). *Mass Media and American Foreign Policy: Insider Perspectives on Global Journalism and the Foreign Policy Process*. Ablex, Nordwood, NJ.
- Olien, C., Donohue, G., and Tichenor, P. (1983). Structure, communication and social power: Evolution of the knowledge gap hypothesis. *Mass communication review yearbook*, 4:455–461.
- Osgood, C. E. (1959). The representational model and relevant research methods. In de Sola Pool, I., editor, *Trends in Content Analysis*, pages 33–88. University of Illinois Press, Urbana, IL.
- Osgood, C. E., Saporta, S., and Nunnally, J. C. (1956). Evaluative Assertion Analysis. *Litera*, 3:47–102.
- Osgood, C. E., Suci, G. J., and Tannenbaum, P. H. (1967). *The Measurement of Meaning*. University of Illinois press, Urbana IL.
- Patterson, T. E. (1993). *Out of Order*. Knopf, New York.
- Pennebaker, J. W. and Francis, M. E. (1996). Cognitive, emotional, and language processes in disclosure. *Cognition and Emotion*, 10:601–626.
- Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). *Linguistic Inquiry and Word Count*. Lawrence Erlbaum Associates, Mahwah, NJ.
- Persson, V. (2004). Framing mediated terrorism before and after 9/11. C-Paper, http://www.statsvet.su.se/mediarum/Media_and_Politics_2/PDF/C-papers/framing_mediated_terrorism.pdf.
- Petrocik, J. R. (1996). Issue ownership in presidential elections, with a 1980 case study. *American Journal of Political Science*, 40:825–50.
- Petrocik, J. R., Benoit, W. L., and Hansen, G. J. (2003). Issue ownership and presidential campaigning, 1952–2000. *Political Science Quarterly*, 118:599–626.
- Pit, M. (2003). *How to Express Yourself with a Causal Connective: Subjectivity and Causal Connectives in Dutch, German and French*. Rodopi, Amsterdam.
- Pleijter, A. (2006). *Typen en logica van kwalitatieve inhoudsanalyse in de communicatie-wetenschap (dissertation)*. Tandem Felix, Ubbergen, the Netherlands.
- Popping, R. (2000). *Computer-assisted Text Analysis*. Sage, Newbury Park / London.
- Price, V., Tewksbury, D., and Power, E. (1997). Switching trains of thought: The impact of news frames on readers’ cognitive responses. *Communication Research*, 24:481–506.
- Prud’hommeaux, E. and Seaborne, A. (2006). SPARQL query language for RDF. W3C Recommendation.
- Quattrone, G. A. and Tversky, A. (1988). Contrasting rational and psychological analyses of political choice. *American Political Science Review*, 82:719–736.

- Rabinowitz, G. and MacDonald, S. E. (1989). A directional theory of issue voting. *American Political Science Review*, 83:93–122.
- Ratnaparkhi, A. (1998). *Maximum entropy models for natural language ambiguity resolution*. PhD thesis, University of Pennsylvania, Philadelphia.
- Reynar, J. C. and Ratnaparkhi, A. (1997). A maximum entropy approach to identifying sentence boundaries. In *Proceedings of the ANLP97*, Washington, D.C.
- Reynolds, A. and Barnett, B. (2003). America under attack: CNN's verbal and visual framing of September 11. In Chermak, S., Bailey, F. Y., and Brown, M., editors, *In Media Representations of September 11*. Praeger, New York.
- Rhee, J. W. (1997). Strategy and issue frames in election campaign coverage: A social cognitive account of framing effects. *Journal of Communication*, 47:26–48.
- Richardson, J. E. (2001). British Muslims in the broadsheet press: A challenge to cultural hegemony? *Journalism Studies*, 2(2):221–42.
- Riloff, E. and Wiebe, J. (2003). Learning extraction patterns for subjective expressions. In *Conference on Empirical Methods in Natural Language Processing (EMNLP-03)*, ACL SIGDAT, pages 105–112, Sapporo, Japan.
- Ritzer, G. (2004). *The Globalization of Nothing*. Sage, Thousand Oaks, CA.
- Roberts, C. W. (1989). Other than counting words: A linguistics approach to content analysis. *Social Forces*, 68(1):147–177.
- Roberts, C. W., editor (1997). *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts*. Lawrence Erlbaum, Mahwah, NJ.
- Robertson, R. (1995). Glocalization: Time-space and homogeneity-heterogeneity. In Featherstone, M., Lash, S., and Robertson, R., editors, *Global Modernities*. Sage, London.
- Rosenberg, S. D., Schnurr, P. P., and Oxman, T. E. (1990). Content analysis: A comparison of manual and computerized systems. *Journal of Personality Assessment*, 54(1&2):293–310.
- Ruigrok, N. (2005). *Journalism of Attachment : Dutch newspapers during the Bosnian war*. Het Spinhuis Publishers, Amsterdam.
- Ruigrok, N. and Van Atteveldt, W. (2007). Global angling with a local angle: How U.S., British, and Dutch newspapers frame global and local terrorist attacks. *The Harvard International Journal of Press/Politics*, 12:68–90.
- Ryan, M. (2004). Framing the war against terrorism: U.S. newspaper editorials and military action in Afghanistan. *Gazette*, 66(5):363–82.
- Said, E. W. (1981). *Covering Islam: How the Media and the Experts Determine How We See the Rest of the World*. Pantheon, New York.
- Schaefer, T. M. (2003). Framing the US embassy bombings and September 11 attacks in African and US newspapers. In Norris, P., Just, M. R., and Kern, M., editors, *Framing Terrorism: Understanding Terrorist Threats and Mass Media*. Routledge, New York.
- Scheufele, D. A. (1999). Framing as a theory of media effects. *Journal of Communication*, 29:103–123.
- Scheufele, D. A. (2000). Agenda-setting, priming, and framing revisited: Another look at cognitive effects of political communication. *Mass Communication & Society*, 3 (2-3):297–316.
- Schoenbach, K. (1982). The issues of the seventies: Computerunterstuetzte inhaltsanalyse und die langfristige beobachtung von agenda-setting-wirkungen der massenmedien [the issues of the seventies: Electronic content analysis and the long range observation of agenda-setting effects of the mass media]. *Publizistik*, 27:129–140.
- Scholten, O., Vasterman, P., Ruigrok, N., Takens, J., and Prins, J. (2007). Media en Mabel: Een onderzoek naar de berichtgeving in vijf landelijke dagbladen over de affaire Mabel Wisse Smit. Report of the first Event Monitor of the Netherlands News Monitor.
- Schrodt, P. A. (2001). Automated coding of international event data using sparse parsing techniques. In *Annual meeting of the International Studies Association*, Chicago.

- Schrodt, P. A. and Gerner, D. J. (1994). Validity assessment of a machine-coded event data set for the Middle East, 1982–1992. *American Journal of Political Science*, 38(3):825–854.
- Schrodt, P. A., Gerner, D. J., and Yilmaz, O. (2005). Using event data to monitor contemporary conflict in the israel-palestine dyad. *International Studies Perspectives*, 6(2):235–251.
- Searle, J. (1969). *Speech Acts*. Cambridge University Press, Cambridge, UK.
- Seki, Y., Evans, D. K., Ku, L.-W., Chen, H.-H., Kando, N., and Li, C.-Y. (2007). Overview of opinion analysis pilot task at NTCIR-6. In *Proceedings of NTCIR-6 Workshop Meeting*, Tokyo, Japan.
- Semetko, H. A. and Valkenburg, P. M. (2000). Framing European politics: A content analysis of press and television news. *Journal of Communication*, 50 (2):93–109.
- Shah, D. V., Watts, M. D., Domke, D., and Fan, D. P. (2002). News framing and cueing of issue regimes: explaining Clinton's public approval in spite of scandal. *Public Opinion Quarterly*, 66(3):339–370.
- Shah, D. V., Watts, M. D., Domke, D., Fan, D. P., and Fibison, M. (1999). Television news, real world cues, and changes in the public agenda 1984–1996. *Journal of Politics*, 61:914–943.
- Shaheen, J. (1981). Images of Saudis and Palestinians: A review of major documentaries. In Adams, W. C., editor, *Television Coverage of the Middle East*. Ablex, Norwood, NJ.
- Shanahan, J. G., Qu, Y., and Wiebe, J., editors (2006). *Computing Attitude and Affect in Text: Theory and Applications*. Springer, Dordrecht, the Netherlands.
- Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423.
- Shannon, C. and Weaver, W. (1948). *A Mathematical theory of communication*. Univ. of Illinois Press.
- Sigelman, L. and Buell, E. (2004). Avoidance or engagement? Issue convergence in presidential campaigns. *American Journal of Political Science*, 48:650–661.
- Simmons, B. K. and Lowry, D. N. (1990). Terrorists in the news, as reflected in three news magazines, 1980–1988. *Journalism Quarterly*, 67:692–96.
- Simon, H. A. (1954). Bandwagon and underdog effects and the possibility of election predictions. *Public Opinion Quarterly*, 18:245–253.
- Sintek, M. and Decker, S. (2002). Triple — A query, inference, and transformation language for the Semantic Web. In *Proceedings of ISWC02*.
- Snow, D. A. and Benford, R. D. (1988). Ideology, frame resonance, and participant mobilization. *International Social Movement Research*, 1:197–217.
- Snow, D. A., Rochford, E. B., Worden, S. K., and Benford, R. D. (1986). Frame alignment processes, micromobilization, and movement participation. *American Sociological Review*, 51:464–481.
- Soroka, S. N. (2006). Good news and bad news: Asymmetric responses to economic information. *Journal of Politics*, 68(2):372–385.
- Sowa, J. F. (1988). Using a lexicon of canonical graphs in a semantic interpreter. In Evens, M. W., editor, *Relational models of the lexicon*. Cambridge University Press, Cambridge UK.
- Sowa, J. F. (2000). *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Brooks/Cole, Pacific Grove, CA.
- Speed, G. (1893). Do newspapers now give the news? *Forum*, 15:705–711.
- Steimann, F. (2000). On the representation of roles in object-oriented and conceptual modelling. *Data and Knowledge Engineering*, 35:83–106.
- Stempel, G. and Culbertson, H. (1984). The prominence and dominance of news sources in newspaper medical coverage. *Journalism Quarterly*, 61:671–676.
- Stone, P. J., Bayles, R. F., Namerwirth, J. Z., and Ogilvie, D. M. (1962). The General Inquirer: A computer system for content analysis and retrieval based on the sentence

- as a unit of information. *Behavioral Science*, 7.
- Stone, P. J., Dunphy, D. C., Smith, M. S., Ogilvie, D. M., and Associates (1966). *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press, Cambridge, MA.
- Tjong Kim Sang, E. and Hofmann, K. (2007). Automatic extraction of Dutch hypernym-hyponym pairs. In *Proceedings of CLIN-06*, Leuven, Belgium.
- Tsvetov, M., Reminga, J., and Carley, K. M. (2003). DyNetML: Interchange format for rich social network data. In *NAACSOS Conference 2003*.
- Tversky, A. (1977). Features of similarity. *Psychological Review*, 84(4):327–352.
- Tversky, A. and Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211:453–458.
- Valkenburg, P. M., Semetko, H. A., and De Vreese, C. H. (1999). The effects of news frames on readers' thoughts and recall. *Communication Research*, 26(5):550–569.
- Van Atteveldt, W., Kleinnijenhuis, J., and Ruigrok, N. (2008a). Parsing, semantic networks, and political authority: Using syntactic analysis to extract semantic relations from Dutch newspaper articles. *Accepted for publication in B. Munroe and P. Schrodt (eds.), Special Issue of Political Analysis on Automated Natural Language Processing and Statistical Analysis of Text in Political Research*.
- Van Atteveldt, W., Kleinnijenhuis, J., Ruigrok, N., and Schlobach, S. (2008b). Good news or bad news? Conducting sentiment analysis on Dutch text to distinguish between positive and negative relations. *C. Cardie and J. Wilkerson (eds.), special issue of the Journal of Information Technology and Politics on Text Annotation for Political Science Research*, 5(1):73–94.
- Van Atteveldt, W., Ruigrok, N., Schlobach, S., and Van Harmelen, F. (2008c). Searching the news: Using a rich ontology with time-bound roles to search through annotated newspaper archives. In *Proceedings of the 58th annual conference of the International Communication Association (to appear)*, Montreal.
- Van Atteveldt, W. and Schlobach, S. (2005). A modal view on polder politics. In *Proceedings of Methods for Modalities (M4M) 2005 (Berlin, 1-2 December)*.
- Van Atteveldt, W., Schlobach, S., and Van Harmelen, F. (2007). Media, politics and the Semantic Web: An experience report in advanced RDF usage. In Franconi, E., Kifer, M., and May, W., editors, *ESWC 2007*, number 4519 in LNCS, pages 205–219, Berlin, Germany. Springer.
- Van Belle, D. A. (2000). New York Times and network TV news coverage of foreign disasters: The significance of the insignificant variables. *Journalism & Mass Communication Quarterly*, 77:50–70.
- Van Cuilenburg, J. J., Kleinnijenhuis, J., and De Ridder, J. A. (1986). Towards a graph theory of journalistic texts. *European Journal of Communication*, 1:65–96.
- Van der Wouden, T., Moortgat, M., Schuurman, I., and Renmans, B. (2002). Syntactische annotatie voor het Corpus Gesproken Nederlands (CGN). *Nederlandse Taalkunde*, 7(4):335–352.
- Van Noije, L. L. J. (2007). *The Democratic Deficit Closer to Home*. PhD thesis, Vrije Universiteit Amsterdam.
- Van Noord, G. (2006). At last parsing is now operational. In Mertens, P., Fairon, C., Dister, A., and Watrin, P., editors, *Verbum Ex Machina, Actes de la 13e conference sur le traitement automatique des langues naturelles*, pages 20–42, Louvain-la-Neuve, Belgium. Presses Universitaires de Louvain.
- Volkmer, I. (2002). Journalism and political crises. In Stuart, A. Z. and London, B., editors, *Journalism after September 11*. Routledge, New York.
- Wasserman, S. and Faust, K. (1994). *Social Network Analysis*. Cambridge University Press, Cambridge, England.
- Watts, M. D., Domke, D., Shah, D. V., and Fan, D. P. (1999). Elite cues and media bias in presidential campaigns: Explaining public perceptions of a liberal press. *Communication Research*, 26(2):144–175.

- Weaver, D. H. and Wilhoit, G. C. (1980). News media coverage of U.S. senators in four congresses, 1953–1974. *Journalism Monographs*, 67:1–34.
- Westholm, A. (1997). Distance versus direction: The illusory defeat of the proximity theory of electoral choice. *American Political Science Review*, 91:865–883.
- Wiebe, J. (2000). Learning subjective adjectives from corpora. In *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-00)*, Austin, TX.
- Wiebe, J., Wilson, T., Bruce, R. F., Bell, M., and Martin, M. (2004). Learning subjective language. *Computational Linguistics*, 30(3):277–308.
- Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing HLT-EMNLP-2005*, Vancouver, Canada.
- Winn, G. W. C. (1994). *The Theatre of Terror: Mass Media and International Terrorism*. Longman, White Plains, NY.
- Woelfel, J. (1993). Artificial neural networks for policy research. *Journal of Communication*, 43(1):63–80.
- Wong, R. K., Chau, H. L., and Lochovsky, F. H. (1997). A data model and semantics of objects with dynamic roles. In *Proceedings of IEEE Data Engineering Conference*, pages 402–411.
- Xu, R., Wong, K.-F., and Xia, Y. (2007). Opinmine: Opinion analysis system by for NTCIR-6 pilot task. In *Proceedings of NTCIR-6 Workshop Meeting*, Tokyo, Japan.
- Zaller, J. R. (1992). *The Nature and Origins of Mass Opinion*. Cambridge University Press, New York.
- Zaller, J. R. (1994). Elite leadership of mass opinion: New evidence from the Gulf War. In Bennett, L. and Paletz, D. L., editors, *Taken by Storm : The Media, Public Opinion, and U.S. Foreign Policy in the Gulf War*. University of Chicago Press, Chicago.
- Zhang, M., Zhou, G., and Aw, A. (2008). Exploring syntactic structured features over parse trees for relation extraction using kernel methods. *Information Processing and Management*, 44(2):687–701.

Samenvatting (Dutch Summary)

In onze democratie spelen de media een belangrijke rol als podium en recensent voor het politiek schouwspel. De meeste informatie over politieke actoren en thema's komen de kiezers via de media te weten. Het is dus van groot belang om te begrijpen hoe de media functioneren en welke invloed zij — bedoeld of onbedoeld — uitoefenen op de gedachten en handelingen van het publiek.

In de sociale wetenschappen wordt het onderzoeken van de inhoud van communicatieboodschappen, zoals televisie-uitzendingen of krantenartikelen, *inhoudsanalyse* genoemd. In de kwantitatieve stroming bestaat deze inhoudsanalyse doorgaans uit het bepalen van de frequentie van interessante variabelen, bijvoorbeeld hoe vaak de verschillende politici worden genoemd, of hoe vaak de berichtgeving op een bepaalde manier *geframed* wordt.

Alhoewel met deze handmatige en 'thematische' inhoudsanalyse succesvol onderzoek is gedaan, kleven er ook een aantal nadelen aan. Handmatige inhoudsanalyse is arbeidsintensief, en daardoor duur en tijdrovend. Een ander nadeel is dat de verzamelde gegevens vaak zeer nauw aansluiten op de onderzoeksvraag. Dit maakt het moeilijk om gegevens opnieuw te gebruiken: als bepaalde thema's of frames op een bepaalde manier geteld zijn, is het niet mogelijk deze tellingen te gebruiken voor een andere of zelfs een enigszins gewijzigde onderzoeksvraag.

Zoals beschreven in hoofdstuk 2 is een alternatieve methode voor inhoudsanalyse de semantische netwerkanalyse. In deze methode wordt niet volstaan met het tellen van bepaalde thema's, maar wordt elke boodschap ontleed in een semantisch netwerk van relaties tussen de relevante actoren en onderwerpen: wie steunt wie, welke standpunten worden in-

genomen, gaat het goed met de actoren en hoe ontwikkelen de issues zich? Aan de hand van deze semantische netwerken wordt vervolgens de oorspronkelijke onderzoeksvraag beantwoord, bijvoorbeeld hoe vaak het nieuws op een bepaalde manier *geframed* wordt.

Door deze loskoppeling van tekstanalyse en onderzoeksvraag, met het semantisch netwerk als 'halffabriek', is het mogelijk om verschillende onderzoeksvragen te beantwoorden met dezelfde gegevens, of gegevens die in verschillende onderzoeken zijn verzameld te combineren tot grotere databestanden. Dit biedt een efficiëntievoordeel en maakt het makkelijker om grote databestanden te ontwikkelen en verschillende theorieën direct te vergelijken op basis van deze bestanden.

Ook semantische netwerkanalyse heeft echter nadelen. Ten eerste is het handmatig extraheren van de netwerken uit teksten een nog grotere inspanning dan handmatige thematische analyse. Daarnaast is het in de praktijk vaak lastig om bestaande netwerken te koppelen, omdat de actoren en onderwerpen in die netwerken vaak niet overeenkomen: nieuwe issues komen op, politici veranderen van rol en partij, nieuwe partijen worden gesticht en ontbonden. Tenslotte is het analyseren van semantische netwerken vaak een complexe bezigheid, wat het moeilijker maakt om deze methode in te zetten.

De drie inhoudelijke delen van dit proefschrift beschrijven een aantal methoden en technieken die zijn ontwikkeld om deze nadelen van semantische netwerkanalyse te verminderen. Deze technieken zijn ruwweg onder te verdelen in extractie (deel II), representatie en bevraging (deel III), en systeembeschrijving (deel IV).

Voor de *extractie* van semantische netwerken zijn een aantal technieken uit de computerlinguïstiek gebruikt om het automatisch extraheren van netwerken te vergemakkelijken. Hoofdstuk 5 kijkt naar hoe het samen voorkomen (co-occurentie) van actoren en onderwerpen kan worden gebruikt als een eerste 'associatief' semantisch netwerk. Dit wordt geïllustreerd met een toegepast onderzoek naar de berichtgeving over Islam en terrorisme. Dit toont aan dat op een relatief simpele manier associatieve netwerken geëxtraheerd kunnen worden uit grote verzamelingen teksten, en dat die netwerken interessante inzichten kunnen opleveren over die teksten.

In hoofdstuk 6 wordt dit verrijkt door de grammaticale analyse van zinnen te gebruiken om een onderscheid te maken tussen bron, handelende actor, en lijdend voorwerp. Een *case study* laat zien dat deze informatie gebruikt kan worden om de media-autoriteit van actoren te onderzoeken. De betrouwbaarheid van deze methode wordt gemeten door op zowel meet- als analyseniveau een vergelijking te maken tussen de automatische analyse en een eerdere handmatige analyse. Hieruit blijkt dat, alhoewel de techniek zeker nog verbeterd kan worden, de betrouw-

baarheid op analyseniveau hoog genoeg is voor onmiddellijk gebruik in de sociale wetenschap.

Ten slotte wordt in hoofdstuk 7 door gebruik van *machine learning* op basis van bestaande gegevens de lading (positief – negatief) van de relatie bepaald. Deze methode blijkt de handmatige analyses redelijk te benaderen en werkt significant beter dan een simpelere ‘baseline’ methode. Ook in een aantal concrete *case studies* heeft de methode een goede correlatie met de uitkomsten van een handmatige analyse, en in de meeste gevallen is de correlatie hoog genoeg om op de resultaten te kunnen vertrouwen voor automatische inhoudsanalyse.

Naast de extractie kijkt dit proefschrift naar het *representeren en bevragen* van de geëxtraheerde semantische netwerken. Hoofdstuk 8 beschrijft hoe de taal RDF van het Semantische Web gebruikt kan worden om zowel het semantisch netwerk zelf te representeren als de benodigde achtergrondkennis om dit netwerk te analyseren. In die achtergrondkennis staat bijvoorbeeld van politici wanneer ze lid waren van welke partij en welke functies zij vervulden. Van de onderwerpen is opgenomen tot welke hoofdonderwerpen ze behoren.

Hoofdstuk 9 beschrijft een relatief simpele ‘querytaal’ waarmee onderzoekers of geïnteresseerden het semantisch netwerk kunnen bevragen om bepaalde patronen te zoeken. De resultaten hiervan kunnen getoond en gevisualiseerd worden op zowel geaggregeerd niveau als op het niveau van de oorspronkelijke artikelen waarin het patroon voorkwam. De geaggregeerde gegevens kunnen weer gebruikt worden voor verdere kwantitatieve analyse.

Het laatste deel van dit proefschrift, hoofdstuk 10, bevat de *systeembeschrijving* van de infrastructuur die is ontwikkeld als deel van dit onderzoek. *AmCAT*, de Amsterdam Content Analysis Toolkit, is een systeem om documenten zoals krantenartikelen op te slaan, te doorzoeken, en te prepareren voor analyse. *iNet* is een programma voor handmatige semantische netwerkanalyse dat gekoppeld is aan het AmCAT systeem om het beheer van grootschalige codeeroperaties te vergemakkelijken en om te zorgen dat de coderingen gekoppeld zijn aan de documenten en ontologie in AmCAT.

De verschillende delen van dit proefschrift beslaan een breed scala aan onderwerpen: extractie van semantische netwerken met natuurlijke-taalverwerking, representatie en analyse van deze netwerken met technieken uit de kennisrepresentatie, en een concreet systeem voor grootschalige handmatige en automatische inhoudsanalyse. Samen genomen betekent dit een belangrijke stap voorwaarts voor semantische netwerkanalyse, wat het makkelijker maakt om deze techniek in te zetten in de communicatiewetenschap.

SIKS Dissertation Series

| 1998 | | 1999-8 | Jacques H.J. Lenting (UM) <i>Informed Gambling: Conception and Analysis of a Multi-Agent Mechanism for Discrete Reallocation.</i> |
|--------|---|---------|---|
| 1998 | | 2000 | |
| 1998-1 | Johan van den Akker (CWI) <i>DEGAS - An Active, Temporal Database of Autonomous Objects</i> | 2000-1 | Frank Niessink (VU) <i>Perspectives on Improving Software Maintenance</i> |
| 1998-2 | Floris Wiesman (UM) <i>Information Retrieval by Graphically Browsing Meta-Information</i> | 2000-2 | Koen Holtman (TUE) <i>Prototyping of CMS Storage Management</i> |
| 1998-3 | Ans Steuten (TUD) <i>A Contribution to the Linguistic Analysis of Business Conversations within the Language/Action Perspective</i> | 2000-3 | Carolien M.T. Metselaar (UVA) <i>Sociaal-organisatorische gevolgen van kennis-technologie; een procesbenadering en actorperspectief.</i> |
| 1998-4 | Dennis Breuker (UM) <i>Memory versus Search in Games</i> | 2000-4 | Geert de Haan (VU) <i>ETAG, A Formal Model of Competence Knowledge for User Interface Design</i> |
| 1998-5 | E.W.Oskamp (RUL) <i>Computerondersteuning bij Straftoemeting</i> | 2000-5 | Ruud van der Pol (UM) <i>Knowledge-based Query Formulation in Information Retrieval.</i> |
| 1998 | | 2000-6 | Rogier van Eijk (UU) <i>Programming Languages for Agent Communication</i> |
| 1999-1 | Mark Sloof (VU) <i>Physiology of Quality Change Modelling; Automated modelling of Quality Change of Agricultural Products</i> | 2000-7 | Niels Peek (UU) <i>Decision-theoretic Planning of Clinical Patient Management</i> |
| 1999-2 | Rob Potharst (EUR) <i>Classification using decision trees and neural nets</i> | 2000-8 | Veerle Coup (EUR) <i>Sensitivity Analysis of Decision-Theoretic Networks</i> |
| 1999-3 | Don Beal (UM) <i>The Nature of Minimax Search</i> | 2000-9 | Florian Waas (CWI) <i>Principles of Probabilistic Query Optimization</i> |
| 1999-4 | Jacques Penders (UM) <i>The practical Art of Moving Physical Objects</i> | 2000-10 | Niels Nes (CWI) <i>Image Database Management System Design Considerations, Algorithms and Architecture</i> |
| 1999-5 | Aldo de Moor (KUB) <i>Empowering Communities: A Method for the Legitimate User-Driven Specification of Network Information Systems</i> | 2000-11 | Jonas Karlsson (CWI) <i>Scalable Distributed Data Structures for Database Management</i> |
| 1999-6 | Niek J.E. Wijngaards (VU) <i>Re-design of compositional systems</i> | | |
| 1999-7 | David Spelt (UT) <i>Verification support for object database design</i> | | |

| 2001 | | 2002-08 | |
|---------|--|---------|---|
| 2001-1 | Silja Renooij (UU) <i>Qualitative Approaches to Quantifying Probabilistic Networks</i> | 2002-09 | Willem-Jan van den Heuvel(KUB) <i>Integrating Modern Business Applications with Objectified Legacy Systems</i> |
| 2001-2 | Koen Hindriks (UU) <i>Agent Programming Languages: Programming with Mental Models</i> | 2002-10 | Brian Sheppard (UM) <i>Towards Perfect Play of Scrabble</i> |
| 2001-3 | Maarten van Someren (UvA) <i>Learning as problem solving</i> | 2002-11 | Wouter C.A. Wijngaards (VU) <i>Agent Based Modelling of Dynamics: Biological and Organisational Applications</i> |
| 2001-4 | Evgueni Smirnov (UM) <i>Conjunctive and Disjunctive Version Spaces with Instance-Based Boundary Sets</i> | 2002-12 | Albrecht Schmidt (Uva) <i>Processing XML in Database Systems</i> |
| 2001-5 | Jacco van Ossenbruggen (VU) <i>Processing Structured Hypermedia: A Matter of Style</i> | 2002-13 | Hongjing Wu (TUE) <i>A Reference Architecture for Adaptive Hypermedia Applications</i> |
| 2001-6 | Martijn van Welie (VU) <i>Task-based User Interface Design</i> | 2002-14 | Wieke de Vries (UU) <i>Agent Interaction: Abstract Approaches to Modelling, Programming and</i> |
| 2001-7 | Bastiaan Schonhage (VU) <i>Diva: Architectural Perspectives on Information Visualization</i> | 2002-15 | Rik Eshuis (UT) <i>Semantics and Verification of UML Activity Diagrams for Workflow Modelling</i> |
| 2001-8 | Pascal van Eck (VU) <i>A Compositional Semantic Structure for Multi-Agent Systems Dynamics.</i> | 2002-16 | Pieter van Langen (VU) <i>The Anatomy of Design: Foundations, Models and Applications</i> |
| 2001-9 | Pieter Jan 't Hoen (RUL) <i>Towards Distributed Development of Large Object-Oriented Models, Views of Packages as Classes</i> | 2002-17 | Stefan Manegold (UVA) <i>Understanding, Modeling, and Improving Main-Memory Database Performance</i> |
| 2001-10 | Maarten Sierhuis (UvA) <i>Modeling and Simulating Work Practice BRAHMS: a multiagent modeling and simulation language for work practice analysis and design</i> | 2003 | |
| 2001-11 | Tom M. van Engers (VUA) <i>Knowledge Management: The Role of Mental Models in Business Systems Design</i> | 2003-01 | Heiner Stuckenschmidt (VU) <i>Ontology-Based Information Sharing in Weakly Structured Environments</i> |
| 2002 | | 2003-02 | Jan Broersen (VU) <i>Modal Action Logics for Reasoning About Reactive Systems</i> |
| 2002-01 | Nico Lassing (VU) <i>Architecture-Level Modifiability Analysis</i> | 2003-03 | Martijn Schuemie (TUD) <i>Human-Computer Interaction and Presence in Virtual Reality Exposure Therapy</i> |
| 2002-02 | Roelof van Zwol (UT) <i>Modelling and searching web-based document collections</i> | 2003-04 | Milan Petkovic (UT) <i>Content-Based Video Retrieval Supported by Database Technology</i> |
| 2002-03 | Henk Ernst Blok (UT) <i>Database Optimization Aspects for Information Retrieval</i> | 2003-05 | Jos Lehmann (UVA) <i>Causation in Artificial Intelligence and Law - A modelling approach</i> |
| 2002-04 | Juan Roberto Castelo Valdueza (UU) <i>The Discrete Acyclic Digraph Markov Model in Data Mining</i> | 2003-06 | Boris van Schooten (UT) <i>Development and specification of virtual environments</i> |
| 2002-05 | Radu Serban (VU) <i>The Private Cyberspace Modeling Electronic Environments inhabited by Privacy-concerned Agents</i> | 2003-07 | Machiel Jansen (UvA) <i>Formal Explorations of Knowledge Intensive Tasks</i> |
| 2002-06 | Laurens Mommers (UL) <i>Applied legal epistemology; Building a knowledge-based ontology of the legal domain</i> | 2003-08 | Yongping Ran (UM) <i>Repair Based Scheduling</i> |
| 2002-07 | Peter Boncz (CWI) <i>Monet: A Next-Generation DBMS Kernel For Query-Intensive Applications</i> | 2003-09 | Rens Kortmann (UM) <i>The resolution of visually guided behaviour</i> |
| | | 2003-10 | Andreas Lincke (UvT) <i>Electronic Business Negotiation: Some experimental studies on the interaction between medium, innovation context and culture</i> |

| | | | |
|---------|--|---------|---|
| 2003-11 | Simon Keizer (UT) <i>Reasoning under Uncertainty in Natural Language Dialogue using Bayesian Networks</i> | 2004-12 | The Duy Bui (UT) <i>Creating emotions and facial expressions for embodied agents</i> |
| 2003-12 | Roeland Ordelman (UT) <i>Dutch speech recognition in multimedia information retrieval</i> | 2004-13 | Wojciech Jamroga (UT) <i>Using Multiple Models of Reality: On Agents who Know how to Play</i> |
| 2003-13 | Jeroen Donkers (UM) <i>Nosce Hostem - Searching with Opponent Models</i> | 2004-14 | Paul Harrenstein (UU) <i>Logic in Conflict. Logical Explorations in Strategic Equilibrium</i> |
| 2003-14 | Stijn Hoppenbrouwers (KUN) <i>Freezing Language: Conceptualisation Processes across ICT-Supported Organisations</i> | 2004-15 | Arno Knobbe (UU) <i>Multi-Relational Data Mining</i> |
| 2003-15 | Mathijs de Weerd (TUD) <i>Plan Merging in Multi-Agent Systems</i> | 2004-16 | Federico Divina (VU) <i>Hybrid Genetic Relational Search for Inductive Learning</i> |
| 2003-16 | Menzo Windhouwer (CWI) <i>Feature Grammar Systems - Incremental Maintenance of Indexes to Digital Media Warehouses</i> | 2004-17 | Mark Winands (UM) <i>Informed Search in Complex Games</i> |
| 2003-17 | David Jansen (UT) <i>Extensions of Statecharts with Probability, Time, and Stochastic Timing</i> | 2004-18 | Vania Bessa Machado (UvA) <i>Supporting the Construction of Qualitative Knowledge Models</i> |
| 2003-18 | Levente Kocsis (UM) <i>Learning Search Decisions</i> | 2004-19 | Thijs Westerveld (UT) <i>Using generative probabilistic models for multimedia retrieval</i> |
| 2004 | | 2004-20 | Madelon Evers (Nyenrode) <i>Learning from Design: facilitating multi-disciplinary design teams</i> |
| | | 2005 | |
| 2004-01 | Virginia Dignum (UU) <i>A Model for Organizational Interaction: Based on Agents, Founded in Logic</i> | 2005-01 | Floor Verdenius (UVA) <i>Methodological Aspects of Designing Induction-Based Applications</i> |
| 2004-02 | Lai Xu (UvT) <i>Monitoring Multi-party Contracts for E-business</i> | 2005-02 | Erik van der Werf (UM)) <i>AI techniques for the game of Go</i> |
| 2004-03 | Perry Groot (VU) <i>A Theoretical and Empirical Analysis of Approximation in Symbolic Problem Solving</i> | 2005-03 | Franc Grootjen (RUN) <i>A Pragmatic Approach to the Conceptualisation of Language</i> |
| 2004-04 | Chris van Aart (UVA) <i>Organizational Principles for Multi-Agent Architectures</i> | 2005-04 | Nirvana Meratnia (UT) <i>Towards Database Support for Moving Object data</i> |
| 2004-05 | Viara Popova (EUR) <i>Knowledge discovery and monotonicity</i> | 2005-05 | Gabriel Infante-Lopez (UVA) <i>Two-Level Probabilistic Grammars for Natural Language Parsing</i> |
| 2004-06 | Bart-Jan Hommes (TUD) <i>The Evaluation of Business Process Modeling Techniques</i> | 2005-06 | Pieter Spronck (UM) <i>Adaptive Game AI</i> |
| 2004-07 | Elise Boltjes (UM) <i>Voorbeeldig onderwijs; voorbeeldgestuurd onderwijs, een opstap naar</i> | 2005-07 | Flavius Frasinca (TUE) <i>Hypermedia Presentation Generation for Semantic Web Information Systems</i> |
| 2004-08 | Joop Verbeek (UM) <i>Politie en de Nieuwe Internationale Informatiemarkt, Grensregionale politieke gegevensuitwisseling en digitale expertise</i> | 2005-08 | Richard Vdovjak (TUE) <i>A Model-driven Approach for Building Distributed Ontology-based Web Applications</i> |
| 2004-09 | Martin Caminada (VU) <i>For the Sake of the Argument; explorations into argument-based reasoning</i> | 2005-09 | Jeen Broekstra (VU) <i>Storage, Querying and Inferencing for Semantic Web Languages</i> |
| 2004-10 | Suzanne Kabel (UVA) <i>Knowledge-rich indexing of learning-objects</i> | 2005-10 | Anders Bouwer (UVA) <i>Explaining Behaviour: Using Qualitative Simulation in Interactive Learning Environments</i> |
| 2004-11 | Michel Klein (VU) <i>Change Management for Distributed Ontologies</i> | 2005-11 | Elth Ogston (VU) <i>Agent Based Matchmaking and Clustering - A Decentralized Approach to Search</i> |

-
- | | |
|--|--|
| <p>2005-12 Csaba Boer (EUR) <i>Distributed Simulation in Industry</i></p> <p>2005-13 Fred Hamburg (UL) <i>Een Computermodel voor het Onderstemen van Euthanasiebeslissingen</i></p> <p>2005-14 Borys Omelayenko (VU) <i>Web-Service configuration on the Semantic Web; Exploring how semantics meets pragmatics</i></p> <p>2005-15 Tibor Bosse (VU) <i>Analysis of the Dynamics of Cognitive Processes</i></p> <p>2005-16 Joris Graaumans (UU) <i>Usability of XML Query Languages</i></p> <p>2005-17 Boris Shishkov (TUD) <i>Software Specification Based on Re-usable Business Components</i></p> <p>2005-18 Danielle Sent (UU) <i>Test-selection strategies for probabilistic networks</i></p> <p>2005-19 Michel van Dartel (UM) <i>Situated Representation</i></p> <p>2005-20 Cristina Coteanu (UL) <i>Cyber Consumer Law, State of the Art and Perspectives</i></p> <p>2005-21 Wijnand Derks (UT) <i>Improving Concurrency and Recovery in Database Systems by Exploiting Application Semantics</i></p> | <p>2006-11 Joeri van Ruth (UT) <i>Flattening Queries over Nested Data Types</i></p> <p>2006-12 Bert Bongers (VU) <i>Interaction - Towards an e-ecology of people, our technological environment, and the arts</i></p> <p>2006-13 Henk-Jan Lebbink (UU) <i>Dialogue and Decision Games for Information Exchanging Agents</i></p> <p>2006-14 Johan Hoorn (VU) <i>Software Requirements: Update, Upgrade, Redesign - towards a Theory of Requirements Change</i></p> <p>2006-15 Rainer Malik (UU) <i>CONAN: Text Mining in the Biomedical Domain</i></p> <p>2006-16 Carsten Riggelsen (UU) <i>Approximation Methods for Efficient Learning of Bayesian Networks</i></p> <p>2006-17 Stacey Nagata (UU) <i>User Assistance for Multitasking with Interruptions on a Mobile Device</i></p> <p>2006-18 Valentin Zhizhkhun (UVA) <i>Graph transformation for Natural Language Processing</i></p> <p>2006-19 Birna van Riemsdijk (UU) <i>Cognitive Agent Programming: A Semantic Approach</i></p> <p>2006-20 Marina Velikova (UvT) <i>Monotone models for prediction in data mining</i></p> <p>2006-21 Bas van Gils (RUN) <i>Aptness on the Web</i></p> <p>2006-22 Paul de Vrieze (RUN) <i>Fundaments of Adaptive Personalisation</i></p> <p>2006-23 Ion Juvina (UU) <i>Development of Cognitive Model for Navigating on the Web</i></p> <p>2006-24 Laura Hollink (VU) <i>Semantic Annotation for Retrieval of Visual Resources</i></p> <p>2006-25 Madalina Drugan (UU) <i>Conditional log-likelihood MDL and Evolutionary MCMC</i></p> <p>2006-26 Vojkan Mihajlovic (UT) <i>Score Region Algebra: A Flexible Framework for Structured Information Retrieval</i></p> <p>2006-27 Stefano Bocconi (CWI) <i>Vox Populi: generating video documentaries from semantically annotated media repositories</i></p> <p>2006-28 Borkur Sigurbjornsson (UVA) <i>Focused Information Access using XML Element Retrieval</i></p> |
|--|--|
-
- 2006
-
- | | |
|---|---|
| <p>2006-01 Samuil Angelov (TUE) <i>Foundations of B2B Electronic Contracting</i></p> <p>2006-02 Cristina Chisalita (VU) <i>Contextual issues in the design and use of information technology in organizations</i></p> <p>2006-03 Noor Christoph (UVA) <i>The role of metacognitive skills in learning to solve problems</i></p> <p>2006-04 Marta Sabou (VU) <i>Building Web Service Ontologies</i></p> <p>2006-05 Cees Pierik (UU) <i>Validation Techniques for Object-Oriented Proof Outlines</i></p> <p>2006-06 Ziv Baida (VU) <i>Software-aided Service Bundling - Intelligent Methods & Tools for Graphical Service Modeling</i></p> <p>2006-07 Marko Smiljanic (UT) <i>XML schema matching – balancing efficiency and effectiveness by means of clustering</i></p> <p>2006-08 Eelco Herder (UT) <i>Forward, Back and Home Again - Analyzing User Behavior on the Web</i></p> <p>2006-09 Mohamed Wahdan (UM) <i>Automatic Formulation of the Auditor's Opinion</i></p> <p>2006-10 Ronny Siebes (VU) <i>Semantic Routing in Peer-to-Peer Systems</i></p> | <p>2007-01 Kees Leune (UvT) <i>Access Control and Service-Oriented Architectures</i></p> <p>2007-02 Wouter Teepe (RUG) <i>Reconciling Information Exchange and Confidentiality: A Formal Approach</i></p> |
|---|---|
-
- 2007
-

- | | |
|--|--|
| <p>2007-03 Peter Mika (VU) <i>Social Networks and the Semantic Web</i></p> <p>2007-04 Jurriaan van Diggelen (UU) <i>Achieving Semantic Interoperability in Multi-agent Systems: a dialogue-based approach</i></p> <p>2007-05 Bart Schermer (UL) <i>Software Agents, Surveillance, and the Right to Privacy: a Legislative Framework for Agent-enabled Surveillance</i></p> <p>2007-06 Gilad Mishne (UVA) <i>Applied Text Analytics for Blogs</i></p> <p>2007-07 Natasa Jovanovic' (UT) <i>To Whom It May Concern - Addressee Identification in Face-to-Face Meetings</i></p> <p>2007-08 Mark Hoogendoorn (VU) <i>Modeling of Change in Multi-Agent Organizations</i></p> <p>2007-09 David Mobach (VU) <i>Agent-Based Mediated Service Negotiation</i></p> <p>2007-10 Huib Aldewereld (UU) <i>Autonomy vs. Conformity: an Institutional Perspective on Norms and Protocols</i></p> <p>2007-11 Natalia Stash (TUE) <i>Incorporating Cognitive/Learning Styles in a General-Purpose Adaptive Hypermedia System</i></p> <p>2007-12 Marcel van Gerven (RUN) <i>Bayesian Networks for Clinical Decision Support: A Rational Approach to Dynamic Decision-Making under Uncertainty</i></p> <p>2007-13 Rutger Rienks (UT) <i>Meetings in Smart Environments; Implications of Progressing Technology</i></p> <p>2007-14 Niek Bergboer (UM) <i>Context-Based Image Analysis</i></p> <p>2007-15 Joyca Lacroix (UM) <i>NIM: a Situated Computational Memory Model</i></p> <p>2007-16 Davide Grossi (UU) <i>Designing Invisible Handcuffs. Formal investigations in Institutions and Organizations for Multi-agent Systems</i></p> <p>2007-17 Theodore Charitos (UU) <i>Reasoning with Dynamic Networks in Practice</i></p> <p>2007-18 Bart Orriens (UvT) <i>On the development an management of adaptive business collaborations</i></p> <p>2007-19 David Levy (UM) <i>Intimate relationships with artificial partners</i></p> <p>2007-20 Slinger Jansen (UU) <i>Customer Configuration Updating in a Software Supply Network</i></p> <p>2007-21 Karianne Vermaas (UU) <i>Fast diffusion and broadening use: A research on residential adoption and usage of broadband internet in the Netherlands between 2001 and 2005</i></p> <p>2007-22 Zlatko Zlatev (UT) <i>Goal-oriented design of value and process models from patterns</i></p> | <p>2007-23 Peter Barna (TUE) <i>Specification of Application Logic in Web Information Systems</i></p> <p>2007-24 Georgina Ramirez Camps (CWI) <i>Structural Features in XML Retrieval</i></p> <p>2007-25 Joost Schalken (VU) <i>Empirical Investigations in Software Process Improvement</i></p> |
| 2008 | |
| <p>2008-01 Katalin Boer-Sorbán (EUR) <i>Agent-Based Simulation of Financial Markets: A modular, continuous-time approach</i></p> <p>2008-02 Alexei Sharpanskykh (VU) <i>On Computer-Aided Methods for Modeling and Analysis of Organizations</i></p> <p>2008-03 Vera Hollink (UVA) <i>Optimizing hierarchical menus: a usage-based approach</i></p> <p>2008-04 Ander de Keijzer (UT) <i>Management of Uncertain Data - towards unattended integration</i></p> <p>2008-05 Bela Mutschler (UT) <i>Modeling and simulating causal dependencies on process-aware information systems from a cost perspective</i></p> <p>2008-06 Arjen Hommersom (RUN) <i>On the Application of Formal Methods to Clinical Guidelines, an Artificial Intelligence Perspective</i></p> <p>2008-07 Peter van Rosmalen (OU) <i>Supporting the tutor in the design and support of adaptive e-learning</i></p> <p>2008-08 Janneke Bolt (UU) <i>Bayesian Networks: Aspects of Approximate Inference</i></p> <p>2008-09 Christof van Nimwegen (UU) <i>The paradox of the guided user: assistance can be counter-effective</i></p> <p>2008-10 Wauter Bosma (UT) <i>Discourse oriented summarization</i></p> <p>2008-11 Vera Kartseva (VU) <i>Designing Controls for Network Organizations: A Value-Based Approach</i></p> <p>2008-12 Jozsef Farkas (RUN) <i>A Semiotically Oriented Cognitive Model of Knowledge Representation</i></p> <p>2008-13 Caterina Carraciolo (UVA) <i>Topic Driven Access to Scientific Handbooks</i></p> <p>2008-14 Arthur van Bunningen (UT) <i>Context-Aware Querying; Better Answers with Less Effort</i></p> <p>2008-15 Martijn van Otterlo (UT) <i>The Logic of Adaptive Behavior: Knowledge Representation and Algorithms for the Markov Decision Process Framework in First-Order Domains.</i></p> <p>2008-16 Henriette van Vugt (VU) <i>Embodied agents from a user's perspective</i></p> | |

- 2008-17** Martin Op 't Land (TUD)
Applying Architecture and Ontology to the Splitting and Allying of Enterprises
- 2008-18** Guido de Croon (UM)
Adaptive Active Vision
- 2008-19** Henning Rode (UT)
From Document to Entity Retrieval: Improving Precision and Performance of Focused Text Search
- 2008-20** Rex Arendsen (UVA)
Geen bericht, goed bericht. Een onderzoek naar de effecten van de introductie van elektronisch berichtenverkeer met de overheid op de administratieve lasten van bedrijven.
- 2008-21** Krisztian Balog (UVA)
People Search in the Enterprise
- 2008-22** Henk Koning (UU)
Communication of IT-Architecture
- 2008-23** Stefan Visscher (UU)
Bayesian network models for the management of ventilator-associated pneumonia
- 2008-24** Zharko Aleksovski (VU)
Using background knowledge in ontology matching
- 2008-25** Geert Jonker (UU)
Efficient and Equitable Exchange in Air Traffic Management Plan Repair using Spender-signed Currency
- 2008-26** Marijn Huijbregts (UT)
Segmentation, Diarization and Speech Transcription: Surprise Data Unraveled
- 2008-27** Hubert Vogten (OU)
Design and Implementation Strategies for IMS Learning Design
- 2008-28** Ildiko Flesch (RUN)
On the Use of Independence Relations in Bayesian Networks
- 2008-29** Dennis Reidsma (UT)
Annotations and Subjective Machines - Of Annotators, Embodied Agents, Users, and Other Humans
- 2008-30** Wouter van Atteveldt (VU)
Semantic Network Analysis: Techniques for Extracting, Representing, and Querying Media Content