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

Wouter van Atteveldt
VU University Amsterdam

Abstract
To understand the relation between media and politics, it is necessary to study the content of politicians’ statements in the news. By using syntactic analysis and topic models, this paper looks at how often politicians are quoted, and whether their media statements are similar to their statements in parliament. While media attention simply follows political power, this is quite different for media statements. The frequency of statements is a matter of journalistic demand (e.g. high during scandals) and political supply (e.g. low during closed-door negotiations). Media statements are most similar to political discourse during the campaign, and for limited-issue parties. Some interesting results were found, with the anti-immigration PVV being relatively dissimilar during the campaign, and possible coalition partners being relatively dissimilar during the coalition talks. This paper is a promising first step into the relatively understudied area of mediated politics.

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. 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. 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 (Strömback, 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.

In this paper I show how syntactic analysis can be used to automatically identify and extract the quotes of politicians in newspaper articles. I apply rule-based graph transformations to syntax graphs to automatically extract quotes, and use a combination of keyword and anaphora resolution to identify political actors. This ‘segments’ the media content into fragments linked to specific politicians and a residual media discourse. In these documents,
topic modeling is used to find out whether the topics used by politicians are more similar to the topics these politicians use in parliament (suggesting a mainly political logic) or to the topics used in the recent (residual) media discourse (suggesting media logic).

**Hypotheses**

The correlation between the topics used by politicians in parliament and the media is analysed before, during, and after the 2012 Dutch parliamentary elections. This allows me to compare the mediated political discourse during elections, cabinet formation, and ‘routine’ politics. Since a different cabinet was formed after 2012 (albeit with the same prime minister and senior coalition partner) it also allows me to compare the discourse in opposition and coalition for the two parties switching roles. Following Hopmann et al. (2012), I expect that during the elections politicians are quoted more often and have more opportunity to discuss their own topics rather than react to outside event:

**H1a:** During election periods politicians are quoted in the media more frequently (relative to total media coverage).

**H1b:** The correlation between politicians’ parliamentary and media statements is stronger during election periods.

The Dutch Parliament has a highly fragmented multi-party system with coalition governments. Parties in the coalition are more newsworthy, since they have more say in the policy process. On the other hand, these parties are less free in their discourse as they are bound to defend the compromises laid down in the coalition agreement, while opposition parties are free to react symbolically to further their own interest (cf Walgrave and Van Aelst, 2006). For this reason, I expect coalition parties to be less able to use their own preferred topics in the media:

**H2a:** Members of coalition parties are quoted in the media more frequently (relative to total media coverage).

**H2b:** The correlation between politicians’ parliamentary and media statements is weaker for members of coalition parties than for opposition parties.

A number of studies confirm that more powerful politicians in general receive more media attention (Schönbach et al., 2001; Tresch, 2009). I expect these politicians to also have more success at ‘playing the media game’, being quoted more often and getting their preferred topics and frames into the media discourse (Wolfsfeld, 1997):

**H3a:** Parties with more seats in parliament are quoted more frequently (relative to total media coverage).

**H3b:** The correlation between politicians’ parliamentary and media statements is stronger for members of parties with more seats in parliament.

**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.

 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 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.

1http://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

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:

```sparql
?speech :lexclass "speech";
   :rel_-- [?lexclass "qpunc"].
?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 mark.

2See Figure 1 for the Dutch original and attribution
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".
?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 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".
?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.
?according :rel_mod|:rel_tag ?quote

Figure 2. Example for Pattern 2: Sentence final 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:
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

?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"].
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 satellite 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’. 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’. 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

---

3Vervolgens 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 geibernooid.’ Source ‘Opgewekte Rutte ware verademing’, De Telegraaf, 2012-09-13

4Ruttes 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

Figure 5. Example for pattern 5: Indirect quote
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 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 co-reference 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’. 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.

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

$$\text{Precision} = \frac{TP}{TP + FP}$$
$$\text{Recall} = \frac{TP}{TP + FN}$$

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 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%.
Table 1

Performance of quote extraction & source identification

<table>
<thead>
<tr>
<th>Metric</th>
<th>Quote extraction</th>
<th>Source identification</th>
</tr>
</thead>
<tbody>
<tr>
<td>True positive</td>
<td>93</td>
<td>16</td>
</tr>
<tr>
<td>False Positive</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>False Negative</td>
<td>21</td>
<td>6</td>
</tr>
<tr>
<td>Precision</td>
<td>0.86</td>
<td>1.00</td>
</tr>
<tr>
<td>Recall</td>
<td>0.82</td>
<td>0.73</td>
</tr>
</tbody>
</table>

Table 2

Performance of quote extraction for different quote types

<table>
<thead>
<tr>
<th>Metric</th>
<th>Quote (1 &amp; 2)</th>
<th>Paraphrase (3 &amp; 4)</th>
<th>Colon (5 &amp; 6)</th>
<th>Multi-line</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Positive</td>
<td>10</td>
<td>9</td>
<td>22</td>
<td>4</td>
</tr>
<tr>
<td>False Positive</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>False Negative</td>
<td>1</td>
<td>0</td>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>Precision</td>
<td>0.91</td>
<td>0.75</td>
<td>0.88</td>
<td>0.57</td>
</tr>
<tr>
<td>Recall</td>
<td>0.91</td>
<td>1.00</td>
<td>0.69</td>
<td>0.67</td>
</tr>
</tbody>
</table>

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 to 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.

**Modeling topics in quotes and speeches**

In order to compare the content of media statements to those in parliament, a topic model was created using Latent Dirichlet Allocation (LDA; Blei et al., 2003). Topic modeling is a form of dimensionality reduction somewhat similar to Latent Semantic Analysis, but it works by bayesian estimation of a generative model. This model assumes that texts are created by selecting a mixture of topics for each text, and drawing words from the mixture of words belonging to each topic. In the estimation or model learning phase, the topics-per-document and words-per-topic documents are estimated that are most likely given the data.

Out of 22,891 newspaper articles, 7,929 quotes from politicians were extracted. These were combined with a random sample of 9,760 speeches in Parliament. All texts were parsed using the Alpino parser, and all lemmata (word stems) were selected that occurred at least 20 times, occurred in less than 5% of the documents, and that belonged to a substantive part of speech (i.e. verbs, nouns, adjectives, adverbs). In all, this yielded a term-document matrix of 2,694 lemmata by 18,231 documents. From this matrix, a topic model with 50 topics was created using the R `lda` package.

On visual inspection, the resulting topics are an interesting mix of very substantive topics and more ‘procedural’ language related to the workings of parliament and media statements.

Figure 7 shows a word-cloud visualization of four topics that are representative of this mix. Figure 7(a) is a purely substantive topic expressing a single issue: the housing market. Although the words are in Dutch, the top words are *corporations* (project housing), *woning* (house), *huurder* (tenant), and related words. Figure 7(b) is also substantive, dealing with the tax service. This topic, however, expresses a specific event or frame, namely the problems
Figure 7. Word clouds representing a selection of LDA topics in media and parliament:

(a) Housing
(b) Tax service
(c) Procedural (Parliament)
(d) Statements
with tax credit fraud. Top words here are *fraude*, *belastingdienst* (tax service), *fout* (error), *pak_aan* (handle), and *rapport* (report).

The next two topics are more procedural in nature. Figure 7(c) is a topic that captures specific word use in parliament. Top words are *voorstel* (proposal), *antwoord* (reply), and *termijn* (term). Finally, Figure 7(d) is a topic that contains words that are specific to making media statements such as *niets* (nothing), *helemaal* (totally), *neem* (take), and *gebeurd* (happened).

In LDA, the substantive and procedural frames can be separated because each document is a mixture of different topics. Thus, a parliamentary speech on housing policy will be a mixture of the housing policy topic and the parliamentary procedural topic. This feature of LDA topic models is extremely useful in the current paper as it separates the substantive topics from the more procedural topics specific to media and parliament. This makes it useful to compare topic use between the two sets of discourse.

**The 2012 parliamentary elections in the Netherlands**

The Dutch parliamentary system is a proportional representation system with a very low threshold of 0.67% or one seat in the 150 seat lower chamber. This results in fragmented party system with a high effective number of parties of 6.7 before the 2012 elections and coalition governments consisting of upwards to five different parties. Before the elections, the far-right anti-immigrant party PVV supported a minority cabinet consisting of the convervative VVD and christian democratic CDA parties. When it became clear that a number of financial setbacks would necessitate extra budget cuts, new negotiations ensued between the three parties. On the 21st of April, it became clear that the negotiations had failed and the PVV decided to withdraw their support for the coalition, leading to new elections planned for September 12.

It was initially predicted additional losses for the traditionally large parties VVD, CDA and PvdA (social democrats). However, after a relatively short election campaign dominated by the horse race between VVD and PvdA both parties gained strongly, caused partly by strategic voting since the largest party traditionally takes the initiative in cabinet formation gets the prime minister position. This resulted in a narrow win for the VVD, with both parties scoring around 20% of the seats. As the far-right PVV with 16% of the seats was no longer considered a serious coalition partner, there were no realistic coalitions without both winning parties, and after a relatively short negotiating period the second cabinet headed by prime minister Mark Rutte (VVD) was sworn in on the 5th of November.

For reference, Table 3 gives an overview of the parties that were represented in the Dutch Parliament in 2010 and 2012.

**Results**

**Who is being quoted?**

Figure 8 shows how often each party is mentioned and quoted, per period. The bars are the total number of articles in which each party is mentioned in that period, while the line (on the secondary axis) is the percentage of those articles in which they were quoted or paraphrase.
Figure 8. How often are parties mentioned (bars) and quoted (lines)?
Table 3

Overview of political parties in the Netherlands

<table>
<thead>
<tr>
<th>Party</th>
<th>Description</th>
<th>Leader</th>
<th>Seats (out of 150)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>2012</td>
</tr>
<tr>
<td>VVD</td>
<td>Conservative</td>
<td>Mark Rutte</td>
<td>41</td>
</tr>
<tr>
<td>PvdA</td>
<td>Social democrat</td>
<td>Diederik Samsom</td>
<td>38</td>
</tr>
<tr>
<td>PVV</td>
<td>Anti-immigrant</td>
<td>Geert Wilders</td>
<td>15</td>
</tr>
<tr>
<td>SP</td>
<td>Socialist</td>
<td>Emile Roemer</td>
<td>15</td>
</tr>
<tr>
<td>CDA</td>
<td>Christian Democrat</td>
<td>Sybrand Buma</td>
<td>13</td>
</tr>
<tr>
<td>D66</td>
<td>Progressive</td>
<td>Alexander Pechtold</td>
<td>12</td>
</tr>
<tr>
<td>ChristenUnie</td>
<td>Progressive Christian</td>
<td>Arie Slob</td>
<td>5</td>
</tr>
<tr>
<td>GroenLinks</td>
<td>Green</td>
<td>Jolande Sap</td>
<td>4</td>
</tr>
<tr>
<td>SGP</td>
<td>Orthodox Christian</td>
<td>Keer van der Staaij</td>
<td>3</td>
</tr>
<tr>
<td>PvdD</td>
<td>Animal rights</td>
<td>Marianne Thieme</td>
<td>2</td>
</tr>
<tr>
<td>50+</td>
<td>Senior</td>
<td>Henk Krol</td>
<td>2</td>
</tr>
</tbody>
</table>

Looking at the number of articles mentioning parties during the campaign as shown in Figure 8(a), we can see that especially the VVD, the senior coalition partner, received a lot of media attention. This is followed by the other large parties (CDA, PvdA, and PVV), who all received around 2,000 mentions. In fact, the attention almost slavishly follows the seat distribution with a correlation of $\rho = 0.94$ ($n = 11, p < 0.001$).

The percentage of quotes yields a less predictable pattern. It seems that all the big parties are quoted in around 10% of articles in which they are mentioned, while the smaller parties are talked about more often instead of talking themselves, with quotes in only around 5% of the articles. Two exceptions to this pattern are D66 and the SGP. The latter is presumably be caused by a scandal that occurred during the campaign sparked by the statement by U.S. senator Akin that ‘genuinely raped’ women would not become pregnant. Since SGP is also anti-abortionist, the SGP leader was asked to comment on that statement, and made a reply that was interpreted as not distancing himself from the Senator, which caused quite a media storm for a party that is normally not at the center of attention.

During the cabinet formation period, as shown in Figure 8(b), this picture is quite different. The two largest parties again get most attention, with the next largest party receiving only around a third of the attention that the largest parties get. Although the distribution of attention seems more skewed, this is a reflection of the less fragmented seat distribution and the correlation between mentions and seats (in the new Parliament) is even stronger at $\rho = 0.98$ ($n = 11, p < .001$).

The quotes follow a completely different pattern, however. Instead of all the larger parties being quoted most, they actually lag at slightly over 5%, while most other parties are quoted even more than during the election, with opposition parties ChristenUnie, 50+, and SP all scoring above 12%. This difference is easy to understand. It was immediately clear that VVD and PvdA would have to form a government, and most of the cabinet formation consisted of negotiations between those parties. Since these negotiations were
behind closed doors, and both parties remained tight-lipped, the media had to find other sources to discuss, or rather speculate about, the ongoing negotiations. Although the VVD and PvdA had a majority in the lower house of parliament, they lacked a majority in the senate, and for that reason there was speculation that a broader coalition would be formed to ensure that legislation would not be halted there. The parties that could have gained the coalition a majority there were CDA or D66 and GroenLinks. Interestingly, these parties also were talked about more than quotes. Thus, it seems that during the cabinet formation, especially those parties who did not participate in the ongoing negotiations were cited.

The ‘routine’ administration period following the coalition formation is shown in Figure 8(c). This shows a similar pattern, with by far most attention going to the two parties in the coalition. In terms of quotes, the patterns is more even, with most parties being quoted in around 6% of articles that they appear in, although the coalition partners seem to be quoted a bit less often compared to the other large parties. Exceptions are the PVV, SGP, GroenLinks parties and the PvdD, who are all quoted in less than 5% of the articles. For the PVV, this can be a result both of their political isolation and of their press strategy, since they give almost no interviews or press conferences.

Table 4 presents a simple pooled OLS regression of the percentage of articles in which a party is quoted. Contrary to H1a, parties are directly quoted in election periods significantly less often then in other periods. Parties in the governing coalition were quoted more often (1.35 percentage points), but this effect is not significant, so H2a can also not be confirmed. Finally, the number of seats had no effect whatsoever on whether a party is cited, rejecting H3a. From this overview it is clear that being quoted is not simply a matter of being politically powerful.

And what do they say?

To determine to what extent the content of parties’ media quotes matches their speeches in Parliament, we present the correlation between the topic agenda in both venues.

Table 5 lists the correlation media and parliamentary discourse aggregated per period. One first thing to not is that all correlations are negative: parties use different topics in the media compared to in parliament. This is at least partially an artifact of the used
Table 5

data correlation between parliamentary speeches and media discourse per period

<table>
<thead>
<tr>
<th>Period</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prelude</td>
<td>-0.26</td>
</tr>
<tr>
<td>Campaign</td>
<td>-0.15</td>
</tr>
<tr>
<td>Formation</td>
<td>-0.20</td>
</tr>
<tr>
<td>Rutte II</td>
<td>-0.25</td>
</tr>
</tbody>
</table>

Table 6

data correlation between parliamentary speeches and media discourse per party

<table>
<thead>
<tr>
<th>Party</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>VVD</td>
<td>-0.32</td>
</tr>
<tr>
<td>PvdA</td>
<td>-0.25</td>
</tr>
<tr>
<td>D66</td>
<td>-0.21</td>
</tr>
<tr>
<td>CDA</td>
<td>-0.06</td>
</tr>
<tr>
<td>GL</td>
<td>-0.05</td>
</tr>
<tr>
<td>SP</td>
<td>-0.03</td>
</tr>
<tr>
<td>SGP</td>
<td>0.04</td>
</tr>
<tr>
<td>CU</td>
<td>0.08</td>
</tr>
<tr>
<td>50plus</td>
<td>0.22</td>
</tr>
<tr>
<td>PvdD</td>
<td>0.51</td>
</tr>
<tr>
<td>PVV</td>
<td>0.56</td>
</tr>
</tbody>
</table>

topic model. As explained above, some of the found topics were of procedural nature and contained words related to the workings of parliament.

If we compare the four periods, the campaign is actually the least negative, confirming H1b. During the campaign period, the media discourse was more similar to the parliamentary discourse than during the other periods. For this analysis we compared the media discourse in a certain period to the parliamentary discourse of the whole period, so it is really the media discourse that is closer to the political discourse in that period.

Table 6 shows the same correlation per party, sorted from most dissimilar to most similar. What is immediate obvious is that the larger parties and coalition parties are actually least similar, in contrast to H2a and H3b. The big exception is the PVV, which is the third biggest party but is also the most similar. Anecdotally, this is probably caused by their clear (populist) discourse wich they use in both media and parliament. The PVV in general is media-seeking rather than policy-seeking, and their parliamentary discourse is probably more at the service of their media discourse than the other way around.

Finally, Table 7 shows the correlation per period and per party. A number of things in this table are quite striking. Although PvdA is overall the second-most dissimilar party, especially in the period before the campaign proper their media discourse was very strongly dissimilar to their parliamentary discourse at −0.42, while during the cabinet formation the correlation was actually slightly positive (0.03).

If we look at the cabinet formation period, it is interesting that the two parties who
were doing the negotiating (VVD and PvdA) were able to stay relatively close to their story in Parliament compared to the campaign period and the period after the formation. The three parties who were potential junior candidates (CDA, D66 and GroenLinks), however, were much more dissimilar during that period.

Finally, it is interesting that the two parties that were seen as ‘losing’ the elections, SP and PVV, were both quite dissimilar in the campaign period compared to the other periods, which is contrary to the general trend. It is an interesting question whether they lose because they are unable to get their preferred statements into the media, or whether as soon as they start losing they are forced to talk about why they are losing, rather than about their issue positions and policy proposals.

### Discussion and Conclusion

The use of the media as a strategic tool by politicians is an understudied aspect of political communication. In this paper, I use syntactic analysis to identify when politicians are cited in Dutch newspapers and extract their statements. By building a topic model on these statements and on their statements in Parliament, I determine to what extent their statements in the media correspond to those in Parliamentary speeches.

Although larger parties and parties in power are mentioned much more in the media, they are not necessarily cited more frequently relative to the number of mentions. In the frequency of citations, an interesting pattern of political supply and journalistic demand emerges. During the campaign, the large parties all receive similar levels of attention, but the party with the largest amount of quotes relative to its media exposure was the orthodox christian SGP party, who were embroiled in a scandal after a clumsy remark about the Akins rape-pregnancy statements. During the cabinet formation after the elections, the pattern is quite different. The parties negotiating the new government choose not to make many (meaningful) press statements, so instead the press goes to the future opposition parties to get their interpretation of the ongoing events.

It turned out that during the election campaign the correspondence between media
and parliament statements was greatest. This corresponds to Hopmann et al. (2012)’s findings that politicians are more able to set the agenda during election periods. This can also be seen as a function of supply and demand: in election times there is a lot of coverage of politics, resulting in more demand for and hence less gatekeeping of political statements. Looking at the different parties, in general the media statements of larger parties are all less similar to their parliamentary statements compared to smaller parties, and especially the limited-issue parties such as the Animal Rights party had highly similar media and parliamentary discourses. Intestingly, while the anti-immigrant PVV party was overall most similar in their media and parliamentary statements, this was not the case during the campaign. This is probably explained by the fact that their preferred topics (immigration and the EU) were not very salient during the campaign, forcing them to make more statements about other topics.

This paper clearly shows that the frequency and content of media statements is a much more complex process than the relatively simple news values that determine how frequently a party is mentioned. This is controlled by structural factors, such as party size and issue breadth, but also by more incidental causes such as the place in the political process. To provide further insight into these mechanisms, I think it is important to expand this study in two ways: (1) by also analysing the content of the extracted topics, and determining whether this content is politically driven (i.e. the own issues of a party) or more media driven (reactions to scandals, etc.); and (2) by comparing the media statements not just to the parliamentary statements, but also to the recent overall media discourse. If these additional variables are determined, a more complete picture of the dynamics of mediated political communication will emerge.

References


