Advance Access publication October 14, 2008

Political Analysis (2008) 16:428–446
doi:10.1093/pan/mpn006

Parsing, Semantic Networks, and Political Authority Using Syntactic Analysis to Extract Semantic Relations from Dutch Newspaper Articles

Wouter van Atteveldt and Jan Kleinnijenhuis
Faculty of Social Sciences, Vrije Universiteit Amsterdam,
De Boelelaan 1081, 1081 HV Amsterdam, the Netherlands
e-mail: wouter@vanatteveldt.com (corresponding author)

Nel Ruigrok
Amsterdam School of Communication Research,
University of Amsterdam, Kloveniersburgwal 48, 1012 CX Amsterdam, the Netherlands
e-mail: p.c.ruigrok@uva.nl

Analysis of political communication is an important aspect of political research. Thematic content analysis has yielded considerable success both with manual and automatic coding, but Semantic Network Analysis has proven more difficult, both for humans and for the computer. This article presents a system for an automated Semantic Network Analysis of Dutch texts. The system automatically extracts relations between political actors based on the output of syntactic analysis of Dutch newspaper articles. Specifically, the system uses pattern matching to find source constructions and determine the semantic agent and patient of relations, and name matching and anaphora resolution to identify political actors. The performance of the system is judged by comparing the extracted relations to manual codings of the same material. Results on the level of measurement indicate acceptable performance. We also estimate performance at the levels of analysis by using a case study of media authority, resulting in good correlations between the theoretical variables derived from the automatic and manual analysis. Finally, we test a number of substantive hypotheses with regression models using the automatic and manual output, resulting in highly similar models in each case. This suggests that our method has sufficient performance to be used to answer relevant political questions in a valid way.

1 Introduction

Political language is omnipresent in the game of politics. The power of words undergirds the persuasive appeal of political speeches, negotiation moves, eyewitness reports, and

Author's note: Wouter van Atteveldt is an interdisciplinary Ph.D. student at the Departments of Communication Science and Artificial Intelligence at the Vrije Universiteit Amsterdam (the Netherlands). He hopes to defend his Ph.D. thesis on Methods for automatically extracting and representing media data in the Fall of 2008. Jan Kleinnijenhuis is a professor of mass communication. His research deals with news selection and news effects. Nel Ruigrok is a researcher at the Netherlands News Monitor and a research fellow in the Amsterdam School of Communications Research at the University of Amsterdam. Her research interests include the role of media in the political arena, especially during times of conflict.

© The Author 2008. Published by Oxford University Press on behalf of the Society for Political Methodology.
All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

https://doi.org/10.1093/pan/mpn006 Published online by Cambridge University Press

media coverage. Hence, content analysis of political texts is an important tool in political analysis to grasp interactions in the political process.

The greatest successes in content analysis have been achieved when the research question could be reduced to a *thematic* question about the *emphasis* on themes (issues) or actors in texts. However, there is more to political language than visibility and salience of actors and issues. Politicians try to position themselves on various issues. Politicians build their profile by attacks and cooperation with others, by debating with some competitors and ignoring others. They strive for political momentum by associating themselves with successes rather than failures. Likewise, journalists portray some politicians as winning, others as losing, some as acting, and others as being acted upon. These relations can be seen as the semantic network of the political actors and issues, and exposing and analyzing this network can be a great contribution to political analysis. Semantic Network Analysis (Krippendorff 2004) is the branch of Content Analysis that directly investigates these aspects of political language by representing text as a network of semantic relations between concepts (p. 292). In political communication, such concepts can be political actors or issues, and the relations can be associations, opinions, or actions.

Computer content analysis affords great advantages over manual content analysis, making it possible to analyze large sets of documents quickly and efficiently and decreasing the possibility of systematic bias. However, most successes of computer content analysis so far have relied on word and co-occurrence counts (cf. Roberts 1989). Automating the extraction of Semantic Networks is more difficult due to the capriciousness and ambiguity of the language used for expressing semantic elements such as affect and intentions. For example, Wiebe et al. (2004) found that subjective language uses low-frequency words much more often than objective texts. Nevertheless, considerable success has been achieved by focusing on specific aspects of texts in specific genres, such as illustrated by the Kansas Events Data System (KEDS; Schrodt 2001), which focuses on the extraction of conflict and cooperation between international actors from newspaper headlines. Additionally, new visualization techniques to reveal the structure of such data have been developed (Brandes, Fleischer, and Lerner 2006). Moreover, enormous progress has been made during the last decades in the field of natural language processing (NLP). For example, it is possible to automatically assign each word to its Part-Of-Speech, such as verb, noun, or adjective, with very high degree of accuracy (Manning and Schütze 2002, 371). Automatic syntactic parsers have been developed for Dutch, English, and other languages that accurately determine the syntactic structure of each sentence: what is the main verb, what are its subject and object, which adjectives and prepositions modify which nouns or verbs, etc (Manning and Schütze 2002, 455; Van Noord 2006). These tools can be used in automatic content analysis. For example, Corman et al. (2002) use syntactic parsing to restrict their search of key subjects in organizational communication to words that co-occur in noun phrases.

In this article, we will show that this progress in NLP, and specifically the existence of high-quality syntactic parsers, can be used to extend the range of political research questions that can be answered automatically. The central research question this article tries to answer is methodological:

Can we use the grammatical structure of text to extract Semantic Network information that is relevant for Political Analysis?

Focusing on the Dutch language, we will automatically extract citations and their sources as well as semantic agent/patient relations, creating a Semantic Network as defined above. To assess the quality of the automation of the method presented here, a semantic network content analysis by human coders of the most recent Dutch election campaign will

https://doi.org/10.1093/pan/mpn006 Published online by Cambridge University Press

be used as the gold standard (Kleinnijenhuis et al. 2007). The availability of human codings allows us to assess the validity of the method at different levels: at the level of measurement (individual statements), at the level of analysis (analytical constructs at the aggregate level), and at the level of modeling interactions with variables external to the text. At each of these levels, we can compare the outcome using the automatic method with the outcome of the manual analysis. This gives us an estimate of whether the quality of the extracted Semantic Networks is sufficient for use in political analysis.

The contribution of this paper is two-fold: First, we show that it is feasible to process the output of available linguistic tools to render a Semantic Network Analysis of texts. We present an automated system for extracting Semantic Relations from syntax graphs for Dutch newspapers articles and show that this system obtains acceptable performance. Second, we show that the output of this system, the Semantic Networks, can be used to answer questions that are relevant for political analysis and that are difficult to answer using thematic content analysis.

This article is part of an ongoing effort for automating Semantic Network Analysis described in (Van Atteveldt 2008). Apart from this method, we developed the open source programs AmCAT (Amsterdam Content Analysis Toolkit) for document management and analysis and iNet for manual Semantic Network Analysis. Moreover, Van Atteveldt et al. (2008) describe a method for automatically determining the sentiment or valence of relations between actors and issues, and Van Atteveldt, Schlobach, and Van Harmelen (2007) describe a system for representing, visualizing, and analyzing Semantic Network data. Taken together, these systems provide the basis for thorough and efficient exploration of political language.

The remainder of this article is structured as follows. The next section will describe the automated Semantic Network Analysis system. The third section describes the methodology for testing the validity of the system, including a small case study on measuring media authority. The fourth section presents the results of these tests, as well as the outcomes of the case study. The final section will discuss these findings and point out limitations and possibilities for future research.

2 Method: Determining Semantic Roles Using Syntax Patterns

Figure 1 gives an overview of the method presented in this chapter. The input of the method consists of the syntax (dependency) graphs of these documents. In the dependency graph of a sentence, each node is a word from that sentence and each edge is a grammatical relation

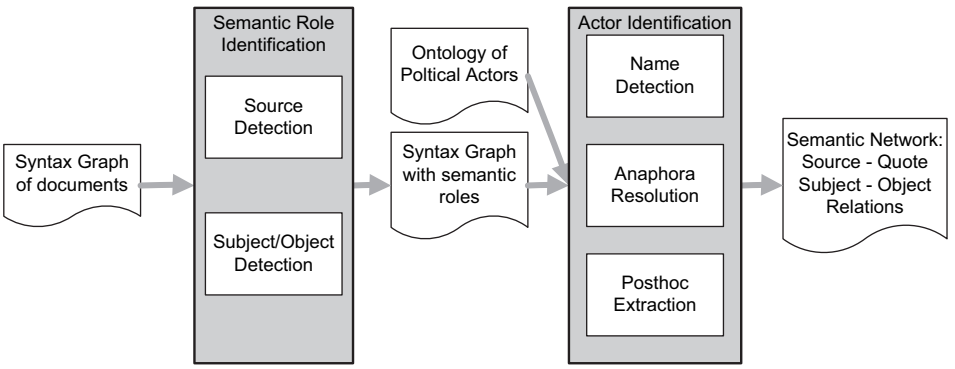


Fig. 1 Overview of the method for extracting Semantic Networks.

https://doi.org/10.1093/pan/mpn006 Published online by Cambridge University Press

between two words. For example, the dependency graph of the sentence ''John likes Mary'' will contain three nodes, one for each word. The node ''John'' will be connected to ''likes'' with the grammatical subject relation, whereas ''Mary'' will be connected to ''likes'' with the direct object relation. In our implementation, we preprocessed all sentences using the freely available Dutch Alpino parser (Van Noord 2006; Van Noord and Malouf 2000).

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. In this graph, the Actor Identification module identifies all relevant political actors as described in the political ontology. This ontology is a rich description of the different political actors and their relevant properties such as gender, party membership, and political function (Sowa 2000; Antoniou and Van Harmelen 2004). The Actor Identification Module detects the occurrence of the actors from 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 actors from the ontology.

2.1 *Source Detection*

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 recognize these patterns from the text and dependency graph using a limited number of pattern. In principle, we are only interested in extracting quotes as assertive speech acts (Searle 1969), not as directives (orders) or commissives (promises).

Figure 2 shows the dependency trees of two fictive newspaper sentences, corresponding to the main patterns used in this module. Figure 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 gray line.

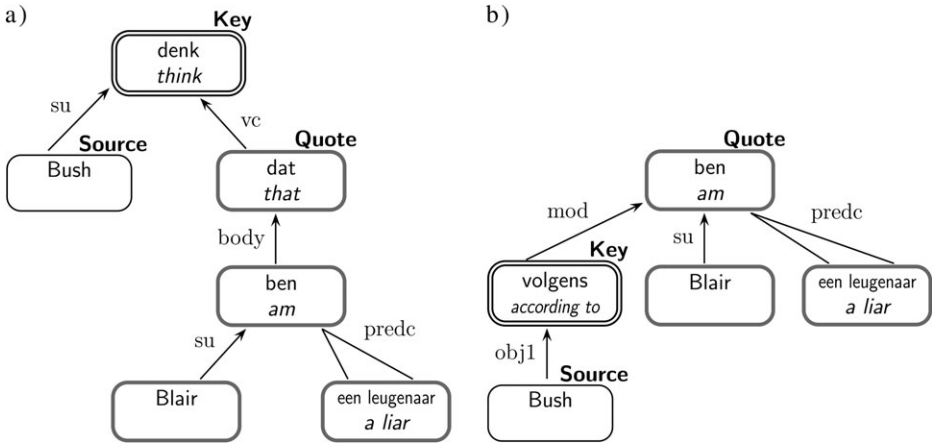


Fig. 2 Example syntax patterns for source recognition. (a) Bush thinks that Blair is a liar. (b) According to Bush, Blair is a liar.

https://doi.org/10.1093/pan/mpn006 Published online by Cambridge University Press

Figure 2b is an example of the second main pattern: [$p_{according}$ 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, whereas the parent of the preposition and its children, except for the preposition itself, forms the quote. This pattern is often used when citing reports (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''). The program containing the full set of patterns for this and the next module is available from the web appendix to this article.

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, whereas the semantic object (or patient) is the actor or entity that it is done to (Dixon 1991). Figure 3 shows some simple example sentences, with the syntactic subject and object indicated above the text and the semantic subject and object below it

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, Mur, and van Noord (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:

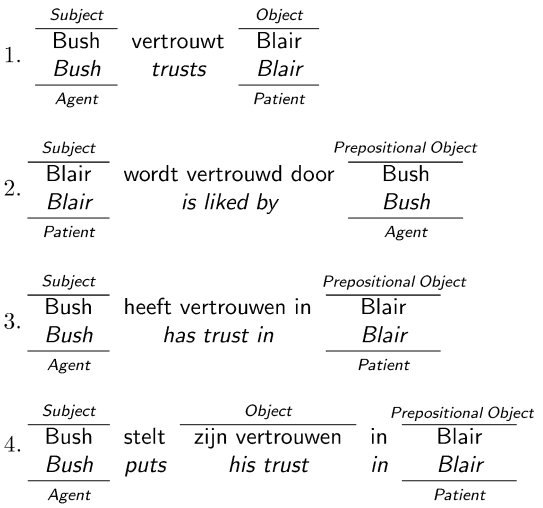


Fig. 3 Example sentences with subject, object and agent, patient.

https://doi.org/10.1093/pan/mpn006 Published online by Cambridge University Press

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.

An alternative approach is using machine learning techniques to extract relations, such as done by Jijkoun (2007) and Zhang, Zhou, and Aw (2008). 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 Fig. 4 of the Dutch sentence reading ''Kerry zegt dat Bush hem een draaikont heeft genoemd'' (Kerry says that Bush called him a flip-flopper). Figure 4a is the raw dependency graph. 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 (obj2) *him*, and a direct object (obj1) *flip-flopper*. Finally, *flip-flopper* has a determiner *a*. Figure 4b 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, whereas the subject *Kerry* is indicated by a solid black circle and

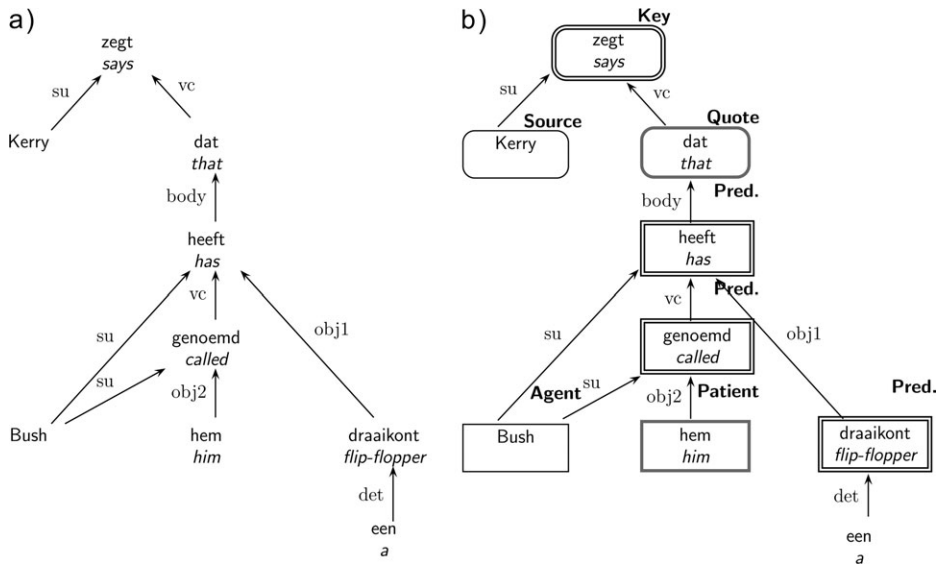


Fig. 4 Example of a parsed sentence with recognized semantic roles. (a) Parsed sentence. (b) Recognized semantic roles.

https://doi.org/10.1093/pan/mpn006 Published online by Cambridge University Press

the quote *that* (and all underlying nodes) by a thick gray circle. The rectangles indicate a subject-predicate–object construction. The predicate is *has called flip-flopper*, which is displayed using a dashed rectangle. The subject, *Bush*, uses a solid line, whereas the line for the object, *him*, is thick gray. Note that *flip-flopper*, although it is the grammatical object, is correctly classified as part of the predicate.

2.3 *Actor Identification*

The task of the next module of the method is recognizing 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 recognizer 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 recognize 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.''

2.3.1 Anaphora resolution

The main difficulty in recognizing 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 using constraints, and rank the remaining candidates using a heuristic function.

From the constraints and heuristics surveyed by Mitkov (2002) and Lappin and Leass (1994), four seem immediately useful for our application: The first constraint in 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, whereas ''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, whereas a reflexive pronoun (zich, *himself*) has to refer to such a noun: In ''Bush likes him,'' him cannot refer to Bush, whereas in ''Bush likes himself'' the pronoun has to refer back to Bush. The first heuristic factor is syntactic parallelism: a noun in the same syntactic position as the anaphora is preferred over other nouns. The second heuristic is a salience heuristic described by Lappin and Leass (1994) that essentially prefers subjects to objects and recent sentences to earlier sentences.

Based on these factors, 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.

https://doi.org/10.1093/pan/mpn006 Published online by Cambridge University Press

2. Candidate selection: A list of possible noun phrases is made from the text preceding the anaphora, using the following ranking:

   Antecedents are ranked from the current sentence to the first sentence in the paragraph.

   Within a sentence, antecedents are ranked by 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 Fig. 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 Nederlanderschap leek te ontnemen. (Hirsi Ali left politics because Minister Verdonk took away 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, who the module knows was minister of immigration at that time. This shows how background knowledge of the domain is required for filtering out possible antecedents. 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.

2.3.2 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.

3 Determining Validity

The purpose of this study is to investigate the performance of a method for extracting the semantic relations from text using grammatical analysis. For normal content analysis studies, the most important indicator of measurement quality is the intercoder reliability. For (rule-based) computer analysis, however, the reliability is always 100% as the output is

https://doi.org/10.1093/pan/mpn006 Published online by Cambridge University Press

fixed deterministically by the input. However, the computer can reliably produce completely invalid results if the rules do not capture the relation between the text and the quantity to measure correctly. Therefore, we need to determine the validity of the automatic method to determine whether we can use it for analyzing political communication. This validity will be estimated in four different ways: the concurrent validity at the levels of measurement, analysis, and predictions and the face validity.

3.1 *Validity at the Level of Measurement*

To determine the validity of the automatic method, we can use the manual Semantic Network Analysis of the 2006 election campaign described in (Kleinnijenhuis et al. 2007) as a Gold Standard. This study analyzed 5,988 newspaper articles, resulting in 15,914 coded statements containing at least one political actors. We assume this manual analysis to be valid based on its reliability and its long history of use in studying political communication. By comparing the automatic output to the manual analysis, we can establish the concurrent validity of the automatic method at the level of measurement.

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. Each item is then assigned to one of four categories: If both the computer and the human coder assigned an item to the relevant set, it is considered a true positive. If the computer assigned it but the human coder did not, it is a false positive. If the human coder assigned it but the computer did not, it is a false negative. If neither assigned it to that set, it is a true negative. From the frequency of true positives and false positives and negatives, we can calculate the precision, recall, and F1 score metrics using the equations below:

$$\text{Precision} = \frac{\text{True positive}}{\text{True positive} + \text{False positive}}. \tag{1}$$

$$\text{Recall} = \frac{\text{True positive}}{\text{True positive} + \text{False negative}}. \tag{2}$$

$$F1 \text{ score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{3}$$

For example, suppose there were four true positives, two false positives, and six false negatives. This gives a precision of 4/6 = .67 and a recall of 4/10 = .4. The $F1$ score is then 2cot 67 × .4/(.4 + .67) = .50. Note that the true negatives are not used in the calculation.

Precision indicates whether, if the computer method states that a certain actor plays a certain role, this was actually correct. Recall indicates how many of the roles actually played by the different actors were found by the computer method. Finally, $F1$ score is the harmonic average of these two measures and gives a good overall estimate of performance. These metrics are more informative than base accuracy and penalize methods that overrepresent the (often large) negative category: since most actors only occur in a small proportion of sentences, not outputing any actors in any of the roles would give a high accuracy, but would not be very meaningful. Since true negatives are not used in the

https://doi.org/10.1093/pan/mpn006 Published online by Cambridge University Press

calculation of the above metrics, such a method would score zero on them. Precision, recall, and $F1$ score are standard performance metrics in Computational Linguistics (Manning and Schütze 2002, 267–271).

3.2 *Validity at the Level of Analysis*

In analyzing (political) communication, we are generally interested in validly measuring theoretical variables at the level of analysis, rather than making sure every individual sentence is analyzed correctly. White noise, that is unbiased errors, at the level of measurement do not affect the results at the level of analysis if the unit of aggregation is large enough. To determine the validity at the level of analysis, we will determine the correspondence of the manual and automatic analysis on the operationalization of theoretical variables that use the extracted semantic roles. These variables depend on the substantive research question, so we need to conduct a specific case study to determine the validity at the level of analysis.

The case study that we present here analyses what Herbst (2003, 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. Donsbach, Mattenklott, and Brosius 1993; Patterson 1993; Donsbach and Jandura 2003). This can be operationalized 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 (Kleinnijenhuis et al. 1998). This definition can be operationalized as the relative frequency with which an actor appears as the semantic agent rather than as patient.

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

3.3 *Validity at the Level of Predictions*

The two measures described above both assume the manual analysis to be correct and any deviation from that to be an error. Due to the ambiguity inherent in language, however, this would not be the case even if human coders would be perfect: some sentences can be interpreted in different ways, meaning that different codings of the same sentence could both be correct. Hence, the validity as described above will be a pessimistic estimate of the real validity, as ambiguous cases or human errors will result in lower scores even if the automatic output is correct. To overcome this limitation, we also determine the concurrent validity at the level of predicting (modeling) the interaction of the extracted media variables with external variables. For this end, 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 validity of the method at the level of modeling or predicting. These hypothesis pairs are as follows:

https://doi.org/10.1093/pan/mpn006 Published online by Cambridge University Press

3.3.1 H1: Indexing theory

Bennett (1990, 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, Donohue, and Tichenor 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 operationalized as the percentage 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 operationalized above. Table 1 in the web appendix to this article lists the parties with seats in the Dutch parliament before the 2006 elections.

3.3.2 H2: Opportune witnesses

The specific selection of sources is a central part of the theory of ''opportune 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 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 pillarized 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. 2007). 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 with the newspaper. Table 1 in the web appendix lists the Dutch national newspapers that we consider to be aligned with the various parties. It should be noted that this operationalization 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 article is not to give the definitive answer on this hypothesis; so even if this operationalization is debatable, it is interesting to investigate whether the method presented in this chapter can be used to test it.

https://doi.org/10.1093/pan/mpn006 Published online by Cambridge University Press

3.3.3 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, Benoit, and Hansen 2003; Benoit 2007). For example, the welfare issue is traditionally owned by left-wing parties, whereas 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, whereas negative economic development is harmful (Lewis-Beck 2006). From these theories, we hypothesize that political actors will be more authoritative on issues they own, in which we include valence issues for the governing parties:

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 United States 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, Park, and Ridout 2006), which makes it worthwhile to test these hypotheses outside the United States. Similar to H2, these hypotheses are tested by regressing authority on issue ownership. Issue ownership is operationalized as a dummy variable, based on a survey of Dutch citizens in September 2006 (Kleinnijenhuis et al. 2007, 113), as indicated in Table 1 in the web appendix.

3.3.4 Combined model

All hypotheses can be measured 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.

3.4 *Face Validity*

As a final indicator of validity, 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 at best be interpreted as a preliminary indication of the face validity of the output of the method.

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

https://doi.org/10.1093/pan/mpn006 Published online by Cambridge University Press

analysis. This section will answer that question by presenting the results of the validity tests, answering the question ''How good/useful is this method?''

4.1 *Validity at the Level of Measurement*

The performance of the method at the article level is given in Table 1. The columns list the precision, recall, and *F*1 score of the method. The first three rows contain the performance on identifying sources, subjects, and objects. The fourth row contains the combined performance, 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 judgment on these numbers. If we treat the aggregate *F*1 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, whereas the human found another answer. In the light of these considerations, we find this performance acceptable, but there is certainly room for improvement.

4.2 *Validity at the Level of Analysis*

Where the previous measure was calculated at the article level, the construct validity of the method depends on measuring the theoretical constructs at the level of analysis. Table 2 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 are 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 canceling 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.

4.3 *Concurrent Validity at the Level of Predictions 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.

Table 1 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

Note. $N = 16{,}454$ found items (1944 sources, 9614 subject, 4896 object).

https://doi.org/10.1093/pan/mpn006 Published online by Cambridge University Press

Table 2 Correlation between method output and human codings at the level of analysis

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

Note. $N = 450$ combinations of issue × newspaper × party.

4.3.1 H1: Indexing theory

H1a: Powerful politicians will be quoted relatively often.

H1b: Powerful politicians will be portrayed as acting relatively often.

Table 3 summarizes the regression analyses of the number of seats in parliament and incumbency, 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, this effect is very small and 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.

4.3.2 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.

Table 3 Regression analysis of indexing: political power and media authority calculated using automatic and human coding

		Automatic		*Human*	
Hypothesis	*Variable*	*Estimate*	*Standard error*	*Estimate*	*Standard error*
H1a (party is source)	% of seats	0.21***	0.034	0.18***	0.032
	Incumbent	6.1***	0.90	5.8***	0.85
	Constant	0.5	0.4	0.9**	0.38
	Adjusted R^2	0.35		0.34	
H1b (party is acting)	% of seats	0.003*	0.001	0.003*	0.0012
	Incumbent	0.03	0.036	0.07	0.033
	Constant	0.63***	0.016	0.74***	0.015
	Adjusted R^2	0.02		0.02	

Note. $N = 450$ combinations of issue × newspaper × party. *$p < .05$, ***$p < .001$.

https://doi.org/10.1093/pan/mpn006 Published online by Cambridge University Press

Table 4 Regression analysis of opportune witnesses: aligned parties on media authority calculated using automatic and human coding

Hypothesis (dependent variable)	*Variable*	*Automatic*		*Human*	
		Estimate	*Standard error*	*Estimate*	*Standard error*
H2a (party is source)	Aligned party	5.6***	0.87	5.6***	0.80
	Constant	2.9***	0.47	2.9***	0.43
	Adjusted R^2	0.10		0.11	
H2b (party is acting)	Aligned party	0.05†	0.028	0.07*	0.028
	Constant	0.66***	0.015	0.74***	0.014
	Adjusted R^2	0.01		0.02	

Note. $N = 379$ combinations of issue × newspaper × party; †$p < .1$, *$p < .05$, ***$p < .001$. Note that the N is lower than 450 because one (new) newspaper does not have a traditionally aligned party.

These hypotheses were tested using univariate regression of the traditional alignment of a newspaper with the dependent variable. Table 4 shows the results of this analysis, which displays the same pattern as the results for the first hypothesis: H2a is confirmed by the strong effect of alignment on source use, whereas the evidence for H2b is much weaker, with a trend according to the output of the method and a weakly significant but small effect in the human coding.

4.3.3 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 5 lists the univariate regression model of issue ownership on 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

Table 5 Regression analysis of issue ownership on media authority calculated using automatic and human coding

Hypothesis (dependent variable)	*Variable*	*Automatic*		*Human*	
		Estimate	*Standard error*	*Estimate*	*Standard error*
H3a (party is source)	Owned issue	–0.1	1.1	0.0	1.0
	Constant	4.2***	0.4	4.2***	0.36
	Adjusted R^2	0.00		0.00	
H3b (party is acting)	Owned issue	0.05	0.035	0.06	0.032
	Constant	0.67	0.012	0.76	0.011
	Adjusted R^2	0.00		0.01	

Note. $N = 450$ combinations of issue × newspaper × party.

https://doi.org/10.1093/pan/mpn006 Published online by Cambridge University Press

that no party will be left alone on any issue (Sigelman and Buell 2004; Kaplan, Park, and Ridout 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.

4.3.4 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 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 as 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, which is one if both dummies are one, and zero otherwise, shown in the bottom part of Table 6. The interaction effect is negative and of the same magnitude as the main effect of incumbency, both of which are highly significant, whereas

Table 6 A multivariate test of indexing, opportune witnesses, and issue ownership calculated using automatic and human coding

		Automatic		*Human*	
Dependent variable	*Variable*	*Estimate*	*Standard error*	*Estimate*	*Standard error*
Party as source	% of seats	0.21***	0.033	0.18***	0.031
	Incumbent	6.3***	0.94	5.7***	0.89
	Own issue	–3.5***	0.88	–3.0***	0.82
	Aligned party	1.2	0.74	1.7*	0.70
	Constant	0.6	0.4	0.9*	0.38
	Adjusted R^2	0.38		0.36	
Party is acting	% of seats	0.0029*	0.0014	0.0028*	0.0012
	Incumbent	0.02	0.038	–0.01	0.035
	Own issue	0.02	0.036	0.04	0.033
	Aligned party	0.02	0.030	0.03	0.028
	Constant	0.63	0.17	0.73	0.015
	Adjusted R^2	0.02		0.02	
Party as source (interaction)	% of seats	0.20***	0.033	0.17***	0.031
	Incumbent	7.7***	1.0	6.9***	0.94
	Own issue	–0.5	1.2	–0.5	1.1
	Own and incumbent	–6.5***	1.7	–5.4***	1.7
	Aligned party	1.2	0.7	1.8*	0.7
	Constant	0.4	0.4	0.7	0.4
	Adjusted R^2	.039		0.38	

Note. $N = 450$ combinations of issue × newspaper × party. *$p < .05$, ***$p < .001$.

https://doi.org/10.1093/pan/mpn006 Published online by Cambridge University Press

the main effect of the owned issue is close to zero and not significant. Hence, incumbent parties are quoted more often only if the topic is not an issue owned by that party, and issue ownership does not have a direct effect. These results are interesting but should be taken with a grain of salt as they can be affected by the inclusion of valence issues in the owned issues for the incumbent parties and the fact that the incumbent parties were all relatively right winged. More importantly, for all investigated models, the ones based on computer output are highly similar to those based on human coding. Since we would draw exactly the same conclusions based on either output, and we assume the human coding to be valid, this makes it likely that the computer method can be validly used to model interactions with external variables.

4.4 *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 automatic 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 judgments 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 judgments that are preferred by human coders, but still judgments which are justifiable. It strengthens our conclusion that the method is sufficiently valid to answer questions relevant to political science.

5 Conclusion

Analysis of political communication is an important aspect of political research. Semantic Network Analysis represents this communication as semantic relations between political concepts such as actors and issues. This article showed how recent advances in computational linguistics can be harnessed to automate the extraction of Semantic Networks. In particular, grammatical analysis was used to distinguish between semantic agent and patient and to identify quotes and their sources. Actors from a political ontology were identified using their names, functions, and using simple anaphora resolution.

The performance of this method was tested by comparing its output to a manual content analysis of the Dutch 2006 elections. Performance at the level of articles was acceptable, and at the levels of analysis and prediction, it was good. Face validity, measured informally by judging a sample of the method output, was also good. These results indicate that the method can be used to automatically analyze political texts and draw valid inferences from them.

Notwithstanding those positive results, automated Semantic Network Analysis as presented here has several limitations. First of all, its performance at the level of separate sentences can be improved further. For this purpose, we conducted an error components analysis, presented in the web appendix to this article, which identified specific aspects on which to improve the system. Moreover, the identification of actors requires a detailed ontology or knowledge base of actors, including information such as name, gender,

https://doi.org/10.1093/pan/mpn006 Published online by Cambridge University Press

and political functions. This information, however, will generally be easily available for political actors and will also be relevant for the analysis of results. Finally, the current system is implemented and tested for newspaper articles in the Dutch language. The fairly general nature of the rules, which enable the identification of semantic roles and the identification of political entities in this case, suggests that it should also perform well on other texts, such as speeches, debates, or policy documents, as long as they as reasonably well structured. Moreover, the approach taken here should also be feasible for other languages, especially for similar languages such as German or English. However, this requires a high-quality syntactic parser for each of these languages, and developing and testing the extraction rules is a nontrivial task.

In sum, although the system for extracting Semantic Relations in this paper has room for improvement, the system is proven to perform acceptably, and the results are immediately useful for conducting political analysis. Moreover, we identified a number of aspects that the system could be improved on, and we are confident that the performance and usefulness of this system will increase in the future.

References

Antoniou, G., and F. Van Harmelen. 2004. *A semantic web primer.* Cambridge, MA: MIT Press.

Bennett, L. 1990. Toward a theory of press-state relations in the United States. *Journal of Communication* 40(2):103–25.

Benoit, W. 2007. *Communication in political campaigns*. New York: Peter Lang.

Bouma, G., J. Mur, and G. van Noord. 2003. Reasoning over dependency relations for QA. *IJCAI'05 workshop on Knowledge and Reasoning for Answering Questions*. Edinburgh: Scotland.

Bouma, G., G. van Noord, and R. Malouf. 2000. Alpino: wide-coverage computational analysis of Dutch. In *Language and computers, computational linguistics in the Netherlands 2000*, eds. W. Daelemans, K. Sima'an, J. Veenstra, and J. Zavrel, 45–59. Amsterdam: Rodopi.

Brandes, U., D. Fleischer, and J. Lerner. 2006. Summarizing dynamic bipolar conflict structures. *IEEE Transactions on Visualization and Computer Graphics* 12(6):1486–99.

Budge, I., and D. Farlie. 1983. *Explaining and predicting elections: issue-effects and party strategies in twenty-three democracies*. London: Allen and Unwin.

Corman, S., T. Kuhn, R. McPhee, and K. Dooley. 2002. Studying complex discursive systems: centering resonance analysis of communication. *Human Communication Research* 28(2):157–206.

Dixon, R. 1991. *A new approach to English grammar, on semantic principles*. Oxford, UK: Oxford University Press.

Donsbach, W., and O. Jandura. 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., A. Mattenklott, and H. Brosius. 1993. How unique is the perspective of television? *Political Communication* 10:41–57.

Galtung, J., and M. Ruge. 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.

Hagen, L. 1993. Opportune witnesses: an analysis of balance in the selection of sources and arguments. *European Journal of Communication* 8:317–43.

Herbst, S. 2003. Political authority in a mediated age. *Theory and Society* 32:481–503.

Jijkoun, V. 2007. *Graph transformations for natural language processing*. University of Amsterdam PhD thesis.

Kaplan, N., D. Park, and T. Ridout. 2006. Dialogue in American political campaigns? An examination of issue convergence in candidate television advertising. *American Journal of Political Science* 50:724–36.

Katz, B., and J. Lin. 2003. Selectively using relations to improve precision in question answering. Presented at the Natural Language Processing for Question Answering (EACL, April 12–17, 2003), 43–50, Budapest, Hungary.

Kleinnijenhuis, J., D. Oegema, J. de Ridder, and P. Ruigrok. 1998. *Paarse Polarisatie: De slag om de kiezer in de media*. Alphen a/d Rijn: Samson.

Kleinnijenhuis, J., O. Scholten, W. Van Atteveldt, A. Van Hoof, A. Krouwel, D. Oegema, J. De Ridder, N. Ruigrok, and J. Takens. 2007. *Nederland vijfstromenland: De rol van media en stemwijzers bij de verkiezingen van 2006*. Amsterdam: Bert Bakker.

https://doi.org/10.1093/pan/mpn006 Published online by Cambridge University Press

Krippendorff, K. 2004. *Content analysis: an introduction to its methodology.* 2nd ed. Thousand Oaks, CA: Sage Publications.

Landis, J., and G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33:159–74.

Lappin, S., and H. Leass. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics* 20(4):535–61.

Lewis-Beck, M. 2006. Does economics still matter? Econometrics and the vote. *Journal of Politics* 68:208–12.

Lijphart, A. 1975. *The politics of accommodation. Pluralism and democracy in the Netherlands*. (2nd ed., revised). Berkeley, CA: University of California Press.

Manning, C., and H. Schütze. 2002. *Foundations of statistical natural language processing*. 5th ed. Cambridge, MA: MIT Press.

Mitkov, R. 2002. *Anaphora resolution.* Harlow, UK: Longman.

Olien, C., G. Donohue, and P. Tichenor. 1983. Structure, communication and social power: evolution of the knowledge gap hypothesis. *Mass Communication Review Yearbook* 4:455–61.

Patterson, T. 1993. *Out of order.* New York: Knopf.

Petrocik, J. 1996. Issue ownership in presidential elections, with a 1980 case study. *American Journal of Political Science* 40:825–50.

Petrocik, J., W. L. Benoit, and G. J. Hansen. 2003. Issue ownership and presidential campaigning, 1952–2000. *Political Science Quarterly* 118:599–626.

Roberts, C. W. 1989. Other than counting words: a linguistics approach to content analysis. *Social Forces* 68(1):147–77.

Ruigrok, N. 2005. *Journalism of attachment: Dutch newspapers during the Bosnian war.* Amsterdam: Het Spinhuis Publishers.

Schrodt, P. A. 2001. Automated coding of international event data using sparse parsing techniques. In Annual meeting of the International Studies Association. Chicago.

Searle, J. 1969. *Speech acts*. Cambridge, UK: Cambridge University Press.

Sigelman, L., and E. Buell. 2004. Avoidance or engagement? Issue convergence in presidential campaigns. *American Journal of Political Science* 48:650–61.

Sowa, J. 2000. *Knowledge representation: logical, philosophical, and computational foundations*. Pacific Grove, CA: Brooks/Cole.

Stempel, G., and H. Culbertson. 1984. The prominence and dominance of news sources in newspaper medical coverage. *Journalism Quarterly* 61:671–6.

Van Atteveldt, W. 2008. Semantic network analysis: Techniques for extracting, representing, and querying media content. Charleston, SC: BookSurge Publishers.

Van Atteveldt, W., J. Kleinnijenhuis, N. Ruigrok, and S. Schlobach. 2008. Good news or bad news? Conducting sentiment analysis on Dutch text to dinstinguish between positive and negative relations. In *Special Issue of the Journal of Information Technology and Politics on Text Annotation for Political Science*, eds. C. Cardie and J. Wilkerson, 5(1): 73–94.

Van Atteveldt, W., S. Schlobach, and F. Van Harmelen. 2007. In *Media, politics, and the semantic web: An experience report in advanced RDF usage*, eds. E. Franconi, M. Kifer, and W. May. Berlin, Germany: Springer.

Van Noord, G. 2006. At last parsing is now operational. In *Verbum Ex Machina, Actes de la 13e conference sur le traitement automatique des langues naturelles*, eds. P. Mertens, C. Fairon, A. Dister, and P. Watrin, 20–42. Louvain-la-Neuve, Belgium: Presses Universitatires de Louvain.

Weaver, D., and G. Wilhoit. 1980. News media coverage of U.S. senators in four congresses, 1953–1974. *Journalism Monographs* 67:1–34.

Wiebe, J., T. Wilson, R. F. Bruce, M. Bell, and M. Martin. 2004. Learning subjective language. *Computational Linguistics* 30(3):277–308.

Zhang, M., G. Zhou, and A. Aw. 2008. Exploring syntactic structured features over parse trees for relation extraction using kernel methods. *Information Processing and Management* 44(2):687–701.

https://doi.org/10.1093/pan/mpn006 Published online by Cambridge University Press