Introduction

‘No-campaign Wilders is a circus for bodyguards and media’
(Wilders’ tour NEE 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
1.1 Introduction

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 Schrodt 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 Schrodt (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-
answer 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. 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-
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, criticised, 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.