Clause analysis: using syntactic information to enrich frequency-based automatic content analysis

Wouter van Atteveldt
VU Amsterdam

Tamir Sheafem, Shaul Shenhav, Yair Fogel-Dror
Hebrew U. Jerusalem

Abstract
This paper shows how syntactic information can be used to automatically extract clauses from text, consisting of a subject, predicate, and optional source. Since the output of this analysis can be seen as an enriched token list or bag of words, normal frequency based or corpus linguistic analyses can be used on this output. Taking the 2008–2009 Gaza war as an example, we show how corpus comparison, topic modelling, and semantic network analysis can be used to explore the differences between US and Chinese coverage of this war.

Introduction

Spoken and written texts are an important source of information for studying politics. For this reason, content analysis is one of the central methodologies in political communication. Conducting content analysis manually, however, brings a high unit cost in terms of time and money, limiting the amount of material that can be analyzed. That makes it difficult to build large data sets for longitudinal or comparative research. The complex and often reciprocal relations found in studying communication, however, necessitate such data intensive research to unearth causation patterns. For example, the Comparative Agendas Project (Baumgartner et al., 2006, 2009) shows how a large scale content analysis of various political agendas can be used to answer a large variety of research questions.\footnote{See e.g. \url{http://www.comparativeagendas.info/?page_id=16} for a list of publications resulting from the content analysis effort} However, even with the help of machine learning the manual coding effort required for projects of that size are out of reach of all but the best-funded groups Grimmer and Stewart (cf. 2013).

Computer content analysis can help overcome these limitations. Automatic content analysis has been used frequently since at least the General Inquirer (Stone et al., 1966).
Such automatic content analysis systems are often dictionary or keyword based. This means that they treat a text as a vector of word frequencies ignoring the ordering of and relations between the words, in other words, they assume that a text can be viewed as a *bag of words*. Such approaches have been successfully used for automatically determining the topic of a text, for example using machine learning (Sebastiani, 2002). In the comparative agendas project, a technique known as active learning, iterative machine learning by manually coding the most problematic cases, was used to accurately determine the topic of large amounts of documents at a fraction of the cost of coding all documents (Hillard et al., 2008; Collingwood and Wilkerson, 2012). Recent work on unsupervised topic models (i.e. without using manual coding) also shows that the topic of a text can be modeled based only on word frequencies (Quinn et al., 2010; Blei et al., 2003). Monroe et al. (2008) review a large number of articles using lexical features and show how techniques from computational linguistics can be used to determine which words are good indicators of the content of political conflict.

These successful applications of dictionary-based or lexical techniques are all aimed at identifying the content of communication. *Framing analysis* moves beyond establishing which issues occur in a text to determining how these issues are presented. How an issue is framed can determine to a great extent how other actors react to an issue, giving political actors a strong interest in controlling the framing of an issue (cf. Riker, 1986; Entman, 2004). In a conflict situation, both sides have a strong interest in persuading the media to adopt their framing of the issue, turning (mediated) public diplomacy into a strategic contest over international frame building (Sheafer and Shenhav, 2010).

Whether framing can be measured using a purely lexical approach depends on the definition of the frames to be identified. If there is a clear distinction between the object that is frames and the aspects that can be emphasized, keyword co-occurrence methods can be used to perform the framing analysis (Ruigrok and Van Atteveldt, 2007; Schultz et al., 2012). If an actor and issue can be both the object of the framing and part of the frame definition, however, co-occurrence analysis in general does not suffice. Especially in conflict situations, the difference between two ways of framing an issue is often a matter of the exact relations between actors expressed in a text. For example, the sentence ‘Israel blames Hamas for starting the conflict’ frames Hamas as attributed with the responsibility and Israel as the source of this frame. If word order is ignored, this might as well be the other way around. In conflict situations, it is crucial to distinguish between an actor doing something and having something done to it. This distinction is generally expressed in the syntactic structure of a text rather than in the word frequencies used in keyword methods.

This paper explores that intuition by using the grammatical structure of text to improve the results of keyword-based analysis. In particular, sentences are split into clauses consisting of a subject, predicate, and optional source. These clauses can then be analysed as ‘enriched’ bags of words, for example by comparing the predicates used by different sources or by analysing which words are used to relate specific actors in the subject position to actors in the predicate.

By showing how automatic syntactic analysis can be used to automatically extract rich information from text, this paper paves the way for richer automatic large scale text analysis. By differentiating between extracting the clauses and analysing these, this allows existing frequency-based tools to be combined with the information from the clause analysis.

Besides describing the clause method in detail, this paper uses the method to do an
explorative analysis of a data set of US and Chinese (English language) coverage of the 2008 Gaza war. This shows how normal frequency based techniques such as co-occurrence, topic modeling, and semantic network analysis techniques can be directly applied to the results of clause analysis to yield interesting and plausible results.

**Extracting clauses using grammatical analysis**

This section gives a detailed description of how clauses are automatically derived from plain text. This process consists of four steps which will be described in sequence: (1) Parsing the sentences; (2) Extracting quotes and paraphrases; (3) Splitting subjects from predicates; and (4) Extracting the enriched token list from the result.

Note that these techniques are currently developed for English and Dutch. Although a smaller language, Dutch has a very active computational linguistics community and offers very good tool support. It also shows that the techniques can be translated relatively easy between (similar) languages: the (python) code specific to each language is around 50 lines of code.

**Parsing**

The first step in clause analysis is processing the sentences using a syntactic (dependency) parser. A dependency parser converts each sentence into a graph where each node represents 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’ be the root of the dependency tree, with John having a (grammatical) subject relation to ‘love’ while Mary has a (grammatical) direct object relation.

For English, we used the Stanford CoreNLP parser, while for Dutch the Dutch HPSG parser Alpino was used. Figure 1 shows the resulting parse tree for the fictional example sentence: “According to Mary, she was attacked by John” (in Dutch: “Volgens Marie werd zij door Jan aangevallen”). As you can see, the two software packages produce different parse trees, although the main structure is the same. In both cases, the main verb is ‘attacked’ (aangevallen), which is modified by agency phrase (‘by John’ / ‘door Jan’) and by the attribution ‘according to Mary’ (volgens Marie). CoreNLP handles the passive voice explicitly, turning the syntactic object ‘she’ into a passive subject (nsubjpass), and recognizing John as the agent (nmod:agent). Alpino simply represents the syntactic structure, so ‘zij’ (she) is the direct object (obj1) of the verb ‘aangevallen’ (attacked). According to Alpino, ‘zij’ (she) is also the subject of ‘werd’ (was). CoreNLP also explicitly labels the attribution modifier as ‘nmod:according_to’. Finally, CoreNLP also identifies that ‘Mary’ and ‘she’ corefer (i.e., are the same person).

**Quotes and Paraphrases**

Quotes and paraphrases are extracted by applying a limited set of patterns to the parse tree. For English, two patterns are used, as shown in Figure 2:

---

2 All code is available open source at http://github.com/vanatteveldt/xtas
According to Mary,

she was attacked by John

Volgens Marie

werd ze door Jan aangevallen

(a) English

she was attacked by John

(b) Dutch

Figure 1. Parse tree of example sentences

(a) The first patterns is very broad: it looks for a verb from a list of speech verbs\(^3\). If such a verb is found, its subject (or passive agent) is the source, while the complement (defined using a list of possible grammatical relations) is the quote.

(b) The second pattern uses the ‘nmod:according_to’ relation identified by CoreNLP: the parent of this relation is the quote, while the child of the relation is the source.

In Dutch journalistic texts, a wider range of patterns is used, as shown in Figure 3. The first two patterns are comparable to the English ‘according to’ pattern:

\(^3\)tell, show, acknowledge, admit, affirm, allege, announce, assert, attest, avow, claim, comment, concede, confirm, declare, deny, exclaim, insist, mention, note, proclaim, remark, report, say, speak, state, suggest, talk, tell, write, add

Figure 2. Syntax patterns for finding English quotes and paraphrases
(a) “Blijkt uit” (is apparent from) is a quote marker similar to according to, but in contrast with the English CoreNLP parser is not marked specifically by Alpino. So, the pattern looks for the word “uit” as a phrasal complement (pc) of “blijkt”. Then, the direct object (obj1) of “uit” is the source, while the subject of “blijkt” is the quote.

(b) “Volgens” and “aldus” (according to) are semantically similar to “blijkt uit”, but have a different syntactic representation. Here, the source is the direct object of “volgens”, while the quote is its parent, “volgens” functioning as modifier or tag.

In contrast to English, the direct speech verbs\(^4\) (“John says ..”) can yield three different syntactic patterns in Dutch. In all cases, the subject (or passive agent) of the verb is the source, but the quote is found differently depending on the presence of complementizer or quote mark:

(c) If the explicit complementizer ‘dat’ (that) is found as the verbal complement (vc) of the speech verb, the body of ‘dat’ is the quote.

(d) If an explicit quotation mark (”) is found as the parent of the speech verb, the quote is found in the nucleus (nucl) of the speech verb.

(e) Finally, if neither is found, the quote is the tag of the speech verb

**Multi-line quotes.** In both English and Dutch, journalists often use quotes spanning multiple lines. In some cases, the first line contains an explicit quotation pattern, such as the speech verb below:

“What the British did was very important, extremely important,” said Jelena Pejic [..].
“It has loosened the Security Council’s ranks.”
(NY Times, 1997-12-14, Legal Experts Agree on [..] a Global Criminal Court)

In other cases, however, the multi-line quote begins without an explicit speech verb:

Pattie [..] has been profoundly affected by her grandmother’s early tragedy. “Although my father was American-English and my schoolfriends were mainly Jewish, I totally identified as Armenian [..]. It was part of being Armenian.”
(The Guardian, 2001-01-27, A People Killed Twice)

\(^4\)accepteer, antwoord, beaam, bedenk, bedoel, beken, beklemtoon, bekrachtig, belijd, beluister, benadruk, bereken, bericht, beschouw, beschrijf, besef, betuig, bevestig, beweer, bewijs, bezweer, biecht, breng, brul, concludeer, confirmeer, constateer, debiteer, Declareer, demonstreer, denk, draag_uit, email, erken, expliceer, explicieteer, fantaseer, formuleer, geef_aan, geloof, hoor, hamer, herinner, houd_vol, kondig_aan, kwetter, licht_toe, maak_bekend, maak_hard, meld, merk, merk_op, motiveer, noem, nuanceer, observere, onderschrijf, onderstreep, onthul, ontslui, ontvouw, oordeel, parafaseer, postuleer, preciseer, presumeer, pretendeer, publiceer, rapporteer, realiseer, refereer, reken, roep, roer_aan, ruik, schat, schets, schilder, schreeuw, schrijf, signaleer, snap, snater, specificeer, spreek_uit, staaf, stel, stip_aan, suggereer, tater, teken_aan, toon_aan, twitter, verbaas, verhaal, verklaar, verklap, verkondig, vermoed, veronderstel, verraad, vertel, vertel_na, verwacht, verwittig, verwonder, verzeker, vind, voel, voel_aan, waarschuw, wed, weet, wijs_aan, wind, zeg, zet_uiteen, zie, twitter
These types of quotes, which are fairly frequent in journalist text, where identified by looking for citation marks in the beginning and end of sentences. A multi-line quote starts with a sentence beginning with a citation mark, and ends with a sentence ending with a citation mark. If a sentence starts and ends with a citation mark (and does not contain other citation marks), it is treated as a multi-line quote as well. To avoid false positives due to incorrect punctuation handling, at most one sentence without any quotes can be included between the initial and final sentence of the multi-line quote.

If the sentence before the multi-line quote contains a normal quote, the source of this quote is taken to be the source of the multi-line quote as well, as exemplified in the excerpt from the NY Times above, where the source of the first sentence (Ms. Pejic) is also the source of the subsequent quote. If the multi-line quote is not preceded by a normal quote, the top-level subject of the previous sentence is taken to be the source of the multi-line quote. For example, in the quote from the Guardian, ‘Pattie’ is the subject of the first sentence, and is taken as the source of the multi-line quote following it.

**Validation.** To determine the performance of these patterns, the quotes extracted by the computer were compared to a manual gold standard. Since multi-line quotes were used as well, validation was based on 307 successive sentences taken from 10 randomly selected Dutch newspaper articles. In the gold standard, these 307 sentences contained 114 quotes, or almost one in every three sentences. Of these quotes, 93 were correctly identified by the computer, giving a recall of 82%. The computer incorrectly identified another 15 quotes, yielding a precision of 86%. This shows that the system performs well for extracting the source and the quote of sentences.

(Note: formal validation for English still in progress.)
Extracting Clauses

The last syntactic step is the extraction of clauses. As explained above, a clause is a \((subject, \text{predicate})\) pair, with the predicate including all indirect and direct objects. So, the trivial example “Mary loves John” contains a single clause, with ‘Mary’ as subject and ‘loves John’ as predicate.

Technically, a clause is formed for every verb that has a subject (or passive agent), unless that subject is the source of a quote extracted earlier. So, the sentence “According to Mary, she was attacked by John” yields a single clause, with John as the subject (since that is the passive agent), and “she was attacked” as the predicate. Additionally, any verb with a passive subject (nsubjpass) with no corresponding passive agent creates a clause with an empty subject. So, in the sentence “Mary said she was attacked”, ‘said’ does not yield a clause since it is the source of the extracted quote, and ‘she was attacked’ is the predicate of a clause without subject, since we don’t know who attacked her.

Validation. In order to validate the clause extraction, 150 sentences containing either an aggression verb\(^5\) were randomly chosen from international coverage of the 2009 Gaza war. In these sentences, the found clauses were checked against a manual coding of whether Israel or Hamas occurred in the subject or predicate of aggression clauses (i.e., were the perpetrator or victim of the aggression). In total, the manual coding yielded 72 clauses, 93% of which were correctly identified by the system, while precision was 86%. Apart from parse and coreference errors made by CoreNLP, the biggest problem was verbs like ‘dying’, which semantically express that something happened to the syntactic subject.

Clauses as Annotated Bags of Words

Figure 4 shows the parse trees of the example sentences. In both sentences, a quotation and a clause were found, marked in the parse tree with a border around the anchor nodes. In the English sentence, Mary is correctly recognized as source of the quotation. Attacked is the anchor node for both the quote and the predicate of the clause, while John is the anchor node for the clause subject.

To determine the full set of words belonging to a role, for clauses all descendants of the anchor node are collected until a node is found that plays a role in a different clause or quotation. For quotations the same technique is used, except that only other quotation roles can block the collection. This means that clauses can be nested within quotes, but not the other way around. In the example sentence, that means that “According to Mary” is the source of the quotation, and the rest of the sentence is the predicate. The clause subject is “by John”, while the predicate is “she was attacked”.

The big advantage of assigning words to non-overlapping clauses and quotations is that the resulting structure can be easily interpreted as a list of tokens (words) enriched with the extra information. Table 1 shows this token list for the (English) example sentence. The first columns give the word, lemma, and part-of-speech (POS) information as returned by the parser. This is followed by two sets of ID and roles for quotation and clause, respectively: The IDs are used to differentiate between separate quotations/clauses, while the role indicates which word plays which role in the clause.

\(^5\)The list of aggression verbs was created for another project by manually selecting relevant words from WordNet synset ‘contend’ and all hyponyms (transitively).
According to Mary, she was attacked by John

Volgens Marie, werd ze door Jan aangevallen

(a) English

(b) Dutch

*Figure 4*. Parse tree of example sentences

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Tokens enriched with quotation and clause columns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word</td>
<td>Lemma</td>
</tr>
<tr>
<td>1</td>
<td>According</td>
</tr>
<tr>
<td>2</td>
<td>to</td>
</tr>
<tr>
<td>3</td>
<td>Mary</td>
</tr>
<tr>
<td>4</td>
<td>,</td>
</tr>
<tr>
<td>5</td>
<td>she</td>
</tr>
<tr>
<td>6</td>
<td>was</td>
</tr>
<tr>
<td>7</td>
<td>attacked</td>
</tr>
<tr>
<td>8</td>
<td>by</td>
</tr>
<tr>
<td>9</td>
<td>John</td>
</tr>
</tbody>
</table>
As will be shown in the next section, the fact that the result of clause analysis can be seen as enrichments on the token list make it very for further analysis. In a sense, the resulting tokens can be seen as 'annotated bags of words'. Where 'bag of words' is the normal term for a collection of documents represented by their word frequencies (i.e. ignoring word order and relations), the 'bags' produced by clause analysis can represent various interesting contexts. For example, one could extract all words in quotations by a certain actor, or all words where one actor does or says something to another specific actor. Because the resulting word vectors can be treated as regular ‘bags of words’, this allows for existing techniques such as corpus analysis and topic modeling to be used for further analysis.

Analyzing clauses: Coverage of the 2009 Gaza War

In order to show how regular corpus and network analytic techniques can be used to analyse the data yielded by clause analysis, this section presents an analysis of the foreign coverage of the 2009 Gaza war.

On December 27, 2008, Israel launched a military operation against Hamas in the Gaza Strip, with the stated aim ‘of stopping the bombardment of Israeli civilians by destroying Hamas’ mortar and rocket launching apparatus and infrastructure’ and ‘of reducing the ability of Hamas and other terrorist organizations in Gaza to perpetrate future attacks against the civilian population in Israel.’ The operation started with a wave of airstrikes launched against targets in the Gaza Strip. During these attacks, Hamas in turn intensified its rocket and mortar attacks against southern Israel, reaching major cities such as Beersheba and Ashdod for the first time. The airstrikes were followed by an operation on the ground, which ended on January 18, 2009, when Israel declared a unilateral ceasefire, followed by Hamas announcing a one week ceasefire twelve hours later. Israel completed its withdrawal from Gaza on January 21. According to various sources, the conflict resulted in between 1,166 and 1,417 Palestinian and 13 Israeli fatalities.

During the war, the Israeli Prime Minister’s Office and the Ministry of Foreign Affairs issued English-language statements targeted at the international community on a daily basis. They distributed their materials as official guidelines for Israel’s public diplomacy efforts to all relevant units (Israeli embassies, government ministers, and spokespersons). Such “organized attempts by a government to exert as much control as possible over the way state policy is portrayed in foreign media” (Entman, 2008, p. 89) are called mediated public diplomacy. It is argued that frame building contests play a central role in mediated public diplomacy (Sheafer and Shenhav, 2010), and that the closer a country is to Israel politically and culturally, the more effective this frame building will be (Sheafer et al., 2014).

The substantive question to be answered in this section is whether we can use clauses to show the difference in coverage of this war between U.S. and Chinese media. For this purpose, all available English language sources from the U.S. and China were downloaded from LexisNexis about the Gaza war were downloaded and processed using the method described above. In total, this yielded 795,620 clauses from 2176 Chinese and 6475 American articles. In these clauses, simple keyword lists were used to identify references to Israel and Hamas.

Table 2

*Mentions of Hamas and Israel in different roles in US and Chinese media (column percentages)*

<table>
<thead>
<tr>
<th></th>
<th>Hamas</th>
<th>Israel</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>U.S.</td>
<td>China</td>
<td>U.S.</td>
<td>China</td>
</tr>
<tr>
<td>All clauses</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Source</td>
<td>6.7%</td>
<td>10.6%</td>
<td>12.0%</td>
<td>11.0%</td>
</tr>
<tr>
<td>Subject</td>
<td>45.2%</td>
<td>38.8%</td>
<td>45.3%</td>
<td>50.0%</td>
</tr>
<tr>
<td>Predicate</td>
<td>48.1%</td>
<td>50.5%</td>
<td>42.7%</td>
<td>39.0%</td>
</tr>
<tr>
<td>N</td>
<td>34837</td>
<td>4188</td>
<td>80121</td>
<td>12768</td>
</tr>
<tr>
<td>Clauses containing aggression</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Source</td>
<td>5.2%</td>
<td>7.3%</td>
<td>12.3%</td>
<td>12.3%</td>
</tr>
<tr>
<td>Subject</td>
<td>49.1%</td>
<td>33.1%</td>
<td>65.4%</td>
<td>70.3%</td>
</tr>
<tr>
<td>Predicate</td>
<td>45.7%</td>
<td>59.5%</td>
<td>22.3%</td>
<td>17.4%</td>
</tr>
<tr>
<td>N</td>
<td>3298</td>
<td>477</td>
<td>9621</td>
<td>1645</td>
</tr>
</tbody>
</table>

Who is quoted?

Journalists are strongly dependent on their sources, and by quoting a source a journalist gives a platform to that source and allows him or her to give a certain perspective on the news. From the literature, we know that journalists tend to quote actors who they consider part of the mainstream political debate, and tend to shut out actors that have views outside this spectrum (Bennett, 1990; Bennett et al., 2007; Wolfsfeld, 2014).

Table 2 gives an overview of how often Hamas and Israel occur in the source, subject, and predicate roles of clauses. The first three rows show the distribution (as a column percentage) of these actors over the three roles in general. If we look at the source role, we see that in both Chinese and American media an Israeli actor is used as source in just over 10% of their overall mentions. For Hamas, however, the picture is different: while in the Chinese media it is quoted almost as often as Israel (relative to overall mentions), but in the American media this is only 6.7% compared to 12% for Israel. So, in the American media Israel is apparently a much more acceptable source of statements. If we consider the distribution over the subject and predicate roles, we see that in the American media both actors occur in both roles with almost even frequency, Hamas being in the predicate slightly more often than Israel. In the Chinese media, however, this difference is much more profound: Hamas is subject in 38% of mentions and part of the predicate in 50.5%, while Israel is subject in 50% and predicate in 39%.

This difference is even stronger if we only consider predicates which contain aggression, shown in the bottom half of the table. There, in the Chinese media Hamas is subject in just over 33% of clauses and in the predicate of almost 60%, while for Israel this is 70% and 17.4%, respectively. Although in the American media the differences are also more pronounced than for all clauses: Both Hamas and Israel occur more often as subject than in the predicate of aggression clauses, but while for Israel this is much stronger than for Hamas, the difference is smaller than in the Chinese media.

From this analysis, we can state that in the American media Israeli actors are much
more likely to be used as a source of statements than Hamas, and both Hamas and Israel are often portrayed as subject (perpetrator) of aggression. In the Chinese media, on the other hand, both actors are about as likely to be used as a source, and Hamas is much more often seen as the object of aggression, while Israel is almost exclusively the subject of aggression. Thus, we can state that the US media seems much more pro-Israeli than the Chinese media in this aspect of their conflict coverage.

Who does what to whom?

Besides looking at the frequency of statements, we can analyse the content of actions conducted by each actor. Taking a purely corpus linguistic approach, we can see which words occur most frequently in the predicates with a specific actor as subject. For each actor, we select these words that occur

To compare American and Chinese media, for each actor we select those words that occur more often in their predicates according to American media than according to Chinese media, and vice versa. Figure 5 shows the results of this analysis, where the most frequent words are shown in a semantic network with ties based on co-occurrence in the same predicate. To reduce the amount of ties, backbone extractions was conducted followed by a simple clustering as indicated by node color.

If we look at the words in predicates where Israel is subject according to the American media (5a), we see that the main cluster is about the rocket assauly (by) Hamas, and the (Israeli) right to defend. A separate cluster talks about allowing supplies. In the Chinese media (5b), Gaza rather than Hamas is the central concept, and the rocket attacks are not included. Rather, the focus is on the (Israeli) military operation, and includes terms like 'aggression' and 'humanitarian'.

A similar difference is found in the words in predicates with Hamas as the subject. In the American media (5c), focus is on Hamas as a 'terrorist organization', the missiles fired, the use of civilians as human shields, and the smuggling of weapons. In the Chinese media on the other hand (5d), these activities are not mentioned, and most attention is devoted to diplomatic efforts and a unilateral cease fire. Also, their actions against Israel are described using military terms as soldier and army.

As above, we see here that the American media is more pro-Israeli, focussing on the supposed terrorist acativity of Hamas as a reason for the Israeli actions, while the Chinese media concentrate on the Israeli military actionrs themselves and on possible diplomatic solutions to the conflict.

A topic model of the predicates

Rather than relying on a keyword-based for analysis of predicate content, we can also employ topic modeling to automatically cluster the predicates. Table 3 shows the result of a 20-topic LDA model on the nouns and verbs in the all found predicates. The four central columns show the usage of each topic in predicates with Hamas and Israel as subject according to the US and Chinese media. The topics are sorted by occurrence, with the topic topics being used by these actors the most.

The top two topics are topic #18, which is a mainly Hamas-used topic about rocket attacks, and #6, which is a mainly Israel topic on the military offensive. Interestingly, the
(a) Israel subject, US media  (b) Israel subject, Chinese media

(c) Hamas subject, US media  (d) Hamas subject, Chinese media

Figure 5. Most frequent words in predicates with Hamas and Israel as subject in Chinese and American media (Visualized as co-occurrence network)
Table 3
LDA topics and occurrence in predicates of Hamas and Israel according to US and Chinese media

<table>
<thead>
<tr>
<th>Topic</th>
<th>Hamas</th>
<th>Israel</th>
<th>Hamas</th>
<th>Israel</th>
<th>Top words in topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>0.29</td>
<td>0.11</td>
<td>0.29</td>
<td>0.12</td>
<td>rocket, Israel, attack, fire, Hamas, use</td>
</tr>
<tr>
<td>6</td>
<td>0.08</td>
<td>0.35</td>
<td>0.05</td>
<td>0.17</td>
<td>Gaza, offensive, ground, Strip, air, operation</td>
</tr>
<tr>
<td>1</td>
<td>0.14</td>
<td>0.09</td>
<td>0.07</td>
<td>0.07</td>
<td>call, ceasefire, Gaza, end, international</td>
</tr>
<tr>
<td>20</td>
<td>0.09</td>
<td>0.09</td>
<td>0.10</td>
<td>0.10</td>
<td>Gaza, border, control, open, israeli</td>
</tr>
<tr>
<td>3</td>
<td>0.07</td>
<td>0.11</td>
<td>0.08</td>
<td>0.11</td>
<td>Gaza, hit, City, home, school, area</td>
</tr>
<tr>
<td>2</td>
<td>0.09</td>
<td>0.05</td>
<td>0.11</td>
<td>0.08</td>
<td>Hamas, Israel, palestinian, group, terrorist</td>
</tr>
<tr>
<td>7</td>
<td>0.06</td>
<td>0.06</td>
<td>0.08</td>
<td>0.07</td>
<td>kill, people, civilian, Palestinians, more, child</td>
</tr>
<tr>
<td>8</td>
<td>0.05</td>
<td>0.05</td>
<td>0.06</td>
<td>0.07</td>
<td>peace, Israel, state, right, achieve, end</td>
</tr>
<tr>
<td>12</td>
<td>0.08</td>
<td>0.05</td>
<td>0.03</td>
<td>0.04</td>
<td>United, talk, hold, meet, States, meeting</td>
</tr>
<tr>
<td>9</td>
<td>0.02</td>
<td>0.04</td>
<td>0.03</td>
<td>0.06</td>
<td>Gaza, humanitarian, aid, allow, need, supplies</td>
</tr>
<tr>
<td>16</td>
<td>0.03</td>
<td>0.02</td>
<td>0.04</td>
<td>0.05</td>
<td>East, Middle, world, conflict, country</td>
</tr>
<tr>
<td>5</td>
<td>0.02</td>
<td>0.02</td>
<td>0.04</td>
<td>0.04</td>
<td>make, time, way, thing, little, same</td>
</tr>
<tr>
<td>13</td>
<td>0.03</td>
<td>0.02</td>
<td>0.03</td>
<td>0.03</td>
<td>year, last, day, week, month, fall</td>
</tr>
<tr>
<td>11</td>
<td>0.04</td>
<td>0.01</td>
<td>0.02</td>
<td>0.03</td>
<td>report, news, follow, story, information, see</td>
</tr>
<tr>
<td>19</td>
<td>0.01</td>
<td>0.01</td>
<td>0.03</td>
<td>0.03</td>
<td>take, Obama, time, first, president, Bush</td>
</tr>
<tr>
<td>15</td>
<td>0.02</td>
<td>0.01</td>
<td>0.03</td>
<td>0.03</td>
<td>see, find, body, home, come, look</td>
</tr>
<tr>
<td>10</td>
<td>0.02</td>
<td>0.01</td>
<td>0.02</td>
<td>0.02</td>
<td>Senate, seat, state, today, vote, Obama</td>
</tr>
<tr>
<td>14</td>
<td>0.01</td>
<td>0.01</td>
<td>0.02</td>
<td>0.02</td>
<td>money, pay, economy, tax, cut, job</td>
</tr>
<tr>
<td>4</td>
<td>0.01</td>
<td>0.01</td>
<td>0.02</td>
<td>0.02</td>
<td>New, protest, York, take, hold, rally</td>
</tr>
<tr>
<td>17</td>
<td>0.00</td>
<td>0.01</td>
<td>0.02</td>
<td>0.02</td>
<td>know, think, want, happen, see, mean</td>
</tr>
</tbody>
</table>

latter topic is associated with Israel much less in American media (15%) than in Chinese media (35%), reflecting the stronger Chinese focus on Israel’s offensive actions. The reverse is seen for topic #7 (seemingly on civilian casualties), which according to the American media is used more if Hamas is subject (8%) than if Israel is subject (7%), while in the Chinese media it is used slightly more if Israel is subject. Finally, it is interesting to consider topic #11, which contains words related to giving information: in the Chinese media this is used much more by Hamas than Israel, while in the American media this is reversed. This is an ‘inductive’ confirmation of the rule-based outcome reported above that the Chinese media are much more likely to quote Hamas actors than the American media.

Clause-based Semantic Network Analysis

In most applications, Semantic Network Analysis or text-based Social Network Analysis creates a graph of words or actors related by co-occurrence in a sliding window or other context unit. This yields an undirected network, or at best a network directed by surface word order (which is not very informative in languages with flexible word order). Using clause-based analysis, we can construct a network of edges from terms or actors in the subject position to those in the predicate.
Figure 6. Semantic Network of subject-predicate links in Chinese and American media
Figure 6 shows the result of this analysis, displaying the semantic network for American and Chinese media. Not surprisingly, in both networks we see a dense central cluster around the words 'Gaza', 'Israel', and 'Hamas'. The US network is more concentrated, and has a large cluster devoted mainly to the military operation (displayed in red; although it also includes the United Nations resolution). Additionally, two clusters around US politics are seen in the top-left. In the Chinese media, on the other hand, the military terms are less central, and diplomatic terms are much more prevalent, also including (other) Arab leaders.

**Conclusion**

This paper shows how grammatical analysis can be used to automatically split sentences into *clauses* consisting of a subject, predicate, and optional source. This uses the output of grammatical analysis to extract useful information of ‘who does what to whom’, making it possible to answer questions that are difficult to answer using bag of words methods.

Since the process of splitting sentences into clauses does not depend on substantive choices, this approach allows the ‘hard part’ of grammatical analysis – dealing with syntax trees – to be standardized and offered as a web service. Researchers that wish to use the method can then use their favourite analysis tool (e.g. R or python) for further analysis. Since the results of clause analysis can be seen as an ‘enriched’ token list or bag of words, all normal corpus linguistic tools and techniques can be applied to these results by using the clause information to split and filter the data in interesting ways, e.g. by comparing predicates from different sources or actors.

In order to show how the results of clause analysis can be used for substantive analysis, I analysed all clauses from the coverage of the Gaza war in American and (English-language) Chinese media. First, analysis of which actors are used as source in the news show that, while American media quote Hamas actors much less frequently than Israeli actors, both are given equal voice in the Chinese media. Additionally, in Chinese media Israel are much more often the subject of clauses, especially of clauses containing aggression verbs, while Hamas are more often seen as the object of aggression.

From an analysis of the relative frequency of words used in predicates with Israel and Hamas as subject, I showed that while in Chinese media Israel is mostly associated with violence and Hamas is (also) associated with diplomacy, in the US media the terrorist actions of Hamas and Israel’s right to defend itself are stressed. This is confirmed with a topic model of the same predicates, which shows that the Chinese media focus more on Israel’s military offensive, while the US media associate Hamas especially with rocket attacks and civilian casualties.

Finally, a semantic network analysis was conducted with links from words in the subject of clauses to words in the predicate. The American media have a strongly clustered network around the conflict itself, with a secondary cluster of political terms; while the Chinese media have a much less concentrated network, focussing also on the diplomatic aspects of the conflict and the role of (other) Arab countries.

These results show that clauses produced by this method can be analysed in various ways, yielding rich and plausible results that are in line with existing literature on the conflict. Although a number of interesting analyses of the data were shown in this paper, it
would also be very simple to use other tools for e.g. sentiment analysis, machine learning, or apply existing disctionaries such as in LIWC or DICTION to these clauses.

Although the substantive results are based on English media, the rules are also developed for the Dutch language with relatively minor adaptations. This shows that it should be possible to extend the system to other languages with relative ease, as long as a good (dependency) parser is available and journalists use predictable patterns for quotes and paraphrases.

References


