To be presented at the Text as Data Conference, London, 2013

Quotes as Data
Extracting Political Statements from Dutch Newspapers by applying Transformation Rules to Syntax Graphs

Wouter van Atteveldt
VU University Amsterdam

To understand the relation between media and politics, it is necessary to study the content of politicians' statements in the news. This paper presents a method to automatically extract such statements by applying graph transformation rules to the syntactic structure of Dutch newspaper sentences. It also shows how politicians can be identified using a dictionary approach and anaphora resolution. The method is validated using manual verification, yielding good precision (86%) and recall (82%) for the extraction of quotes and decent recall (73%) for the identification of politicians. This shows that the method presented here performs sufficiently for investigating political statements in the news on a large scale.

Introduction: The power relation between media and politics

Media and politics have a complicated relationship of mutual interdependence, both at the institutional and at the individual level. Modern democratic institutions depend on the mass media to communicate policy and political positions to citizens to allow them to participate, either actively or through voting (Dahl, 1998). Individual politicians (and political parties) depend on the media to be visible to the electorate (Gans, 1979; Sheafer, 2001). Likewise, the media depends on politics as a source for a large part of its content. Individual actors (reporters but also media outlets) depend on contact with politicians in order to get scoops and in general to get enough information for their daily stories with the limited resources available to a reporter (Bennett, 1990; Bennett et al., 2007).

An area where the relation between press and politics has been extensively studied is that of Agenda Setting. From a comprehensive survey of the literature on the setting of the political agenda by the media, however, Walgrave and Van Aelst (2006) conclude that "the results are contradictory" (p.89), with scholars such as Soroka (2002) finding considerable media impact on the political agenda while others (e.g. Pritchard and Berkowitz, 1993; Kleinnijenhuis et al., 1997) find limited influence. Van Noije et al. (2008) show that the direction of influence depends on the policy area. Kleinnijenhuis et al. (2003) shows that if the aggregate agenda of political actors in the media is taken into account, the media follow this expressed political agenda during election campaigns. These mixed results are not altogether surprising: the power balance between politics and the 'fourth estate' of the media is not tipped entirely one way or another, so which party has the upper hand in a given situation depends on a variety of contingent factors. Walgrave and Van Aelst (2006) propose a model with a number of contingency factors that influence the agenda setting process, stressing the importance of five political context variables, including the institutional and political context. They also note the importance of personal characteristics of the politician, including their ability and propensity to 'play the media game and go along with the media logic' (p.103).

One factor that is left out of these studies that can explain the mixed evidence is the use of the media by politicians. Agenda setting studies investigate the content of the press and political discourse, and implicitly assume that an influence at the level of content reflects an influence or power relation at the institutional level. In essence, agenda setting studies portray politicians as almost passive objects of media attention, minding their own business in Parliament and hoping that the media take notice. In reality, however, we know that politicians use the media as a strategic tool. For example, Cook (2005) argues that politicians use the media to communicate either with their peers or their constituency as needed to fulfill their (policy or electoral) goals. Likewise, Wolfsfeld (1997) sees the contest over the media as part of the general struggle for political control. As stated by Sheafer (2001), politicians who "invest their creativity, initiative, and energy" in Parliamentary activity rather than in playing the media game, "are missing the point or wasting their time" (p.730).

If politicians use the media as one of the tools at their

disposal, we need to examine this use of the media as a platform to understand the power relation between press and politics. We have to figure out whether politicians are used by the media as a source to fit their format, as suggested by media logic (Strömbäck, 2008); or whether politicians use the media to further their own agenda. In the words of Wolfsfeld and Sheafer (2006), we need to examine the issue of "who drives the news" (p. 350).

Wolfsfeld and Sheafer (2006) show how powerful and charismatic politicians can 'ride' media waves to get media attention. Other studies confirm that well-connected political actors are best able to gain media attention (Tresch, 2009; Schönbach et al., 2001; Sellers and Schaffner, 2007). It stands to reason that those more powerful or better connected actors will also be able to choose the context and content of their media statements. However, to the knowledge of the author this has not been investigated quantitatively. Moreover, to understand the role that these statements play in the larger media and political discourse, it is necessary to investigate to what extent the content of these statements can influence the other media and political content.

Due to the necessary reciprocal relations between media content and political statements inside and outside the media, finding such influences requires time series analysis with relatively large data sets. Therefore, it would be highly beneficial to have a method for automatically identifying and extracting quoted or paraphrased statements by politicians. This paper presents a method than can extract such information from newspaper articles using grammatical analysis. By manually verifying the automatically extracted statements, it shows that the performance of this method is sufficient for use in the large scale analyses necessary for studying the role of such statements in political communication.

Automatically Extracting Quotes

This paper uses rule-based transformation of the grammatical dependency structure of Dutch newspapers sentences in order to extract quotes and identify their sources. The main procedure for the extraction and identification of quotes and sources is as follows:

1. Sentences are parsed into a dependence graph and this graph is preprocessed to add lexical markers and resolve conjunctions.
2. Syntactic patterns are used to identify quotes and their sources.
3. Multi-line and full-sentence quotes are identified using a regular-expression based procedure.
4. Politicians are identified in the identified sources using dictionary look-up and anaphora resolution.

The remainder of this section will elaborate upon each of these steps.

Parsing and preprocessing

The first step in the processing of sentences is syntactic parsing. In this paper, the Dutch HPSG parser Alpino is used, which can represent the grammatical structure of a sentence in a dependency graph (Van Noord, 2006). In a dependency graph, each node represent a word, and the edges express grammatical dependency relations between the nodes. For example, the dependency structure of the sentence "John loves Mary" would have the verb 'love would be the root of the dependency tree, with John having a (grammatical) subject relation to 'love' while Mary has a (grammatical) direct object relation.

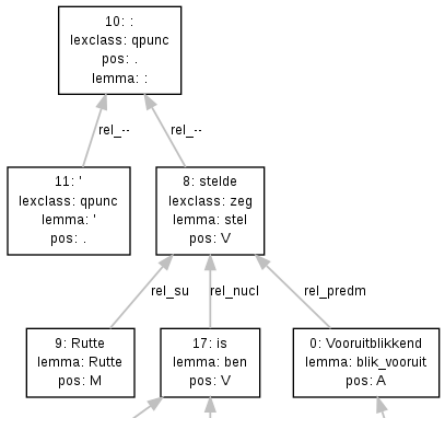
Before the actual processing starts, the dependency graphs are first preprocessed in two ways. First, the graph is enriched with lexical information, marking words as belonging to two predefined lexical categories: direct speech verbs, such as to say and to state, which indicate direct speech by a source, and attribution words, such as 'according to', which indicate indirect speech. Second, minor graph transformations are used to e.g. resolve conjunctions and to add lexical markers to multi-word speech words such as 'laten weten' (let is be known that): Neither word is by itself a speech word, but taken together they are a common way of marking a paraphrase.

Both the preprocessing described here and the actual processing described below was conducted using graph transformations. First, the dependency graph and all information about the words was translated into an RDF graph (Antoniou and Van Harmelen, 2004). In RDF, a graph description language developed in the context of the Semantic Web, graphs are represented as subject, predicate, object triples. Each grammatical dependency was represented as a triple, and moreover the word, lemma, and Part-of-Speech of each word was represented with triples with a string literal as object. This allows the lexical enrichments described above to be seen as a graph transformation, as adding a lexical marker entails inserting a new triple pointing to the string literal representing the lexical class. These graph transformations are represented as SPARQL 1.1 UPDATE statements (Prud'hommeaux and Seaborne, 2006), and executed using the Fuseki engine[1].

Citation Patterns

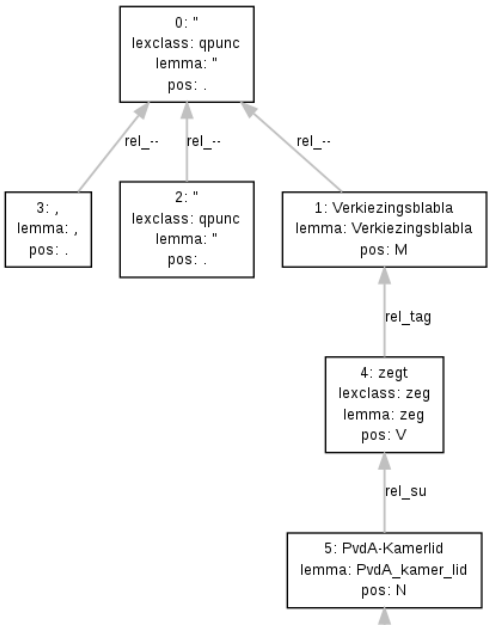
In order to process the variety of ways in which journalists can quote or paraphrase politicians, six syntactic patterns were developed. These patterns are presented and explained below, starting with the most explicitly

[1] `http://jena.apache.org/documentation/serving_data/`



Vooruitblikkend naar de Tweede Kamerverkiezingen van 12 september stelde Rutte: 'Een stem op de PVV is een verloren stem' (*Reflecting on the parliamentary elections of September 12, Rutte stated: 'a vote on the PVV is a lost vote*). Source: Openlijke verwijten luiden verkiezingscampagne in, de Volkskrant, 2012-05-14

Figure 1. Example for Pattern 1: Direct quote



"Verkiezingsblabla", zegt PvdA-Kamerlid Kuyken (*"Electoral blabber", says PvdA MP Kuyken*). Source: 130-plan krijgt tegengas, De Telegraaf, 2012-07-21

Figure 2. Example for Pattern 2: Sentence final quote

marked quotes. For each pattern, an example structure is shown alongside the SPARQL code that is used to identify quotes conforming to that pattern.

(1) Direct Quotes. *S says: 'Q'*

The most straightforward way to express a quote is to state explicitly that the source is saying something, followed by a colon and/or a text enclosed by quotation marks.

For example, take the sentence fragment *Rutte stated: 'a vote on the PVV is a lost vote*[2]. The relevant part of the dependency parse tree of this sentence is displayed in Figure 1.

The main verb *stelde* (to state) has lexical class *speech*, identifying it as an explicit speech marker. Although the quotation marks are represented by Alpino as a single mark at the top of the tree, this provides sufficient evidence that the nucleus of the verb is a quote stated by the subject of the verb. More formally, this pattern is represented in SPARQL as follows:

```
?speech :lexclass "speech";
   :rel_-- [?lexclass "qpunc"].
?quote :rel_nucl ?speech.
?source :rel_su ?speech
```

SPARQL patterns are basically lists of triples to be found in the graph. A node name preceded by a question mark, such as `?speech`, is a variable that can be reused. In this case, we are looking for a node which has lexical class 'speech', and which has a punctuation relation (`rel_--`) with a node marked as *qpunc*, or quote punctuation. Finally, we look for the nucleus and subject of this node, and mark these as the quote and the source.

(2) Sentence-final quote. *Q, says S*

A variation on the first pattern is where the quotation occurs at the end of the sentence. For example, take the sentence *"Electoral blah blah," says PvdA MP Kuyken.* The parse for this sentence is shown in Figure 2

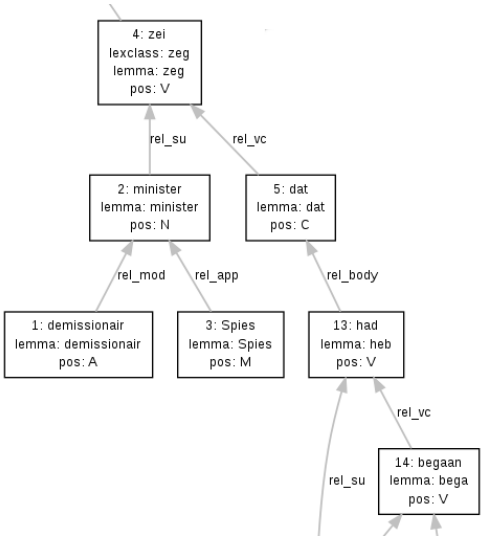
Although superficially this resembles the first pattern, in Alpino it is represented differently, with the speech verb marked as a discourse `tag` of the head of the quote. The pattern to find these is relatively simple: we look for a speech verb which is the tag of the quote, the source of the quote being the grammatical subject of the speech verb:

```
?speech :lexclass "speech".
?speech :rel_tag ?quote.
?source :rel_su ?speech
```

(3) Paraphrase. *S says that Q*

In the third pattern, the quote is indicated with an overt lexical speech marker but without the supporting punctuation. This generally indicates a paraphrase rather than a real citation. Since the content of the paraphrase is still attributed to the source, this distinction is

[2]See Figure 1 for the Dutch original and attribution



[..] minister Spies zei dat de PvdA een 'kapitale blunder' had begaan [..] (*[..] minister Spies said that the PvdA had made a 'capital blunder' [..]*). Source: see Figure 1

Figure 3. Example for pattern 3: Paraphrase

currently ignored. For example, take the sentence fragment *[..] minister Spies said that the PvdA had made a 'capital blunder'* Figure 3 shows the parse tree for this fragment of the sentence. Note that this sentence contains a paraphrase with an embedded literal quote; this embedded quote is ignored in the current version of the rule set.

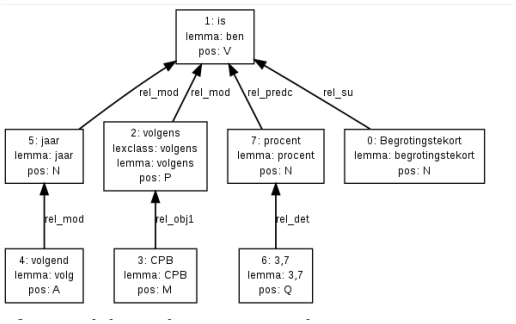
To detect such paraphrases, we look for a speech verb that has a verbal complement (`vc`) with the lemma "dat" (that). The body of the complementizer is the quote, while the subject of the speech verb is the source:

```
?speech :lexclass "speech".
?source :rel_su ?speech.
?quote :rel_body ?that.
?that :lemma "dat"; :rel_vc ?speech
```

(4) Attribution. *Q, according to S*

A fourth pattern is used for attributing a quote using an explicit paraphrasing marker such as 'according to'. For example, in the sentence *According to CPB, the deficit will be 3.7%*, the claim about the deficit is attributed to the planning agency CPB using the marker 'according to' (volgens). In the Alpino parse, such as shown in Figure 4, the source is the direct object (obj1) of the attribution marker (volgens), which is itself either a modifier or discourse tag of the main sentence verb, which is marked as the quote:

```
?according :lexclass "volgens".
?source :rel_obj1 ?according.
```



Begrotingstekort is volgens CPB volgend jaar 3,7 procent (*According to CPB, the deficit will be 3.7% next year*). Source: ?

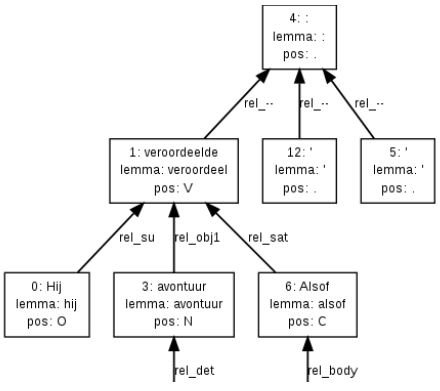
Figure 4. Example for pattern 4: Attribution

```
?according :rel_mod|:rel_tag ?quote
```

(5) Indirect quote:. *S was clear: 'Q'*

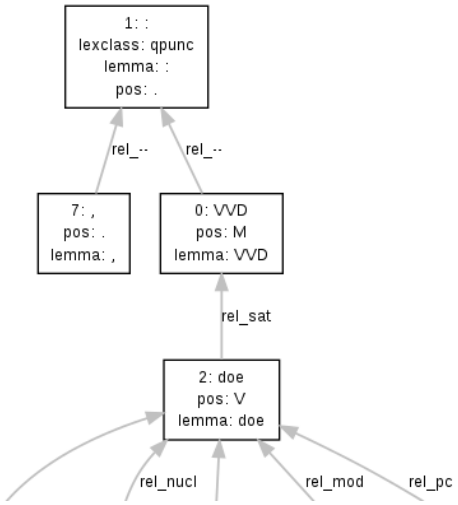
The final two patterns are more indirect, relying on the use of a colon to start the quote without an explicit speech or attribution verb. For example, take the sentence *He condemned the adventure: 'as if the Netherlands are a political laboratory*, as shown in Figure 5. The verb *to complain* is not an explicit speech marker, although it is semantically related. Such verbs are used indirectly to lead in the quote after the colon.

In the dependency tree, such sentences are marked by a quote-punctuation (colon or quotation mark), with the main verb a direct child of the punctuation. The source and quote are the subject and satellite discourse unit (`sat`), respectively. In the current pattern, the lexical class of the main verb is not checked:



Hij veroordeelde het avontuur: 'Alsof Nederland een pollitiek laboratorium is' (*He condemned the adventure: 'as if the Netherlands are a political laboratory'*). Source: See Figure 1

Figure 5. Example for pattern 5: Indirect quote



VVD: doe recht aan alle werkenden, betaald of onbetaald (*VVD: consider all workers, paid or unpaid*). Source: Letter to the editor, De Volkskrant, 2012-11-06

Figure 6. Example for pattern 6: Colon-quotes

```
?source :rel_su ?top.
?quote :rel_sat ?top.
?top :rel_--/:lexclass "qpunc"
```

(6) Colon-quote:. *S: Q*

The final pattern seems the simplest: the speaker (name or noun phrase) followed by a colon and a quote. For example, take the sentence *VVD: Consider all workers, paid or unpaid* as shown in Figure 6 (note the lack of quotation marks in this example). In such sentences, the part before the colon does not have a main verb, causing the quote to be attached to the source directly, in this case as a satellite. This leads to the following SPARQL query:

```
?source :rel_-- [:lexclass "qpunc].
?quote :rel_nucl|:rel_sat ?source.
FILTER NOT EXISTS {?source :pos "."}
FILTER NOT EXISTS {[] :quote ?quote}
```

The source is a direct dependent of the quotation punctuation, and the quote is either the nucleus or sattelite discourse unit of the source. The final two sentences rule out mistakes where the source is itself punctuation, or when the quote is already identified by another pattern. These filters are needed since there is no explicit 'key' for the pattern such as normally formed by the speech or attribution verb.

Multi-line quotes

The patterns listed above are used to find quotes that are contained within a sentence. Often, however, quotes can span multiple sentences. This can occur after a normal quote, i.e. where the quote does not end with the sentence. For example, take the sentences: *Wilders complained about Rutte: 'if you do what he says, he puts his arm around you. If you choose for the Netherlands and hold your ground, you are threatened and intimidated'*.[3] In this example, the second sentence is part of the quote started in the first sentence.

In other cases, the quote is not part of a syntactic pattern as discussed above. For example, consider the sentence *Wilders rejected Rutte's attack that a vote on the PVV would be a lost vote. 'The PVV will be among the largest parties'*.[4] In this example, the second sentence forms a full sentence quote, with the subject of the previous sentence (Wilders) as its source.

Both these cases are dealt with using a character-based approach. If a sentence has an open quote, the following sentences are included until the quote is closed. A quote can be left open either by a syntactically found quote that misses a closing quotation mark, or by a sentence starting with a quotation mark that does not contain a syntactic quote. If a multi-line quote started with a syntactic quote, the source of that quote is the source of the following lines as well. If a multi-line quote or full sentence quote does not follow a syntactic quote, the subject of the previous sentence is seen as the source. In the first example above, the quote 'If you choose ... intimidated' is attributed to Wilders since he is the source of the syntactic quote in the preceding sentence. In the second example, Wilders is also seen as the source of the quote 'the PVV will be among the largest parties', in this case because he is the subject of 'to reject' in the preceding sentence.

Recognizing sources

After the quotes have been identified, it is necessary to determine the identity of the sources. In the target domain of this paper, the set of sources is a closed list of politicians or other actors that are known beforehand. This makes it easy to use dictionary based methods to recognize the sources. After that, anaphora resolution is used to identify later anaphoric references to the recognized sources.

The dictionary-based look-up is primarily based on the last name of the politician. Since sometimes last

[3]Vervolgens beklaagde Wilders zich over Rutte: 'Als je doet wat hij zegt, slaat hij een arm om je heen Als je voor Nederland kiest en je rug recht houdt, dan word je bedreigd en geÃŕntimideerd.' Source 'Opgewekte Rutte ware verademing', De Telegraaf, 2012-09-13

[4]Ruttes aanval dat een stem op de PVV een verloren stem is, verwierp Wilders 'De PVV wordt de grootste of een van de grootste partijen.' Source: See Figure 1

names are not unique (for example, the current Dutch prime minister (Mark Rutte) has the same surname as Arno Rutte, an MP of the same party. To resolve this, the first name, party name, or political function of the politician is required to occur within within 5 words of the last name at least once in the article. This reflects the habit of journalists to use a full description of a source the first time, and using only the last name or an anaphoric reference in further mentions.

Many references in newspaper articles are anaphoric. For English, the Stanford parser has a built-in coreference resolver. For Dutch, however, there is no off-the-shelf anaphora system. For the existing systems published by Mur and van der Plas (2006), to the knowledge of the author no current implementations exist.

To overcome this, we apply a simplification of the anaphora resolution algorithm published by Lappin and Leass (1994). For each reference, a list of candidate referents is made by looking at the previous five sentences for proper names, where preference is given to names in the subject position. This preference stems from the general preference for the reference and referent to be in the same grammatical position, and the source of a citation is generally in the subject position. If the candidate referent is a known politician, a final check is made that the genders of the referent and the anaphora match.

As an example, consider the sentences: *From the party convention ChristenUnie leader Arie Slob called on the VVD and CDA to distance themselves from the PVV. He condemned the adventure: 'as if the Netherlands are a political laboratory'*.[5] In the second sentence, 'he' is the source of the quote. To resolve this, the system looks at the previous sentence, which has mentions of Arie Slob and the party names ChristenUnie, VVD, CDA and PVV. Since Slob is the only name in the subject position, and since Slob and the pronoun 'he' as both masculine, Slob is (correctly) chosen as the referent.

Validation

The methodology described in this paper was developed in the context of an ongoing substantive inquiry into the role of statements by politicians in the media discourse. For this reason, the methods were validated by manually verifying the automatic analysis on a sample of newspaper articles that contain at least one reference to a political party or party leader. Although the methods presented above are essentially sentence-based, the anaphoric references and multi-line citations made it necessary to consider the context of sentences as well. The validation presented here are based on 307 successive sentences from 10 articles randomly selected from Dutch national newspaper articles in the period of April 2012 to June 2013.

For these sentences, the two outcomes of the methodology were verified independently. First, it was judged whether the source and quote were extracted correctly. Second, in those sentences where a source was extracted by the computer, it was judged whether the computer correctly identified the political actor (if any) that the source referred to, directly or through anaphoric reference. The latter verification was explicitly limited to national politicians (MPs and cabinet members) and party names.

For both validation steps, the performance of the system is expressed using the metrics *precision* and *recall*. These metrics are standard in the fields of information retrieval and extraction (Manning and Schütze, 2002). Both methods assume that the gold standard contains a set of items that need to be extracted. The *recall* of the method is then the proportion of actual items that were correctly extracted. Its twin measure *precision* is the proportion of extracted items that was correct.

More formally, we can define the metrics as follows for extracting the quotes (and analogously for identifying the source). Let TP (True Positives) be the number of sentences where the quote was correctly found by the system; FP (False Positives) the number of sentences where an incorrect quote was extracted; and FN (False Negatives) the number of sentences where a quote was missed by the system. Then, precision and recall can be calculated as follows:

$$\begin{aligned} Precision &= \frac{TP}{TP+FP} \\ Recall &= \frac{TP}{TP+FN} \end{aligned}$$

In manual content analysis, reliability is generally measured using inter-coder reliability metrics such as Cohen's kappa (1960) or Krippendorff's alpha (2004). For determining the validation of automatic content analysis methods, the precision and recall metrics described above are superior for at least two reasons. First, inter coder reliability metrics assume two equivalent coders, where in validating automatic content analysis we are comparing the machine 'coder' to a gold standard that we assume is correct. By presenting precision and recall separately rather than a single number, it can be seen whether the automatic coding is too strict or too lenient, which can tell us something about possible biases

[5] Vanaf zijn partijcongres riep ChristenUnie-lijsttrekker Arie Slob zaterdag VVD en CDA op zich te distantiÃńren van de samenwerking met de PVV Hij veroordeelde het avontuur: 'Alsof Nederland een politiek laboratorium is.' Source: see Figure 1

Table 1
Performance of quote extraction & source identification

Metric	Quote extraction	Source identification
True positive	93	16
False Positive	15	0
False Negative	21	6
Precision	0.86	1.00
Recall	0.82	0.73

Table 2
Performance of quote extraction for different quote types

Metric	Quote (1 & 2)	Paraphrase (3 & 4)	Colon (5 & 6)	Multi-line
True Positive	10	9	22	4
False Positive	1	3	3	3
False Negative	1	0	10	2
Precision	0.91	0.75	0.88	0.57
Recall	0.91	1.00	0.69	0.67

in the end result. Second, in a task such as the present the true negatives, i.e. cases where the computer correctly identified that no source was present, are not very informative. By not taking true negatives into account but rather looking explicitly at errors of omission and commission, precision and recall give a more informative estimate of the performance of the automatic method.

It has to be noted that this verification should probably be enlarged, especially as only 22 political actors were present as quotes in the investigated sentences. However, the total number of sentences is large enough to give a good first estimate of the validity of the system.

Outcomes

The performance of the quote extraction and source identification are listed in Table 1. For the quote extraction, both precision and recall are above 80%, with precision slightly higher than recall. This shows that the system performs well for extracting the source and the quote of sentences. For source identification, only 22 of the 108 (correctly and incorrectly) extracted sources contained either an actual or an identified national political actor. From this set, 16 were identified correctly while 6 references were missed, giving a recall of 75% and a precision of 100%.

Error analysis

When doing the manual verification, the real type of each quote was also determined for a subset of the sentences. This used a slightly different set than the grammatical patterns presented above, combining patterns 1 and 2 (quotes), 3 and 4 (paraphrases) and 5 and 6 (implicit quotes using colons). Also, cases where a multi-line or whole sentence quote was present were recorded. Table 2 shows the performance of the system for each of these types. Unsurprisingly, the system has much more trouble with the more implicit quotes using colons than with the sentences containing more explicit quote markers. For the explicit quotes, the errors that were present were largely lexical. For example, the word 'melden' (to report) was used as the speech marker in two cases, but this word was not present in the lexicon.

Especially the recall of the colon-quotes was low, indicating that the current set of rules is too strict. The reason for the current strict rules is that, although colons often indicate quotes, they can also be used the mark the second part of the sentence as an example, and the use of quotation marks is not always consistent. Moreover, although the main verb of the first part is often speech related, such as complain or deny, the class of 'speech related' verbs is much larger and opener than the class of direct speech verbs. That said, it would probably be good to use a list of speech-related verbs, e.g. from WordNet or EuroWordNet (Miller, 1995; Vossen, 1999), and have more lenient rules in case such a verb is found. Apart from this, a small error was found where single-word names followed by a colon were not correctly identified as a source.

For the multi-line quotes, both precision and recall were problematic. This makes sense since the current implementation relies on the correct use of quotes for both starting and ending a multi-line quote. Even if journalists are correct in their use of quotes (which they usually are), tokenization and encoding problems cause problems here, with quotes dropping to the next sentence or being encoded as commas. Sometimes, the quote symbols simply disappear due to Unicode problems. These mainly technical difficulties should be sorted out before more substantive gains can be made. Interestingly, human readers have little problem identifying the start and end of a quote, mainly because of stylistic and/or pragmatic clues. Although capturing these clues into an algorithm will be difficult, some sort of machine learning might be beneficial here since the 'neat' quotes can be identified quite easily, and it can be expected that language use in quotes will be different from the language used in the rest of the article.

For the source identification, a formal error analysis was largely useless due to the low count. Moreover, visual inspection quickly revealed that all 6 errors were caused by dictionary problems: the extracted source

contained the last name of a known politician, but the politician was not identified because the text did not fulfill the disambiguation criteria described above: none of the mentions of the last name in the text also contained the first name, party name, or function within five words. Combined with the high precision of the method, although calculated on a small sample, this suggests that the disambiguation should be made more lenient.

Conclusion and Discussion

This paper demonstrated a method to automatically extract quoted and paraphrased statements of politicians from Dutch newspaper articles. The method uses rule-based graph transformation of the syntactic structure to extract the quotes and their sources. Political actors are then identified using a dictionary based method augmented with anaphora resolution. The method is shown to have sufficient accuracy, with precision and recall being over 80% for quote extraction and over 70% for identification. A qualitative analysis of the found errors showed a number of avenues for improving this performance.

The method is developed and tested using Dutch newspaper articles and a Dutch syntax parser. Moreover, the transformation rules are language- and parser-specific, both in the specific grammatical patterns that are looked for and in the way that these patterns are represented by the parser. However, the methodology used here should be easily transplanted to other languages with a similar grammatical structure and good wide-coverage automatic parsers. In fact Sheafer et al. (2013) show how similar transformation of syntactic structures can be used to identify the framing of conflict in English language newspapers.

Besides differences between languages, syntactic patterns can also be expected to vary greatly between different domains. While news content can be expected to be fairly regular and grammatically correct, no such guarantees can be given for other textual content such as personal communication or social media. Although automatic parsers are getting better every year, less regular language use in the target domain will necessarily translate to lower performance of a rule based method.

Another limitation of the method is that developing the rules is time-consuming and requires expertise on the grammatical structure of the target language. Moreover, as the rules become more complex it becomes more difficult to understand, maintain, and improve the rules. Also, due to the variations in grammatical structure between different languages and parsers, it will be difficult to maintain a single set of rules, which means that for each language to be studied a separate rule set needs to be developed and maintained. Once this work is done, however, the method can be used to analyze an arbitrary amount of text.

It should be noted that both the automatic parsing and graph transformations are computer-intensive tasks. For a long sentence, the process can take up to a minute to complete. Fortunately, since all articles can be processed in parallel, it is easy to use a number of cores or even computers to process texts quickly. For an ongoing substantive study using this method, over 23,651 articles containing 935,806 sentences were processed. The parsing was conducted on the SurfSARA LISA supercomputing cluster[6], while the graph transformations were done on a single core on a modern computer. Both tasks were finished in around 24 hours, showing that processing time is not a barrier to large scale analysis using this method.

The importance of this method lies mainly in enabling large-scale quantitative research into the role of politicians' media statements in the broader political discourse. As such, the real validation of this method will be its successful application to a problem that was difficult to solve without the method. The first step in doing this will be an investigation of the agenda dynamics of the statements by different actors and the remainder of the media content. Do (powerful) politicians choose what they want to say in the media, or do journalists select a quote to support the point he or she is making? Do the statements of politicians decide the flow of the media discourse, and does this translate to political discourse and, ultimately, policy? The method presented in this paper is a first step to finding en empirical answer to these tantalizing questions.

References

Antoniou, G. and Van Harmelen, F. (2004). *A Semantic Web Primer*. MIT Press, Cambridge, Ma.

Bennett, L. (1990). Toward a theory of press-state relations in the United States. *Journal of Communication*, 40(2):103–125.

Bennett, W., Lawrence, R., and Livingston, S. (2007). *When the press fails: Political power and the news media from Iraq to Katrina*. University of Chicago Press.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

Cook, T. (2005). *Governing with the news: The news media as a political institution (2nd ed.)*. University of Chicago press.

[6] `http://lisa.sara.nl`

Dahl, R. (1998). *On Democracy*. Yale University Press, New Haven.

Gans, H. J. (1979). *Deciding what's news: A study of CBS evening news, NBC nightly news, Newsweek, and Time*. TriQuarterly Books.

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., Oegema, D., de Ridder, J. A., Van Hoof, A. M. J., and Vliegenthart, R. (2003). *De puinhopen in het nieuws*, volume 22 of *Communicatie Dossier*. Kluwer, Alphen aan de Rijn (Netherlands).

Krippendorff, K. (2004). *Content Analysis: An Introduction to Its Methodology (second edition)*. Sage Publications, Thousand Oaks, CA.

Lappin, S. and Leass, H. (1994). An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–561.

Manning, C. and Schütze, H. (2002). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, fifth printing edition.

Miller, G. (1995). *WordNet: a lexical database for English*. ACM Press, New York.

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)*.

Pritchard, D. and Berkowitz, D. (1993). The limits of agenda-setting: The press and political responses to crime in the United States, 1950-1980. *International Journal of Public Opinion Research*, 5(1):86.

Prud'hommeaux, E. and Seaborne, A. (2006). SPARQL query language for RDF. W3C Recommendation.

Schönbach, K., de Ridder, J., and Lauf, E. (2001). Politicians on tv news: Getting attention in dutch and german election campaigns. *European Journal of Political Research*, 39:519–31.

Sellers, P. J. and Schaffner, B. F. (2007). Winning coverage in the u.s. senate. *Political Communication*, 24:377–91.

Sheafer, T. (2001). Charismatic skill and media legitimacy. *Communication Research*, 28(6):711–736.

Sheafer, T., Shenhav, S., Takens, J., and Van Atteveldt, W. (2013). Relative political and value proximity in mediated public diplomacy: The effect of state-level homophily on international frame building. *Accepted for publication in Political Communication*.

Soroka, S. (2002). *Agenda-setting dynamics in Canada*. University of British Columbia Press.

Strömbäck, J. (2008). Four Phases of Mediatization: An Analysis of the Mediatization of Politics. *The International Journal of Press/Politics*, 13(228).

Tresch, A. (2009). Politicians in the media: Determinants of legislators' presence and prominence in swiss newspapers. *The International Journal of Press/Politics*, 14.

Van Noije, L., Kleinnijenhuis, J., and Oegema, D. (2008). Loss of Parliamentary Control Due to Mediatization and Europeanization: A Longitudinal and Cross-Sectional Analysis of Agenda Building in the United Kingdom and the Netherlands. *British Journal of Political Science*, 38(03):455–478.

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 Universitatires de Louvain.

Vossen, P., editor (1999). *EuroWordNet: a multilingual database with lexical semantic networks for European languages*. Kluwer, Dordrecht, the Netherlands.

Walgrave, S. and Van Aelst, P. (2006). The contingency of the mass mediaâĂŹs political agenda-setting power: Towards a preliminary theory. *Journal of Communication*, 56:88–109.

Wolfsfeld, G. (1997). *Media and political conflict: News from the Middle East*. Cambridge University Press, Cambridge, U.K.

Wolfsfeld, G. and Sheafer, T. (2006). Competing actors and the construction of political news: The contest over waves in Israel. *Political Communication*, 23(3):333–354.