Clause analysis: Using syntactic information to automatically extract source, subject, and predicate from texts with an application to the 2008-2009 Gaza War *

Wouter van Atteveldt
VU University Amsterdam

Tamir Sheafer, Shaul R. Shenhav, Yair Fogel-Dror
Hebrew University of Jerusalem, Israel

Abstract

This article presents a new method and open source R package that uses syntactic information to automatically extract source–subject–predicate clauses. *This improves on frequency based text analysis methods by dividing text into predicates with an identified subject and optional source, extracting the statements and actions of (political) actors as mentioned in the text. The content of these predicates can be analyzed using existing frequency based methods, allowing for the analysis of actions, issue positions and framing by different actors within a single text. We show that a small set of syntactic patterns can extract clauses and identify quotes with good accuracy, significantly outperforming a baseline system based on word order. Taking the 2008–2009 Gaza war as an example, we further show how corpus comparison and semantic network analysis applied to the results of the clause analysis can show differences in citation and framing patterns between U.S. and English-language 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. The complex and often reciprocal relations found in political communication, however, require analyzing large corpora of text to unearth causation patterns. Large digital text archives are increasingly available, but the high cost and complexity of large scale manual text analysis is out of reach of all but the best-funded groups (cf. Grimmer and Stewart, 2013).

For this reason, many researchers turn to automatic text analysis. As showcased in a recent virtual issue on Innovations in Text Analysis in this journal (Roberts, 2015), the methodological toolkit now includes methods such as automatic scaling of latent traits (Lowe and Benoit, 2013), topic modeling (Blei et al., 2003; Grimmer, 2010), automatic text classification using machine learning (D'Orazio et al., 2014), and automatic event coding (Schrodt and Gerner, 1994, 2000).

With the exception of event coding, all these methods are frequency based, treating text as a 'bag of words' and ignoring the way in which words are combined to express relations between the actors and issues mentioned in a text. This makes these methods very useful for determining the topic of a text, or for measuring ideology or tone as expressed by the whole text. Texts such as news items, however, usually mention multiple actors that make state-

*Corresponding author: wouter@vanatteveldt.com.
The data and R scripts for replicating the validation and substantive analyses are published in the Harvard Dataverse (Van Atteveldt et al., 2016). Please cite this paper as: Van Atteveldt, W., Sheafer, T., Shenhav, R.S., & Fogel-Dror, Y. (forthcoming). Clause analysis: using syntactic information to enrich frequency-based automatic content analysis. Accepted for publication in *Political Analysis*

ments about different other actors and issues. For such documents, frequency based methods are less well suited as they do not distinguish between the statements by different actors, or between an actor that makes a statement or an actor that is being talked about. For analyzing such documents, it is often necessary to not just identify actors, but specify what role they play, and what their relation is to each other and to mentioned frames, issues, or evaluations. This requires looking beyond word frequency and zooming in on the semantic relations expressed in the text.

This article presents a new method, *clause analysis*, that uses automatic syntactic parsing to enable more informative automatic text analysis.[1] High quality parsers now exist for many languages, yielding a syntactic representation of each sentence that shows the verbs used and their subject, object(s), and modifiers. From this we automatically extract clauses consisting of an optional source, a subject, and a predicate, where the predicate contains both the action and its object and modifiers. Each clause can be seen as an action or event, showing which words are used to describe the action, and which role each actor plays. This method can be combined with existing methods such as reviewed above: by dividing each document into different contexts, e.g. with a specific actor as source or subject, frequency-based analysis can be performed on these contexts. For example, a topic model can be made of all quotes from a political actor, the vocabulary of different actors can be compared, or the ideology of quoted actors determined using scaling.

After reviewing the relevant literature in political methodology and computational linguistics, the method is described in more detail in the third section. This is followed by a validation of the method by comparing it with both manual coding and with a baseline that uses the word order to separate subject from object. Besides this micro-level validation, in the fifth section we present a macro-validation where we use the method to analyze conflict coverage. In particular, we compare the way in which the U.S. and (English-language) Chinese press covered the 2008-2009 Gaza War. We know from the literature that Chinese and U.S. media differed markedly in how they covered the Gaza War, with the U.S. media being more receptive to the official Israeli framing that the attack was a justified response to Palestinian terrorism (Sheafer et al., 2014). This makes it an interesting case to show how, in a context in which almost every article mentions both Israel and Hamas, grammatical analysis can show systematic bias in who is used as source, who is portrayed as aggressor, and how the actions of both parties are framed.

[1]This method is implemented as an open source R package that can be applied directly to the output of the syntactic parser. The package is published at `http://github.com/vanatteveldt/rsyntax`.

Automatic Analysis of Political Communication

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, meaning 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 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), or on a combination of word frequencies and metadata (Grimmer, 2010).

Frequency-based approaches have also been used to analyze other aspects of political communication such as tone and frame. For ex-

ample, dictionary based approaches have been used to measure political affect (e.g. Young and Soroka, 2012) and framing (Kellstedt, 2003; Ruigrok and Van Atteveldt, 2007). 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. Automatic scaling techniques have been used for estimating party positions either from unannotated text (Slapin and Proksch, 2008; Lowe and Benoit, 2013) or by using reference texts to anchor the political dimensions (Laver et al., 2003).

These methods are all frequency-based and as such do not capture the relations between actors and issues expressed in texts such as news reports. A notable exception is the work of Phillip Schrodt and colleagues on automatic event coding (Schrodt and Gerner, 1994, 2000; Schrodt et al., 2005). They have developed the TABARI automatic coding system and its current successor system, PETRARCH, which combine linguistic data with a sophisticated dictionary for coding (conflict) events and their subject and object (Schrodt, 2014). While TABARI relies on replacement patterns to differentiate subject and object, PETRARCH uses full syntactic parsing, making it similar in goal and general methodology to the clause analysis method proposed in this paper. There are some important differences, however. First, PETRARCH is aimed specifically at coding a predefined set of events, and uses a manually crafted and fine-tuned dictionary for identifying and classifying these events, outputting subject, event type, and object. In contrast, clause analysis does not classify the event (action) or specify the object, instead using the grammatical analysis only to separate the source and subject of the action from the predicate. Thus, rather than concentrating on high-precision extraction of very specific information, clause analysis decomposes text into smaller units (the predicates) with a recognized subject and optional source. The extracted predicates can then be analyzed with normal frequency based methods such as topic modeling, sentiment analysis, or semantic network analysis. As will be illustrated in the application section, this can be used to show what actors talk about and how they are described.

The system is also related to the work described by Van Atteveldt et al. (2008) who use syntactic information to attribute sentiment expressions to actors, and the study presented by Fogel-Dror et al. (2015), who use a machine learning approach on top of syntactic relations for sentiment attribution. Compared to these systems, the current system has a different focus: it is aimed at enhancing frequency based methods to allow framing by incorporating the relations between actors, rather than analyzing the sentiment expressed in a text. Moreover, the current system is simpler and easier to translate, since it only uses a limited number of syntactic patterns rather than relying on complicated syntax rules or statistical models trained on annotated data.

Within the field of computational linguistics, this system can be seen as a specialized and simplified version of Semantic Role Labeling (SRL; Carreras and Màrquez, 2005). SRL is usually performed using machine learning based on large data sets manually annotated with multiple possible semantic roles for each action. For example, the Semafor SRL system (Chen et al., 2010) has actions such as `shoot_projectiles`, for which it lists roles like `agent`, `projectile`, and `purpose`. Given the large variety of possible actions and roles, however, systems based on these manually crafted resources almost always have data scarcity problems. The method described here avoids these problems by not trying to identify all possible roles for an action, but only separating source and subject from the rest of the predicate.

Extracting clauses using grammatical analysis

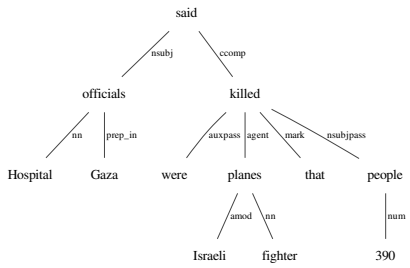
The process of automatically extracting clauses and quotes consists of four steps: (1) Parsing the sentences; (2) Extracting quotes and

paraphrases; (3) Extracting subject-predicate clauses; and (4) Extracting the enriched token list from the result.

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' at the root of the dependency tree, with John having a subject relation to 'love' while Mary has a direct object relation.

We used the Stanford CoreNLP parser to automatically parse English sentences. The CoreNLP parser is open source wide-coverage dependency parser developed at Stanford University (De Marneffe et al., 2006). Figure 1 shows the resulting parse tree for the (somewhat abridged) example sentence *Hospital officials in Gaza said that 390 people were killed by Israeli fighter planes.*[2] The root of the tree is the main verb, 'said'. To the left is its subject (*nsubj*), 'officials', which is a compound noun (*nn*) with 'Hospital'. It is modified with the prepositional phrase 'in Gaza', which Stanford represents as a single *prep_in* relation with the word 'Gaza'. On the right hand side is the clausal complement (*ccomp*) of 'said'. This clause is a passive construction with 'killed' as the main verb. CoreNLP explicitly represents the passive structure: the *agent* is the one doing the killing ('Israeli fighter planes'), while the passive subject (*nsubjpass*) '900 people' are being killed.



Hospital officials in Gaza said that 390 people were killed by Israeli fighter planes

Figure 1: *Example sentence parsed by the Stanford CoreNLP dependency parser.*

Quotes and Paraphrases

After the sentences are parsed, the syntactic structure is used to identify quoted and paraphrased sources. As can be seen in the example presented above, the source-quote relation that needs to be extracted is expressed quite directly in the syntactic structure. In fact, the clause analysis method relies on two relatively simple syntactic patterns to identify quotes:

(a) The first pattern identifies sentences with an explicit speech verb, like the 'X said that Y' pattern shown in the example sentence. Technically, this pattern looks for a verb from a fixed list of speech verbs derived from WordNet.[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 is used for source attributions with the common pattern 'According to X, Y'. This is represented by CoreNLP with a special relation *nmod:according_to*, with the parent of this relation being the quote, while the child of the relation is the source.

In the example sentence, the first pattern matches the speech act verb 'says', the source being the subject 'officials', and the quote the complement 'killed'.

[2] *Israel rejects Gaza cease-fire*; New York Times, January 1st 2009.

[3] The dictionary containing the speech verbs as well as the aggression words and indicators for Israel and Hamas used below are available online in the supplemental materials.

Extracting Clauses

The next step is extracting the subject–predicate clauses that express actions. A clause is a *(subject, 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. The clause analysis method described here uses 4 patterns for identifying clauses:

(a) The most common pattern is the straightforward case where a (non-speech) verb has a subject (or passive agent). In this case, a clause is formed with that subject or agent as clause subject, and the remainder of the verb phrase as the predicate.

(b) A second pattern is used for clausal complements which do not have a subject (*xcomp*). In that case, the direct object of the main clause is the subject of the predicate in the clausal complement. For example, in the phrase 'John told Mary to go', Mary is the predicate subject (actor) of the verb 'to go'.

(c) A third pattern is used for nominal actions. This requires a list of nouns that can be used to express actions, for which lexical resources such as WordNet and FrameNet can be used (Miller, 1995; Fellbaum, 1998; Baker et al., 2003). For the case described in this article, a list of aggression actions like 'invasion' and 'attack' was derived from all WordNet synonyms and hyponyms of attack. If such a word is found outside an existing clause, a clause is formed with the action as predicate, and the possessive child as subject (e.g., 'Israeli' in the phrase 'The Israeli invasion').

(d) Finally, these existing patterns are extended to deal with conjunctions and subordinate wh-clauses, where the subject of the clause is copied to the conjunct or wh-clause, respectively. For example, in the (fictional) sentences 'Hamas fired rockets at Israel, killing 20 civilians' Hamas is seen as subject of both firing and killing; while in the (fictional) sentence 'Hamas fired rockets at Israel, who responded by bombarding Gaza' Israel is in the predicate of 'fired', while it is the subject of 'respond [by bombarding]' in the wh-clause 'who responded by'.

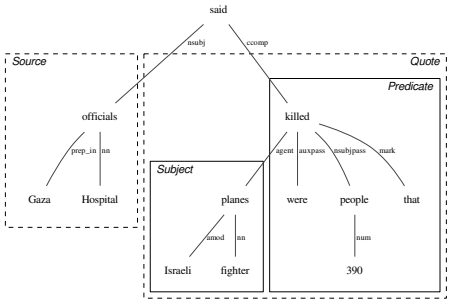


Figure 2: *Example sentence with quote and clause identified by the syntactic rules*

Extracting Quotes and Clauses as Annotated Bags of Words

The steps above identify source–quote and subject–predicate pairs as relations between the main words in each role. In the final step, these words are expanded to the complete phrases headed by these main words. Technically, each role is expanded to contain all descendants of the main word that was identified, but excluding all (descendants of) words identified as main words for other clauses or sources. Sources are expanded in the same way, but the expansion is only stopped by other sources, not by clauses, so a quote can contain one or more clauses, but a clause never contains quotes.

Figure 2 shows the parse tree of the example sentence, in which a single quote and a single clause were identified by the system. For the quote, 'officials' was identified as the source, which is expanded to include 'Hospital' and '(in) Gaza'. The word 'killed' anchors both the quote and the predicate of the clause. The

quote is expanded to contain the rest of the sentence ('that ... fighter planes'). The predicate, however, does not contain the phrase 'Israeli fighter planes', since 'planes' was identified as the subject of the predicate, blocking the expansion of the predicate.

An 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 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 IDs 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.

As will be illustrated in the next section, the fact that the result of clause analysis can be seen as enrichments of the token list makes it convenient 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, topic modeling and scaling to be used for further analysis.

For reasons of clarity, an example sentence was chosen that contains only a single quote and a single clause. Most sentences in newspaper articles, however, contain multiple nested and juxtaposed clauses. For example, consider the following (not even exceptionally long) sentence: *Since December 27, when Israel launched its massive assault on the Hamas rulers of the Gaza Strip, Palestinian militants have fired approximately 600 rockets and mortar shells into southern Israel, killing four people and wounding dozens more, according to the Israeli army.*[4] This sentence contains two clauses ('Israel launched its assault' and 'militants fired rockets'), both attributed to the Israeli army as source. Such sentences shows how difficult it is to understand who is doing what in complicated conflict situations using only word order, while the (syntactic) verb structure of a sentence quite directly expresses the needed information: the only verbs with subjects are launched and have (fired), each heading their respective clause which is expanded to contain the target of the actions, allowing the syntactic patterns to correctly identify both clauses and their source.

Validation

The system was validated by comparing the automatically generated results to a sample of sentences that was annotated manually.[5] For this case, a random sample of 250 sentences that contained an aggression word and a reference to Israel and/or Hamas was drawn from international coverage of the Gaza war. All sentences were manually coded by identifying the aggression mentioned in the sentence, and determining whether the aggressor, victim, and (optional) attributed source referred to Israel, Palestinians, or another actor. From the resulting coded aggressions, all unique (aggressor, victim, source) triples were kept, yielding in total 204 coded aggressions from 183 relevant sentences.[6] In order to estimate reliability of the manual coding, a sample of 35 sentences were also coded independently by a second coder, yielding a Krippendorff's alpha of .72 for source and .79 for both aggressor and victim.

[4] *Two soldiers wounded as Gaza rocket hits Israel*, AFP, 2009-01-08.

[5] The data and R scripts for replicating the validation and substantive analyses are published in the Harvard Dataverse (Van Atteveldt et al., 2016)

[6] The method presented in this article is aimed at identifying the subject and source of actions, not at identifying whether an action constitutes an aggression. For that reason, sentences not containing aggression were deemed irrelevant and manually removed, and the coreferences yielded by Stanford were manually corrected.

Table 1: *Tokens in the example sentence as parsed by CoreNLP enriched with quote and clause information*

				Quotation		clause	
	Word	Lemma	POS	ID	Role	ID	Role
1	Hospital	hospital	Noun	1	subject		
2	officials	official	Noun	1	subject		
3	Gaza	Gaza	Name	1	subject		
4	said	say	Verb				
5	that	that	Pronoun	1	quote	1	predicate
6	390	390	Quantity	1	quote	1	predicate
7	people	people	Noun	1	quote	1	predicate
8	were	be	Verb	1	quote	1	predicate
9	killed	kill	Verb	1	quote	1	predicate
10	Israeli	israeli	Adjective	1	quote	1	subject
11	fighter	fighter	Noun	1	quote	1	subject
12	planes	plane	Noun	1	quote	1	subject

These manually coded aggressions were then compared to the automatically extracted quotes and clauses. For each automatically extracted clause, it was identified whether one of the words in the subject of predicate referred to either Israel or Palestinians. If the clause was contained in a quote, the actor identified was included as source, yielding (aggressor, victim, source) triples similar to the manual coding.

For both clause and source extraction the automatically extracted results were compared to the manual analysis by computing precision and recall scores, the standard scores used in Information Retrieval. Precision is defined as the percentage of extracted clauses/sources that was correct (i.e., that was also included in the manually coded triples), and is a measure of the accuracy of the extraction. Recall is defined as the percentage of manually coded clauses/sources that was found by the system, giving a measure of the coverage of the extraction. The F1 score is the harmonic mean of precision and recall and is often used as an overall performance measure (e.g. Grossman and Frieder, 2012).

The results of this validation are listed in Table 2. As can be seen in the table, clause extraction has fairly good precision and recall, both being around 0.7. The rules for source identification have very good precision (0.95), while recall is moderate (0.61). This shows that if a source is identified it is almost always correct, but the system misses relatively many sources.

Table 2: *Performance of clause and source extraction using syntactic rules compared to word order baseline*

Method	Clause Extraction			Source Identification		
	Pr	Re	F1	Pr	Re	F1
Clause analysis	.70	.72	.71	.95	.61	.74
Baseline	.36	.35	.35	.50	.62	.55

Pr.=Precision (percentage of found items that were correct); Re.=Recall (percentage of correct items that were found). F1=F1 Score (harmonic mean of precision and recall)
Note: N=204 manually identified clauses from 183 randomly selected sentences mentioning either Israel or Hamas and an aggression word, of which 74 contained a quoted or paraphrased source.

To investigate whether syntactic rules are indeed needed for analyzing conflict coverage, the system is compared to a baseline system that relies on the relatively fixed English word order to distinguish between actor and object: for each actor pair, the actor occurring first in the sentence is seen as the aggressor, while the actor occurring later is seen as the victim. As can be seen from the table, this performs sub-

stantially worse than the syntactic rules, giving low precision and recall scores of around 0.35.

The source extraction was similarly compared to a word-order based baseline. In this baseline system, the same speech verbs were identified as in the syntactic rules, with an actor occurring before the speech verb identified as the source and the statements occurring after the speech verb identified as quote. As shown in the table, this gives recall that is comparable to the syntactic rules (0.62), but much lower precision (0.5), also resulting in a lower F1 score (0.55 compared to 0.74).

Error analysis

The results presented here show that the syntactic rules can be used to extract clauses and sources with a fairly good accuracy, strongly outperforming a word-order baseline. However, the system still makes a number of errors at the sentence level. From a qualitative inspection of the sentences used for validation, it turns out that a large part of these mistakes are attributable to three systematic problems. The first is simply parse errors: although modern automatic syntax parsers are fairly robust, parsing is a complicated process and mistakes are still made; and this is compounded by the fact that digital archives often contain noise.

The second group of errors is caused by the fact that clauses are seen as separate and non-overlapping, while in reality clauses in a sentence are generally connected. For example, take the (abridged) sentence *'Israel [..] vowed to destroy every building used by Hamas [..]'*, from which the system identifies two clauses: (Israel / vows to destroy) and (Hamas / used buildings). Another example illustrates cases where one actor calls on another actors to (not) do something, or accuses the other actor of doing something: *'The carnage [..] drew mounting international pressure for Israel to end the offensive against Hamas.'* In this case, Israel is part of the predicate of the clause about international pressure, but it is not identified as the subject of the subordinate clause about the offense.

Apart from causing identification errors, treating clauses as independent also discards potentially interesting information. For example, in the sentence *'European leaders have promised technical assistance to stop Hamas shipping weapons into Gaza'*, currently no relation between the EU and Hamas is extracted since the subordinate clause (Hamas / ships weapons) is excluded from he main clause (EU / promises assistance). Correctly identifying the nature of the relation between clauses can potentially add a lot of interesting information to the analysis.

A third group of errors is caused by the subject or object of an action being expressed by prepositional phrases. For example, in the sentence *'Israeli air strike in Hamas-controlled Gaza killed [..]'*, Hamas is seen as part of the subject since the 'in [..] Gaza' prepositional phrase is attached to the air strikes, not the killing. Conversely, in the sentence *'[Israel] is concerned about Tel Aviv coming within range of rockets launched from Gaza'*, Gaza is seen as part of the predicate rather than as the actor of the rocket launches.

For the extracted sources, many of the missed quotes are due to two reasons. A number of sentences express quotes in more 'creative' ways, for example paraphrasing Hamas in the sentence *'It added that Hamas reports of kidnapping an Israeli soldier stemmed from [..]'*. A second pattern missed by the current system is sentences where the source is not expressed using words, but only with punctuation. Although it would be easy to detect quotation marks in the source text, care would need to be taken since quotation marks and colons are also used for other purposes, and it is not always trivial to identify the source of the quote.

Substantive use case: Using clause analysis to analyze international framing of the 2008–2009 Gaza War

This section will provide an additional validation of the clause method. The previous section showed that the clause method can extract

source, subject, and predicate of sentences with sufficient accuracy and that it substantially outperforms a word-order baseline method. By using clause analysis to analyze the foreign coverage of the 2008–20009 Gaza War, this section will show how the method can be used to extract plausible and meaningful substantive results. Moreover, these results will be compared to results from the baseline method to determine whether the better sentence-level accuracy of the clause method translates to better substantive results at the macro level. Finally, this section will show how a regular corpus analytic technique, Semantic Network Analysis (Van Atteveldt, 2008), can be used to analyze the predicates yielded by clause analysis.

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.'[7] The operation started with a wave of airstrikes 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. 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. 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, in which frame building contests play a central role (Sheafer and Shenhav, 2010). Based on the homophily hypothesis we expect U.S. media to be more receptive to the Israeli frame building effort than Chinese media because of their differing proximity to Israel: The US is one of the most proximate countries to Israel both culturally and politically, while China is one of the most distant countries to Israel on these dimensions (Sheafer and Shenhav, 2010; Sheafer et al., 2014). Thus, if there are differences in framing to be found anywhere it would be between the press in these countries.

[7] Israeli Ministry of Foreign Affairs, `http://www.mfa.gov.il/GazaFacts` [accessed July, 2010].

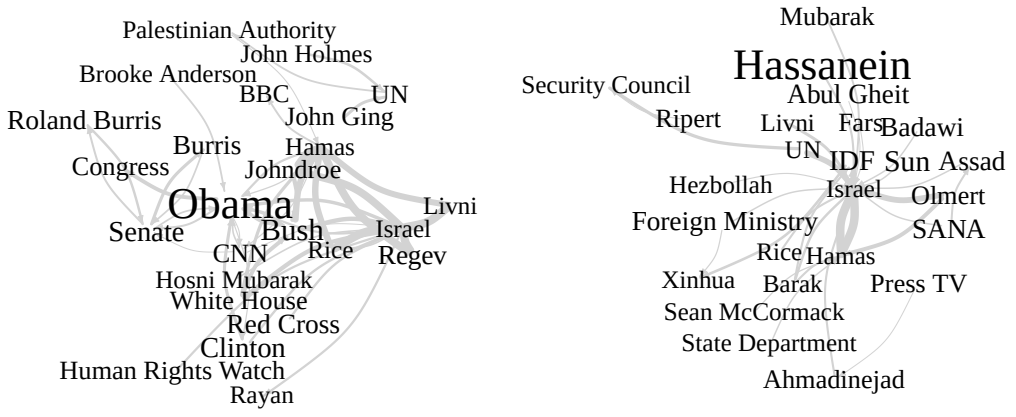
Table 3: *Usage of Hamas and Israel as source and overrepresentation of Israel in all clauses and in clauses containing aggression; in U.S. and Chinese media coverage of the 2008–09 Gaza War*

Source	U.S. Media		Chinese Media	
	All	Aggr.	All	Aggr.
Israel	8,224	4,126	1,225	677
Hamas	1,824	704	439	193
Overrepresentation Israel	4.5	5.9	2.8	3.5

For this analysis, all articles mentioning 'Gaza' published between December 27, 2008, and January 20, 2009 in all available English language sources from the U.S. and China were downloaded from LexisNexis and processed using the clause method described above. In total, this yielded 795,620 clauses from 2,176 Chinese and 6,475 American articles. In these clauses, the keyword lists described above were used to identify references to Israel, Hamas, and aggression.

Who is quoted?

The first question we can answer is whether there is a citation bias: is either actor quoted more often in the U.S. media compared to the Chinese media? Table 3 shows how often Hamas and Israel occur as source of statements. The first two columns list the number of time each actor was identified in the source in the U.S. media, listing all clauses and only those clauses containing an aggression word, respectively. The next two columns give the same information for the Chinese media. As shown by the third row, Israel is overrepresented as a source in the media of both countries. As expected based on the cultural and



(a) *U.S. Media* **(b)** *Chinese Media*

Figure 3: *Citation Network in Chinese and U.S. media coverage of the 2008–2009 Gaza war*
Note: N=81,592 (U.S.) / 12,141 (Chinese) unique quotes. Arrow width indicates how often the source node mentioned the target node in a quote. Label size indicates overrepresentation in that country's media.

political proximity of the U.S. to Israel, this overrepresentation is much larger in the U.S. media than in the Chinese media (4.5 versus 2.8). For the clauses containing aggression, this is even more pronounced, with Israel being used as a source almost six times as often in the U.S. media, compared to 3.5 times as often in the Chinese media. Chi-squared tests show that both differences are highly significant ($\chi^2(1) = 61.5.7, p < .01$ for all clauses, $\chi^2(1) = 31.6, p < .01$ for aggressions). Thus, these results show that although both countries favor Israeli sources, the U.S. media show a more pro-Israeli citation bias than the Chinese media.

Who talks about whom?

Besides this quantitative result we can also take a more qualitative look at the identified sources. Figure 3 shows the citation network extracted from the U.S. and Chinese media, where the edge width indicates how often the source actor mentioned the target actor in a quote. To show the difference between the media in both countries, the network only shows the actors that are overrepresented as source in that country, including other actors only if they are mentioned by these overrepresented actors. In the displayed network, node size indicates overrepresentation, and backbone extraction is used to limit the amount of edges shown.[8]

Looking at the citation patterns, we can see that in the U.S. media Israel and Israeli actors such as Tzipi Livni and Mark Regev are frequently given a podium to talk about Hamas. Besides these Israeli actors, most of the cited actors are U.S. politicians, especially president Obama. In the Chinese network we see that there is more room for Palestinian sources to discuss Israel as well as the other way around. The most overrepresented source is Moaweya Hassanein, the Palestinian health minister who made frequent statements about (Palestinian) casualties. Additionally, although American sources such as the State Department and Secretary of State Condolleeza Rice are also given a podium, the Chinese media quote a much wider range of international actors, including the Syrian and Iranian presidents and Egyptian diplomat Abul Gheit.

[8] All graph analyses and visualization were made in R using `igraph` and `semnet` (`https://github.com/kasperwelbers/semnet`).

Table 4: *Number of times Hamas and Israel are mentioned as aggressor or victim in U.S. and Chinese media coverage of the 2008–2009 Gaza war*

Role	U.S. Media				Chinese Media			
	Israel		Hamas		Israel		Hamas	
	N	%	N	%	N	%	N	%
Aggressor	16,028	46%	6,613	43%	3,312	46%	701	34%
Victim	18,708	54%	8,859	57%	3,953	54%	1,378	66%
Total	34,736	100%	15,472	100%	7,265	100%	2,079	100%

Who does what to whom?

Next we turn to the difference in framing the war: who is the aggressor, and who is the victim? In modern public diplomacy, belligerent parties will generally try to frame the other as the aggressor, focusing on the means employed by the opponent and framing their own actions in terms of the goals they are meant to achieve (Sheafer et al., 2014). A first indication of framing bias between the Chinese and U.S. media can be gauged by looking at how often the two belligerent parties are identified in the subject and predicate of aggression clauses.

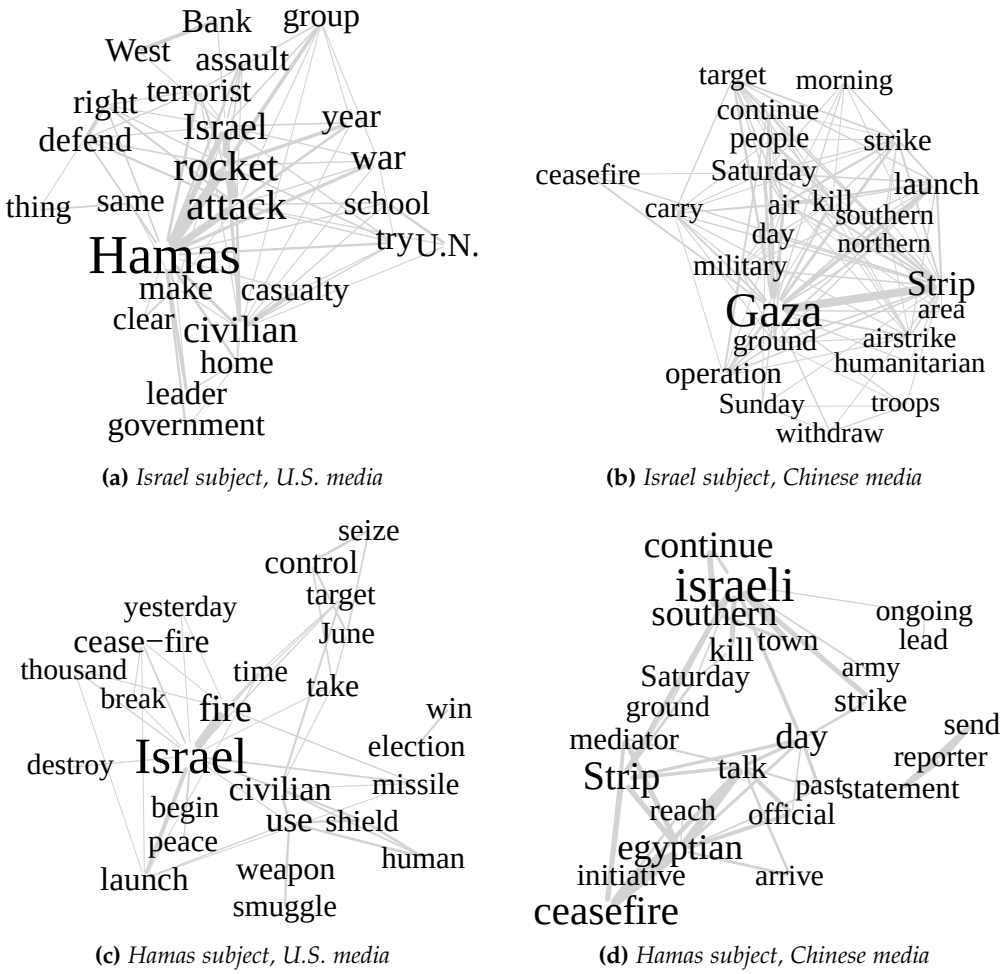
Table 4 shows how often Hamas and Israel are identified in the subjects and predicates of clauses containing aggressions, indicating how frequently they are framed as aggressor and as victim. As shown in the first four columns, in the U.S. media Hamas and Israel are portrayed as subject of aggression with approximately the same frequency (43% and 46%, respectively). In the Chinese media Israel is portrayed as aggressor with the same frequency (46%). Hamas, however, is much more often portrayed as victim of aggression (66% of clauses mentioning Hamas). A chi-squared test confirms that this difference in role for Hamas between the U.S. and Chinese media is highly significant ($\chi^2(1) = 61.0, p < .01$). Given that being portrayed as the perpetrator of aggression is generally undesirable, and that a central part of the Israeli framing of the war was that they were the victim of Hamas' rocket attacks, the U.S. media have a more pro-Israeli framing of the conflict than the Chinese media.

And what do they do?

Besides looking at the frequency of statements, we can analyze 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. We can display this as a co-occurrence based semantic network, linking words that frequently co-occur in the predicate of a clause. Similar to the citation network above, these networks are contrasted by taking the words that are overrepresented in the media coverage of each country, and backbone extraction is used to simplify the network.

Figure 4 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. If we look at the words in predicates where Israel is subject according to the U.S. media (4a), we see that the main cluster is about the rocket assault (by) Hamas, and the (Israeli) right to defend. A separate cluster talks about allowing supplies. In the Chinese media (4b), 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 the human and humanitarian consequences.

A similar difference is found in the words in predicates with Hamas as the subject. In the U.S. media (4c), focus is on actions against civilians like launching missiles at civilians and the use of civilians as human shields, and the smuggling of weapons. Cease-fire is mentioned, but co-occurs with breaking. In the Chinese media on the other hand (4d), these ac-



(a) *Israel subject, U.S. media*

(b) *Israel subject, Chinese media*

(c) *Hamas subject, U.S. media*

(d) *Hamas subject, Chinese media*

Figure 4: *Most frequent words in predicates with Hamas and Israel as subject in Chinese and U.S. media, visualized as co-occurrence network*

Note: N=30,921 (a), 5,122 (b), 13,782 (c) 1,500 (d) unique clauses (560,603 words in total). Label size shows overrepresentation in predicates for the subject in that country compared to the other country. Edges indicate co-occurrence within predicates.

tivities are not mentioned, and more attention is devoted to attacks on the Israeli army. Also, there is a large cluster devoted to diplomatic efforts, including mentioning the cease-fire.

As above, we see here that the U.S. media are more pro-Israeli, focusing on the supposed terrorist activity of Hamas as a reason for the Israeli actions (Israeli goals discourse), while the Chinese media concentrate on the Israeli military actions themselves (Israeli means discourse) and on possible diplomatic solutions to the conflict.

Comparison to Baseline

To determine whether the superior per-sentence accuracy of the clause method compared to the baseline presented in the previous section translates to better substantive results, the analyses above were also performed for the baseline method.[9] According to the base-

[9]Full results of the baseline methods on the Gaza war coverage are included in the online supplemental materi-

line source analysis, as shown in Table A1 in the (online) supplemental materials, the Chinese media actually show a larger overrepresentation of Israeli sources, with Israeli sources quoted 3.4 times more often than Hamas compared to 2.6 for U.S. media. Given the theoretical considerations and previous findings in the literature, these results are much less plausible than the results of the clause method.

On the analysis of predicates the baseline results are more in line with the results of the clause analysis (Table A2): Israel is portrayed as aggressor more often in both U.S. (56%) and Chinese media (60%), while Hamas is most often portrayed as victim (55% in U.S. vs. 61% in Chinese media). Assuming that the errors made by the baseline method are not biased towards either actor, it is not surprising that in the aggregate it finds similar patterns as the clause analysis method. However, the lower accuracy reported in the previous section implies that if we use the method to make more fine-grained selections it will show more errors. This is confirmed by a manual inspection of sentences where the two methods differ show that in most cases the baseline method is mistaken. An illustrative example is the following sentence: *Four Israelis have been killed by Hamas rockets in three days* (CNN, 29 December 2008). According to the baseline, Israel is the aggressor (since it occurs to the left of 'killed') and Hamas is the victim, while the clause method correctly analyses the passive structure and identifies Hamas rockets as the semantic subject.

Finally, Figure A1 shows the semantic networks of the the predicates extracted with the baseline method. Although the overall vocabulary is similar to the results of the clause analysis, there are some remarkable differences: in the U.S. media, references to Israel's right to defend are missing, as are the use of human shields and weapons smuggling by Hamas, while in the Chinese media the references to Hamas' preferred diplomatic solution are also missing. All in all, these results show that using the baseline method we would reach the same conclusion as using clause analysis with respect to the framing bias of aggressor and victim. However, the baseline method has a much less plausible result for the source extraction, and the strong pro-Israeli framing elements are absent from the semantic network analysis of U.S. coverage.

als.

Conclusion

This article presents a method and open source software package for using grammatical analysis to automatically extract *clauses* from newspaper articles, splitting sentences into subject–predicate clauses and identifying a possible quoted or paraphrased source. This uses the output of grammatical analysis to extract 'who does what to whom', making it possible to answer questions that are difficult to answer using frequency or co-occurrence based methods. A comparison with a manually coded gold standard shows that both clauses and quotes can be identified accurately, and that using the syntactic rules substantially outperforms a word-order based system.

In order to show how the results of clause analysis can be used for substantive analysis, we analyzed all clauses from the coverage of the Gaza war in U.S. and (English-language) Chinese media. This showed that the U.S. media more often quote Israeli and American political sources, while the Chinese media give more platform to Palestinian sources and global leaders and diplomats. This suggests a pro-Israeli citation bias in the U.S. news, which was also confirmed by looking at how Israel and Hamas are portrayed: Relative to the Chinese media, U.S. media portray Hamas more often as aggressor. Moreover, the U.S. media focus more on the Israeli goals, while the Chinese media emphasize the military means. These findings are in accordance with the expectation based on the greater cultural and political proximity of the US to Israel compares with China's proximity to Israel (Sheafer et al., 2014). These results were also more plausible than results from applying the word-order baseline method to the same case, which found a reverse citation

bias and missed some of the strong framing elements found in the semantic network analysis.

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, resulting in an enriched token list or bag of words. From here, normal corpus linguistic tools and techniques can be applied to these results, using the clause information to split and filter the data in interesting ways. The substantive results presented in the second half of this study are only a small sample of what is possible. For example, it is equally possible to use techniques such as topic modeling, machine learning, or sentiment analysis to further analyze the words on the dyads between actors in the subject and predicate positions, resulting in a labeled semantic network; or to identify which actors are most able to determine the framing in the news by analyzing the frames used in the quotes of each source, and how these frames are diffused through the news.

These results notwithstanding, there are a number of ways in which the system can be improved. One obvious limitation is that the system described here is designed for English. However, since the only language dependency is in the relatively simple syntax patterns it is easy to translate the results to different languages as long as a good dependency parser is available. In earlier work, we have shown that a similar set of rules also work for quote extraction in Dutch (Van Atteveldt, 2013). Moreover, it is likely that word-order approaches will work even worse in languages like German, which have a less strict word order and often more complex sentence structures.

The error analysis performed as part of the validation also highlighted the strong and weak parts of the system. Part of the errors were due to non-verbal action patterns such as 'the Israeli invasion of Gaza'. Since the system depends on verbal patterns to extract predicates, such nominal patterns are only dealt with by a special pattern relying on a dictionary of actions, and identify the subject correctly only if it is listed in a possessive structure such as in this example. This can possibly be alleviated with either better lexical resources or with a machine learning system such as presented by Fogel-Dror et al. (2015), but both solutions lack the simplicity and transparency of the clause analysis presented here. A second class of errors was caused by relations between clauses. The current system analyses all clauses in isolation, e.g. if 'Israel invaded Gaza after Hamas fired rockets at civilians', it recognizes that Israel invaded Gaza, and Hamas fired rockets, but no link is found between these two statements. Since such discursive relations are often expressed by very explicit markers (such as 'after', 'because', 'causes'), it would be very interesting to develop rules for these patterns as well.

In sum, clause analysis is a useful addition to the text analysis toolkit for analyzing texts in which multiple actors can make different statements about other actors and issues. The limitations discussed above show that the method is especially useful if a researcher is mainly interested in source use and relations that are expressed as verbs, and if the researcher is interested in these direct relations rather than in the overall argumentative structure of a text. Within these limitations, clause analysis opens up possibilities for sophisticated analysis of statements and actions in political text.

References

Baker, C., C. Fillmore, and B. Cronin (2003). The structure of the framenet database. *International Journal of Lexicography 16*(3), 281–296.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). Latent dirichlet allocation. *the Journal of machine Learning research 3*, 993–1022.

Carreras, X. and L. Màrquez (2005). Introduction to the conll-2005 shared task: Semantic role labeling. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, pp. 152–164. Association for Computational Linguistics.

Chen, D., N. Schneider, D. Das, and N. A. Smith (2010). SEMAFOR: Frame argument resolution with log-linear models. In *Proceedings of the 5th international workshop on semantic evaluation*, pp. 264–267. Association for Computational Linguistics.

Collingwood, L. and J. Wilkerson (2012). Tradeoffs in

accuracy and efficiency in supervised learning methods. *Journal of Information Technology & Politics 9*(3), 298–318.

De Marneffe, M., B. MacCartney, and C. Manning (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, Volume 6, pp. 449–454.

D'Orazio, V., S. T. Landis, G. Palmer, and P. Schrodt (2014). Separating the wheat from the chaff: Applications of automated document classification using support vector machines. *Political Analysis*, mpt030.

Entman, R. M. (2008). Theorizing mediated public diplomacy: The U.S. case. *International journal of Press/Politics 13*, 87–102.

Fellbaum, C. (Ed.) (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Fogel-Dror, Y., T. Sheafer, S. R. Shenhav, and W. van Atteveldt (2015). Real-time sentiment analysis in the context of a political conflict. In *Annual Meeting of the American Political Science Association. San-Francisco, CA*.

Grimmer, J. (2010). A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis 18*(1), 1–35.

Grimmer, J. and B. M. Stewart (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis 21*, 267–297.

Grossman, D. A. and O. Frieder (2012). *Information retrieval: Algorithms and heuristics*, Volume 15. Springer Science & Business Media.

Hillard, D., S. Purpura, and J. Wilkerson (2008). Computer assisted topic classification for mixed methods social science research. *Journal of Information Technology and Politics 4*(4), 31–64.

Kellstedt, P. M. (2003). *The mass media and the dynamics of American racial attitudes*. Cambridge University Press.

Laver, M., K. Benoit, and J. Garry (2003). Extracting policy positions from political texts using words as data. *American Political Science Review 97*(2), 311–331.

Lowe, W. and K. Benoit (2013). Validating estimates of latent traits from textual data using human judgment as a benchmark. *Political Analysis*, mpt002.

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

Monroe, B. L., M. P. Colaresi, and K. M. Quinn (2008). Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis 16*(4), 372–403.

Quinn, K. M., B. L. Monroe, M. Colaresi, M. H. Crespin, and D. R. Radev (2010). How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science 54*(1), 209–228.

Roberts, M. E. (2015). Introduction to the virtual issue: Recent innovations in text analysis for social science. *Political Analysis 23*, 254–277.

Ruigrok, N. and W. Van Atteveldt (2007). Global angling with a local angle: How U.S., British, and Dutch newspapers frame global and local terrorist attacks. *The Harvard International Journal of Press/Politics 12*, 68–90.

Schrodt, P. A. (2014). TABARI: Textual Analysis by Augmented Replacement Instructions, version 0.8.4b3. `http://eventdata.parusanalytics.com/software.dir/tabari.html`.

Schrodt, P. A. and D. J. Gerner (1994). Validity assessment of a machine-coded event data set for the Middle East, 1982–1992. *American Journal of Political Science 38(3)*, 825–854.

Schrodt, P. A. and D. J. Gerner (2000). Cluster-based early warning indicators for political change in the contemporary levant. *American Political Science Review 94(4)*, 803–818.

Schrodt, P. A., D. J. Gerner, and O. Yilmaz (2005). Using event data to monitor contemporary conflict in the israel-palestine dyad. *International Studies Perspectives 6*(2), 235–251.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys 34*(1), 1–47.

Sheafer, T. and S. R. Shenhav (2010). Mediated public diplomacy in a new era of warfare. *The Communication Review 12*, 272–283.

Sheafer, T., S. R. Shenhav, J. Takens, and W. Van Atteveldt (2014). Relative political and value proximity in mediated public diplomacy: The effect of state-level homophily on international frame building. *Political Communication 31*(1), 149–167.

Slapin, J. B. and S.-O. Proksch (2008). A scaling model for estimating time-series party positions from texts. *American Journal of Political Science 52*(3), 705–722.

Stone, P. J., D. C. Dunphy, M. S. Smith, D. M. Ogilvie, and Associates (1966). *The General Inquirer: A Computer Approach to Content Analysis*. Cambridge, MA: MIT Press.

Van Atteveldt, W. (2008). *Semantic Network Analysis: Techniques for Extracting, Representing, and Querying Media Content (dissertation)*. Charleston, SC: BookSurge.

Van Atteveldt, W. (2013). News media: platform or power broker? a study of political quotes in newspaper content using syntactic analysis. In *Presented at the New Directions in Analyzing Text as Data Workshop, LSE, 27–28 September*.

Van Atteveldt, W., J. Kleinnijenhuis, and N. Ruigrok (2008). Parsing, semantic networks, and political authority: Using syntactic analysis to extract semantic relations from Dutch newspaper articles. *Political Analysis 16*(4), 428–446.

Van Atteveldt, W., T. Sheafer, S. R. Shenhav, and Y. Fogel-Dror (2016). Replication data for: Clause analysis: Using syntactic information to automatically extract source, subject, and predicate from texts with an application to the 2008-2009 gaza war. doi:10.7910/DVN/DZZXAD, Harvard Dataverse, V1 [UNF:6:IdSlgh3RYlPHO1Hq0pCahQ==].

Young, L. and S. Soroka (2012). Affective news: The automated coding of sentiment in political texts. *Political Communication 29*(2), 205–231.