Online Tools for Content Analysis

In: The SAGE Handbook of Online Research Methods

By: Roel Popping

Edited by: Nigel G. Fielding, Raymond M. Lee & Grant Blank
Pub. Date: 2017
Access Date: August 29, 2019
Publishing Company: SAGE Publications Ltd
City: 55 City Road
Print ISBN: 9781473918788
Online ISBN: 9781473957992
DOI: https://dx.doi.org/10.4135/9781473957992
Print pages: 329-343

Online Tools for Content Analysis

Roel Popping

Introduction

Content analysis is a systematic reduction of a flow of text to a standard set of statistically manipulable symbols representing the presence, the intensity, or the frequency of some characteristics, which allows making replicable and valid inferences from text to their context. In most situations the source of the text is investigated, but this should not exclude the text itself, the audience or the receivers of the text. It involves measurement. Qualitative data are quantified for the purpose of affording statistical inference. In the past it was performed by human coders. Based on a so-called codebook these coders noted whether what was mentioned in the variables involved was found in the texts under study or not.

This changed very much when the computer came in in the 1960s. The software, the General Inquirer, was an aide for the quantification of texts and transcripts, it could code faster and more consistently than humans (Stone *et al.*, 1966). Now the emphasis came to be on the occurrence of particular themes. A list of search entries per theme was to be developed. These search entries are words or phrases that are understood as indicating the occurrence of the corresponding theme. The themes and their search entries are kept in a so-called dictionary. Texts and dictionary are input for the software to be used, output is a data matrix with the themes in the columns and the texts in the rows. In a cell, one finds how many times the theme was mentioned in the corresponding piece of text.

At first, content analysis was used primarily to draw conclusions regarding the source of the message. Communication is broader, however; it also involves the message, the channel, and the audience. The four aspects of communication represent the most common contextual variables used in content analysis. A fundamental characteristic of content analysis is that it is concerned with the communicative act post hoc.

Researchers investigating the message usually look at the occurrence of specific themes in the texts. Today this approach is known as the thematic approach. A dictionary is often used, which informs how search entries, words or word phrases refer to themes one is interested in as they occur in a text. The frequency of occurrences or co-occurrences of themes is the basis for the analyses that should give an answer to the research question. This is the approach that is followed in most research and it can be performed automatically.

Language can be very ambiguous – to start with, a word can have different meanings. The meaning of a word can even change over time; however, the correct meaning usually becomes clear when the context is taken into account. Software only recognizes search entries; therefore, the dictionary should contain guidelines to overcome potential problems. The word ‘bank’ for example is often rephrased as bank#1 to indicate the place where one brings his or her money, as bank#2 to refer to the edge of a river, and eventually as bank#3 to refer to that one can sit on (as in ‘banks of seats'). There are more possibilities as will be shown later.

Two other approaches also receive increasing attention: the semantic and the network approach. These approaches involve not only the identification of alternative themes, but also the encoding of relations among themes in texts. These relational methods for encoding texts are strikingly similar. In each case, a Subject–Verb–Object (S–V–O) syntax is applied during the encoding process, or even a Subject–Valence–Verb–Object (S–V–V–O) syntax, in which the valence can reflect negation, evaluation, intensity, etc. These ‘clause-based content analyses’ afford inferences about how texts’ sources use words in their speech or writings. The methods associated with the new approaches differ primarily according to the research purposes to which each one's relationally encoded texts can be applied. In the semantic approach, variables indicate interrelations that themes may have in texts. The network approach methodologies afford variables that characterize entire networks of semantically related themes.

In all three approaches texts are coded, and in all three the researcher might understand the texts instrumentally; they might be interpreted in terms of the researcher's theory. This is called the *instrumental* approach (Shapiro, 1997). This approach is generally followed. It is differentiated from the *representational* approach, where the source's perspective is used to interpret the texts under study. Here the intended meaning of the source must be identified. This usually demands an interpretation by a human coder. Regardless of whether the source's or researcher's perspectives are used in interpreting texts, these perspectives must be made explicit for the reader to evaluate the validity of conclusions that are made.

Many researchers who prefer machine coding based on a dictionary follow the instrumental approach. This method of coding is very fast and it is therefore no problem to analyse all available texts. But one needs to be cautious because it often looks as if the texts used are an ad hoc population of texts. The population should at least be indicated and motivated. This concern is especially relevant to analyses of blogs or Tweets that become available (see also Hookway and Snee, this volume). On the other hand, the volume of texts can never be an argument against manual coding. Sampling is the solution to this ‘problem'.

This approach, which focuses on generating a data matrix to be used in statistical analysis, differs from qualitative analysis, which is a collective noun for various approaches. Software for this type of research focuses on theory building, text base management, coding and retrieving for descriptive and interpretative analysis. It is not directed to quantification for statistical inferences. Berelson (1952: 116ff) makes the following two remarks with regard to qualitative analysis:

> (1) Much ‘qualitative’ analysis is quasi-quantitative… Just as quantitative analysis assigns relative frequencies to different qualities (or categories), so qualitative analysis usually contains quantitative statements in rough form. They may be less explicit but they are nonetheless frequency statements about the incidence of general categories… (2) ‘Qualitative’ analysis is often based upon presence–absence of particular content (rather than relative frequencies).

Content analysis is also used in the fields of linguistics and information retrieval. This will not be considered here.

My goal is to describe more recent developments within the three approaches mentioned. Here we will see that coding can be performed from the two perspectives just mentioned; we will also see that the computer plays an important role in the process each time. The three approaches – thematic, semantic and network – will subsequently be discussed, each initially regarding manual coding, and then machine coding. Extra attention will be given to the so-called modality analysis, which is seen as a very challenging development. Finally, some issues that might affect the use of content analysis will be outlined.

Thematic Approach to Content Analysis

Thematic content analysis is the term for any content analysis in which variables indicate the occurrence (or frequency of occurrence) of particular themes or concepts. In this section we specify what themes and concepts are, followed by how these can be recognized in a text, but we also cover difficulties with the approach. In the sub-sections, machine coding and manual coding are discussed, as well as the problem of ambiguity in texts. Machine learning is dealt with as part of machine coding.

A concept is ‘a single idea, or ideational kernel, regardless [of whether] it is represented by a single word or a phrase’ (Carley, 1993: 81). Practitioners of thematic content analysis usually reserve the term ‘theme’ for broader classes of concepts. The theme is usually concentrated on a specific referent (e.g. the president, the U.S., British foreign policy, communism). Thematic content analysis allows the researcher to determine what, and how frequently, themes (co-) occur in texts. The method is particularly useful when the researcher is interested in the prominence of various themes in texts, possibly reflecting broad cultural shifts. With respect to a certain research question, therefore, one also needs context variables between which perspectives can be compared, for example the (type of) newspaper in which the text has been published. The first software for content analysis was designed for thematic content analysis based on a dictionary. The analysis is based on a deductive rule-based approach to operationalization. The approach can only be used effectively if a complete theory is available of how the theoretical themes of interest manifest themselves in natural language. This theory, which is the researcher's theory, becomes visible in the dictionary. For this reason, the instrumental approach to coding is followed.

The dictionary-based methods have hardly changed since the development of the first software. For these methods to work well, the scores attached to words must closely align with how the words are used in a particular context. If a dictionary is developed for a specific application, then this assumption should be easily justified. But when dictionaries are created in one substantive area and then applied to another, problems can occur. Dictionaries, therefore, should be used with substantial caution. Scholars must either explicitly establish that word lists created in other contexts are applicable to a particular domain, or create a problem-specific dictionary. In either instance, scholars must validate their results. However, measures from dictionaries are rarely validated, and instead standard practice in using dictionaries is to assume the measures created from a dictionary are correct and then apply them to the problem.

In thematic content analysis one can report occurrences and co-occurrences of themes. Occurrences indicate the prominence of themes. When compared across contexts they can afford inferences about culture's changing themes, ideas, issues and dilemmas or differences between media in representing news content about the same issue. Looking at co-occurrences means looking at associations among themes. This analysis is known as contingency analysis. In this type of analysis, the goal is to calculate associations among occurrence measures and to infer what the resulting pattern of association means. Problems occur if these inferences are about how themes are related. Assume the following text block is investigated: ‘The man likes detective stories, but his wife prefers love themes'. The themes MAN (represented by ‘the man') and LOVE THEME (represented by ‘love themes') co-occur in this block, but no relation between the two is specified. For such inferences relations should have been encoded a priori, not via ad hoc post hoc looks at the texts. Texts should therefore be divided into distinct blocks, for example chapters, paragraphs, sentences, clauses (sentences or part-of-a-sentence that explicitly or implicitly contain an inflected verb, an optional subject and/or object, plus all modifiers related to this verb, subject, and object). Now the co-occurrences can be investigated on the level of the clause. If not, this would lead to ecological fallacy in the interpretation of the data because inferences about the nature of clauses would be deduced from inference from the text to which the clause belongs.

An example of a thematic text analysis is found in Namenwirth (1969) who studied differences between British prestige and mass newspapers with respect to some orientational dimensions. The orientations are considered as marks of distinction with respect to the newspapers, but also with respect to cognitive styles of elites and masses in general. Lots of themes have been investigated. To reduce their number in the analysis, principal component analysis has been applied.

Machine Coding

Looking at publications in which text analysis is used, it turns out that most investigators use the machine to do the coding based on a dictionary. Researchers using full automated software are imposing the software developer's theoretical perspective on the speaker's/author's words. If that is what they intend, then its developers need to have reduced this theory to concrete algorithms in their content analysis software and to have made their theory–algorithm relations clear to users (otherwise users might end up naively believing that the software has somehow ‘revealed’ a perspective-free [i.e. incontestably true] rendering of the texts).

Today, some software offers the possibility to enter one's own dictionary, others use dictionaries provided by the developer or by a team around this developer. Constructing a dictionary is a challenging task. The dictionary should be valid; this implies among others that maintenance deserves a lot of attention. Dictionaries usually cover specific fields.

An example of a dictionary that is well maintained by the researcher is the one that is part of the LIWC software (Linguistic Inquiry and Word Count) for measuring people's physical and mental health (Pennebaker *et al.*, 2003). The software allows users to determine the degree to which any text uses positive or negative emotions, self-references, causal words and 70 other language dimensions.

In recent years, there have been two innovations that receive serious attention: supervised learning and analysis of co-occurrence of words. Both are extensions of the thematic instrumental approach.

Machine Learning

Researchers using dictionaries follow a deductive approach. Today, however, inductive approaches are also becoming available. This happens when machine learning is used. A machine learning algorithm is trained with known data and derives rules by which the given decisions can be reproduced. An algorithm takes texts and their correct coding assignments as inputs, derives a ‘probabilistic dictionary’ (Pennings and Keman, 2002) from this data, and uses this information for the coding of new texts.

The method is a purely statistical approach, which can be used in any language and with any topic category. No assumptions are made about syntax; any text is treated as a simple bag of words. The approach is solely based on superficial, i.e. lexical, features of a text and the assumption that single words or word combinations provide enough information for thematic coding. The training process resembles conventional coder training; it is heavily based on example documents. The computer classifier is treated like any human coder, but with limited language skills and no contextual knowledge. The method is seen by its users as an ideal complement and extension to classic thematic content analysis. Compared to traditional methods of automated content analysis, supervised learning does not require different operationalization strategies; however, one has to remember that any automatic classification is only as good as its training material. Making a correct decision often depends on a lot of context knowledge.

For more details on the methods used in machine learning, see Grimmer and Stewart (2013). Schrodt (2012) reports on actual software developments. The Comparative Party Manifesto Project, in which different aspects of party performance as well as the structure and development of party systems is studied, is an example of where this method is used. The project is based on quantitative content analyses of parties' election programmes from many countries. Laver *et al.*, (2003) contains an extended introduction into this project, including a mathematical model of the way the learning process works. A kind of state-of-the-art of the project is presented in Geminas (2013).

Analysis based on Co-occurrence

One way to reduce the amount of information in texts has been to apply principal component analysis or multidimensional scaling based on co-occurring words or co-occurring themes. In this way, Miller (1997) could demonstrate stakeholder influence on news and patterns of change in frames across time.

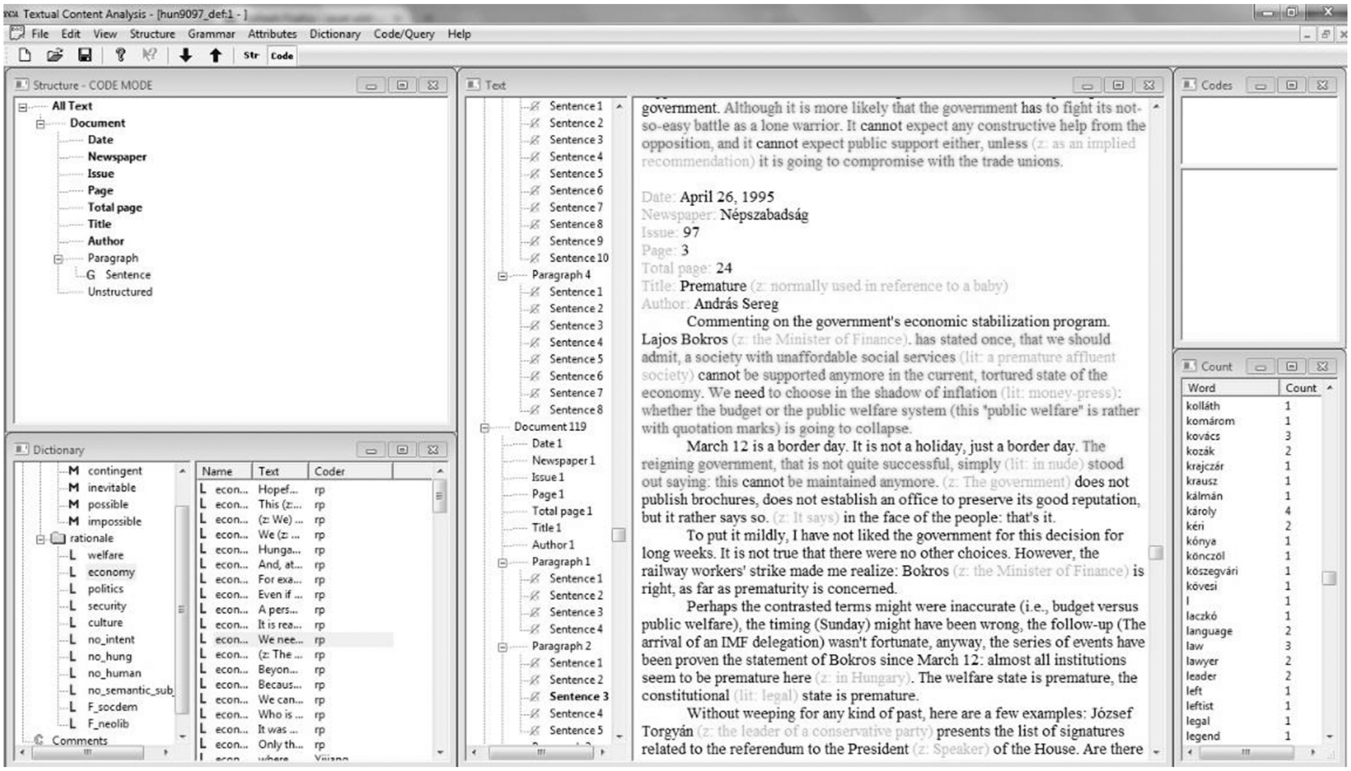
Today natural language processing is becoming used to analyse texts. One method is Latent Dirichlet Allocation (LDA), a hierarchical Bayesian technique that automatically discovers topics that these texts contain. It is based on the idea that each document is a mixture of a small number of topics and that each word is attributable to one of these topics (Blei *et al.*, 2003). Following an iterative process, a descending hierarchical classification method decomposes classes until a predetermined number of iterations fails to result in further divisions. The result is a hierarchy of classes, which can then be schematized as a tree diagram.

An example of a study using such a technique is Schonhardt-Bailey (2005), who analysed speeches on National and Homeland Security by presidency candidates Bush and Kerry in 2004. The software found seven classes of groups of words that could be labelled. The groups could further be reduced to two groups, one containing speeches that were especially US-specific, the other containing speeches dealing with the global order. The first group contained nearly all the speeches by Kelly, the other group speeches by Bush.

Manual Coding

Most machine coding is based on a dictionary. A dictionary usually does not catch the latent meaning of text, and usually there will also be problems with ambiguous texts. These problems might be solved when manual coding is applied, at least as a supplement to machine coding. This way of coding became possible from the moment operating systems and programs made it possible that complete texts could be made visible on the screen of the computer or terminal and allowed users to indicate which search entry or which sentence had been coded. At first, this software was developed for performing qualitative research, but later software for quantitative research also appeared. Figure 19.1 shows what appears on the screen when the software Textual Content Analysis (TCA) for manual coding is used (Roberts, 2008).

Figure 19.1Example Textual Content Analysis window



Name	Text	Coder
L econ...	Hopef...	rp
L econ...	This (z:...	rp
L econ...	(z: We) ...	rp
L econ...	We (z: ...	rp
L econ...	Hunga...	rp
L econ...	And, at...	rp
L econ...	For exa...	rp
L econ...	Even if ...	rp
L econ...	A pers...	rp
L econ...	It is rea...	rp
L econ...	We nee...	rp
L econ...	(z: The ...	rp
L econ...	Beyon...	rp
L econ...	Becaus...	rp
L econ...	We can...	rp
L econ...	Who is ...	rp
L econ...	It was ...	rp
L econ...	Only th...	rp
L econ...	where...	Yiiang

government. Although it is more likely that the government has to fight its not-so-easy battle as a lone warrior. It cannot expect any constructive help from the opposition, and it cannot expect public support either, unless (z: as an implied recommendation) it is going to compromise with the trade unions.

Date: April 26, 1995
Newspaper: Népszabadság
Issue: 97
Page: 3
Total page: 24
Title: Premature (z: normally used in reference to a baby)
Author: András Sereg

Commenting on the government's economic stabilization program. Lajos Bokros (z: the Minister of Finance). has stated once, that we should admit, a society with unaffordable social services (lit: a premature affluent society) cannot be supported anymore in the current, tortured state of the economy. We need to choose in the shadow of inflation (lit: money-press): whether the budget or the public welfare system (this "public welfare" is rather with quotation marks) is going to collapse.

March 12 is a border day. It is not a holiday, just a border day. The reigning government, that is not quite successful, simply (lit: in nude) stood out saying: this cannot be maintained anymore. (z: The government) does not publish brochures, does not establish an office to preserve its good reputation, but it rather says so. (z: It says) in the face of the people: that's it.

To put it mildly, I have not liked the government for this decision for long weeks. It is not true that there were no other choices. However, the railway workers' strike made me realize: Bokros (z: the Minister of Finance) is right, as far as prematurity is concerned.

Perhaps the contrasted terms might were inaccurate (i.e., budget versus public welfare), the timing (Sunday) might have been wrong, the follow-up (The arrival of an IMF delegation) wasn't fortunate, anyway, the series of events have been proven the statement of Bokros since March 12: almost all institutions seem to be premature here (z: in Hungary). The welfare state is premature, the constitutional (lit: legal) state is premature.

Without weeping for any kind of past, here are a few examples: József Torgyán (z: the leader of a conservative party) presents the list of signatures related to the referendum to the President (z: Speaker) of the House. Are there

Word	Count
kolláth	1
komárom	1
kovács	3
kozák	2
krajczár	1
krausz	1
kálmán	1
károly	4
kéri	2
kónya	1
könczöl	1
köszegvári	1
kövesi	1
l	1
laczkó	1
language	2
law	3
lawyer	2
leader	2
left	1
leftist	1
legal	1
legend	1

Here, several windows are opened. At the top left, we see how the data are structured. At the bottom, the themes used are listed. As can be seen, these themes might refer to words or phrases that are literally present in the text (manifest), but also to interpretations by the coder (latent). The window to the right of this window shows all sentences that have been coded. Further to the right is an overview of how the texts are structured (in this case: document, paragraph, sentences and attributes to indicate characteristics of the text, like sequence number, date, title, etc.). The texts follow, here coded parts are indicated by shading. Finally, at the very right is a word count. When the mouse is moved over a coded text (either word or sentence) the window right at the top shows how it has been coded. Specific words in the text can be looked for in a ‘key word in context’ (KWIC) window, here search entries can be indicated and coded at once. When coding has finished a data matrix is generated that can be entered into software for statistical analysis. Today, software for qualitative research also produces a data matrix (see Silver and Bulloch, this volume). The main difference to software for qualitative research is that the goal is different: a data matrix is generated using known themes. Qualitative researchers look for themes and relations between these themes in the context of theory development or a detailed description in a case study. Themes, but also units, can therefore be defined on several levels (e.g. referential versus factual code) and there are often facilities for memoing and transcribing texts.

Coding in this way allows exploration of textual representation. A lot of ambiguity can be overcome and it is easier to code the latent meaning of what is expressed. This is possible because the context is taken into account; however, it is necessary that the motivation for choices is explained (Popping and Roberts, 2009). This does not mean that human coding is flawless. Mikhaylov *et al.* (2012) reported dramatic figures for the reliability of the human coding in the Comparative Manifestos Project. Leites *et al.* (1951) investigate speeches on the occasion of Stalin's seventieth birthday with different Soviet Politburo members. Here choices for assigning specific codes are motivated in great detail. This helps very much in understanding the coding process. Therefore it is good for the validity of the study and probably would have had a positive effect on the reliability if it were computed.

Ambiguity

A big problem with the use of sentences is that they contain ambiguity, i.e. there is doubt or uncertainty about meaning or intention. Several types of ambiguity can be distinguished. A good solution is currently unavailable and the types can be recognised when manual coding is performed. Lexical ambiguity refers to the situation in which a word or a phrase is a homonym; it has more than one meaning in the language to which it belongs. For example, ‘bank', as was mentioned before, can refer to a bench, but also to a financial institution and to the edge of a river. The context in which an ambiguous word is used often makes it evident which meaning is intended. Sometimes the correct reading can be found automatically. Word sense disambiguation is an algorithmic method that automatically associates the appropriate meaning with a word in context (Navigli, 2009). The method is instrumental and based on probabilities, and the information that is used comes from an explicit lexicon or knowledge base or it is gained by training on some corpus. Today, lexical ambiguity in text can be resolved with a reasonable degree of accuracy when this method is used. Schrodt (2012) reports the implementation of these algorithms in software. Where the correct meaning is to be indicated in the text and one does not want to rephrase the word, it is often possible to replace the word with a synonym that has only one meaning. Another solution is to look for homonyms in the texts and to replace them by synonyms that are not ambiguous or by alternative words, for example ‘bank#1’ to indicate the financial institution.

Syntactic ambiguity arises when a complex phrase or a sentence can be parsed in more than one way. The phrase ‘The Dutch sociology teacher’ leaves open whether the sociology teacher is a Dutchman or whether the teacher teaches Dutch sociology.

Semantic ambiguity arises when a word or theme has an inherently diffuse meaning based on widespread or informal usage. Here one might distinguish idiomatic ambiguity, which characterizes expressions that lack clear meaning for those who are ‘outsiders’ to a particular social group, for example ‘You'll eat your words!'. Such expressions are found when data from blogs or Tweets are used which contain a very informal content setting and slang.

Illocutionary ambiguity characterizes statements with meanings that vary as a function of statements made prior to them in context, for example ‘Stop!’ (What I am doing, or how I am doing it?) and ‘Pete bought that software’ (an awful purchase, or just a purchase?).

Besides ambiguity, typographical errors also occur in texts. For the greater part, such errors are corrected by applying a grammar checker. Search engines that identify textual attributes through exact matching of queries and character strings cannot identify expressions with typographical errors or misspellings; this can cause omissions of potentially relevant texts. The problem is addressed by fuzzy search options that take a variety of similarities into account: typographical, phonetic and stemming similarities. A stem is the root of a word (a form which is not further analysable), together with any derivational affixes, to which inflectional affixes are added.

At this moment, no standard methods (coding rules) are available to overcome these problems. Word sense disambiguation is still under development. In my view, the other problems are usually ignored or even not recognized. An exception is Popping and Roberts (2009) who explain in detail how they made their choices with regard to types of modal auxiliary verbs and rationales as used in studies they performed, among which is Roberts, Popping and Pan (2009).

Semantic Approach to Content Analysis

Semantic content analyses yield information on how themes are related according to an a priori specified semantic grammar. This type of analysis expands the types of questions that a researcher can answer. Referring to propaganda techniques in making this point, Roberts (1989: 169) notes that in a thematic analysis a possible research question would be: ‘What themes are mentioned in propaganda that are not mentioned in other communications?'. Using the semantic approach, the question can be extended to: ‘What syntactic strategies are used by political leaders when their policies fail (or succeed)?'. Unlike the former question, the latter asks about concrete relations among themes used within different social contexts.

This section introduces the semantic approach and attention is given to the way in which coding of semantic relations is performed automatically and manually. A new source of ambiguity is also introduced later.

Semantically encoding data requires that one fits themes that occur in a clause into a semantic grammar. Usually valence information (regarding negation, evaluation, etc.) is subsumed under the verb component. For this reason, one sometimes refers to a semantic grammar as having an S–V–V–O form. By taking relational characteristics of the text into account, semantic content analysis improves upon thematic content analysis methods and overcomes many of its problems. Based on a thematic content analysis, co-occurrence of subject and object can be identified. In semantic content analysis, the relation is specified and can be investigated.

Sometimes thematic text analysts will wrongly interpret co-occurrences of themes (i.e. correlations between word frequencies) as indicative of specific semantic relations among these themes. In the thematic approach, themes are counted and nothing is specified with respect to any co-occurrence. In the semantic approach, relations among themes are also encoded. This overcomes the limitations of the contingency analysis for inferences about theme occurrences.

A second point is that practitioners of contingency analysis assume that their thematic categories can be used to capture what words mean at their face value, as it were. In parallel fashion, one might be tempted to map thematic relations according to their surface grammatical relations (e.g. S–V–O). This approach can work if one's texts are highly descriptive. More generally, however, one should take into account that the intended meanings of most natural language expressions are inherently ambiguous.

The information contained in a verb can refer to four formal properties (Carley, 1993: 94 ff): directionality (one- or two-way), strength (defined as intensity, certainty, frequency, and so on), sign (positive, negative) and meaning (the content). Classes of meaning can be generated, for example:

- Similarity: indicates that one theme is identical with or looks like (a part of) another, for example 'The boy resembles his brother'.
- Causal: indicates a cause-effect relation, for example 'Car driving causes pollution'.
- Relation: indicates an association, an ordering, an evaluation, or a realization, for example 'The number of students has increased'.
- Classification: indicates a genus–species relation, for example 'A bike is a vehicle'.
- Structure: indicates a part–whole relation, for example 'The roof is a part of the house'.
- Affective: indicates a judgment of the subject about the object, for example 'Bill has a bad relationship with his boss'.

Machine Coding

Software that automatically codes clauses in text uses a parser, a tool that analyses text according to the rules of a formal grammar. It is a method of understanding the exact meaning of a sentence. It usually emphasizes the importance of grammatical divisions such as subject and predicate. The parser allows coding of simple S–V–O statements. For each of the three parts dictionaries are available, and therefore the texts will be coded instrumentally.

In this way, Gottschalk was able to measure psychological states such as hostility, depression and hope, according to his perspective on states that are reflected in how people relate words (Gottschalk and Bechtel, 1989).

Schrodt (2012), using highly descriptive simple texts (lead sentences in news service articles on international conflict), was able by using his KEDS (now Tabari) software to automatically analyse event data. His data are on the level of the clause. He found that the theme relations follow sufficiently fixed, descriptive formulae and their surface relations are nearly always unambiguous. Note, this unambiguity holds for this research project.

The development of advanced parsers is still going on. A difficulty is that the general framework of semantic analysis is language- and topic-agnostic; the actual computerized text processing is not. The parsing and coding is tailored for a domain-specific research question. Van Atteveldt, Kleinnijenhuis and Ruigrok (2008) contains a good overview of the state of the art with regard to developments in parsing as it is today. Parsers are able now to recognize semantic roles.

Manual Coding

The coder, when applying a generic semantic grammar to relatively unstructured texts, is not supposed to identify surface grammatical relations of themes, but rather to identify each theme's role within the functional form(s) appropriate to its clause of origin. Such identifications can only be made after selecting the appropriate functional form. This requires the coder to look beyond the clause. The coder has to understand both the source's intentions and the social context within which the clause appeared. Coding in a semantic text analysis based on this grammar takes a representative approach to texts. Coding clauses might become very complex. This becomes visible when the distinction between main and subordinate clauses is made. Subordinate clauses are ordinarily those related to a main clause by conjunctions (‘because', ‘since', ‘when'), relative pronouns (‘which', ‘who', ‘that') or proxies. Proxy clauses replace either the subject or the object of a clause.

The TCA software mentioned earlier contains a template for graphical mapping. Here it is possible to define an S–V–O structure and to code the parts separately. A window at the right-upper corner and below the window containing the actual codes shows the coding according to the structure defined when the cursor is moved over the coded text. Depending on the research question the source of the text and the audience might also be mentioned.

Other software that allows the coding of S–V–O tuples is PC-ACE (Franzosi *et al.*, 2013), which allows users to code the data in such a way that they can be entered into a relational database package. The database consists of a set of relations, each of which contains one or more attributes. The coding task to be performed in this software is too complex to have it performed automatically – software for qualitative analysis cannot handle this complexity.

Most programs for qualitative analysis only use hierarchical relations (themes are split into subthemes), but Atlas.ti uses horizontal relations that can be compared to a verb.

Ambiguity

A generic semantic grammar to facilitate coders’ disambiguation of illocutionary ambiguities in natural language was developed by Roberts (1989), who distinguished a four-fold semantic grammar that enables the unambiguous encoding of clauses. The four forms are:

- The description of a state of the art
- The description of a process
- The evaluation of a state of the art
- The evaluation of a process

As soon as a clause is translated into one of these forms, coding can be performed in a correct way. When machine coding is applied the possible ambiguity due to the structure of the sentence is not detected. This problem has not yet been solved, although today artificial intelligence and parsers are being used to assist in performing a correct text analysis.

A form of ambiguity that occurs very often is found in sentences that are in passive voice. Passive voice is used when the focus is on the action. It is not important or not known, however, who or what is performing the action. An example is: ‘My bike was stolen', which can be rewritten as ‘Someone stole my bike'. The sentence must be formulated in active form. Now one must note that the subject of the passive voice becomes the object of the active sentence and that the form of the verb is changed from *to be* + past particle to finite form. The types of ambiguity as mentioned before will continue to exist. A detailed example is found in Roberts (1997).

Modality Analysis

Opinion statements appearing in newspaper editorials or in speeches are very interesting because they are about the need or desirability of some action. This need or desirability becomes visible in the use of modal auxiliary verbs. These are verbs that are usually used with the infinitive form of another verb to express possibility, inevitability, impossibility or contingency. In each modal usage there are two verbs associated with the subject, namely a modal auxiliary verb (e.g. want, hope, ought, refuse) and a main verb in infinitive form (an action). These usages are not intended to convey facts or to describe events; they are used to communicate something about the likelihood of the S–V–O relation. A semantic content analysis that investigates such uses of modal auxiliary verbs is called a ‘modality analysis’ (Roberts *et al.*, 2010). The semantic grammar has the Subject–Modal-auxiliary-verb–Verb–Object (or S–M–V–O) form. This can be presented by using a different terminology: Agency–Position–Action–Object:

Subject	Agency	the initiator of an activity;
Modal-auxiliary verb	Position	the position regarding the agency's activity;
Verb	Action	the activity under consideration;
Object	Object	the target of this activity.

Positions are only taken by intentional agents; the agent cannot be any arbitrary subject. It can only be a person (or institution represented by persons) and not a ‘thing'. The position commonly involves the use of a modal auxiliary verb. For example, an editorial that states ‘Our politicians ought to cooperate more’ would be encoded as politicians (agent) ought (position) to cooperate (action) with politicians (object).

A modal clause is recognizable whenever it conveys intentionality in a way that can be transformed (in a manner agreeable to a native speaker) to a form that includes a modal auxiliary verb. The coder's challenge now is to capture how a text's author understands others’ motivations (thereby getting into the mind of someone who is getting into someone else's mind, as it were). Because modal auxiliary verbs convey intentionality, they can be used to learn about people's motivations, their ideas about a future society and thus about their ideological shifts as individuals or groups, e.g. political parties. They can also be used to learn about motivations that exist with respect to certain specific persons or groups.

The semantic grammar used in a modality analysis always has two parts at its core. There is a modal form, indicating possibility, impossibility, inevitability or contingency, and an associated rationale. These modal forms can be used to understand human motives during interactions and to distinguish subtle nuances in discourse. Many examples are presented in Popping and Roberts (2009).

Through modal usage, a text's source (i.e. its author or speaker) socially constructs what constitutes the possible, the impossible, the inevitable and the contingent regarding the agent–action–object relation. Moreover, it is always reasonable for the source of a modal clause to be queried as to the rationale or explanation of ‘why’ the agent is able, required, permitted, etc. regarding the clause's predicate. For example, a politician might follow his statement ‘We had to impose austere economic measures’ with the rationale ‘otherwise our economy would have stagnated'.

The TCA software has a window that shows the graphical mapping: ‘There is a [political, economic, cultural, security] reason why something is [possible, impossible, inevitable, contingent] for a Hungarian.’ This mapping is used in Roberts, Popping and Pan (2009: 512). Their study on Hungarian society is based on the premise that social systems are justified via the discursive use of modal statements and their associated rationales. Within authoritarian states such modal discourse usually reflects a relatively coherent ‘modality of permission'; however, when the citizens unite to overthrow the totalitarian leaders, their activities are typically justified in terms of a ‘modality of achievement’ (based on market justice among competitors) versus a ‘modality of necessity’ (based on social justice for the masses). These three discursive modalities can be differentiated using content analysis. An analysis of editorials during Hungary's first years of post-Soviet democratization suggests that as late as 1997 Hungarian political discourse was heading toward a modality of necessity, more like the predominant political modality in Western Europe than the achievement modality that characterizes political discourse in the U.S.

Network Approach to Content Analysis

Network content analysis originated with the observation that after one has encoded semantic relations among themes, one can proceed to construct networks of semantically related themes. When a theme represents a person, for example, one can now investigate the position of that person in the network by applying statistical indices for networks. More generally, when themes are depicted as networks, one is afforded more information than the frequency at which specific themes are related in each block of text, and can characterize themes and/or linkages according to their position within the network. A relation between themes might refer to various properties, as was indicated before. By using scores on these properties the information in networks can be represented. This constitutes the representation of the model. The data can be analysed statistically. Attention is first given to machine coding, and then to manual coding.

Machine Coding

A type of study that is currently receiving more and more attention is the one in which detailed sociocultural ethnographies are conducted based on characteristic descriptions from texts and fusing the results from varied sources. Tambayong and Carley (2013) focus on changes in political networks in Sudan. They were interested in themes that were aliases of political agents. By allowing their software, the AutoMap program, to filter out these agents and the relations between them, they were able to construct a network based on these agents and by using network statistics, they could indicate the relevance of each agent, even from different perspectives. This type of study is increasing, especially based on data from Tweets and blogs. The impression is that themes that can be related are sufficient for the investigators who perform these studies. Questions about the research problem and the design of the study as addressed earlier are often ignored. The same holds for the sample or even population that is investigated.

Manual Coding

For many years, two network methods that allow statistical inferences received most attention: network evaluation approaches and cognitive mapping. These methods start from different positions. In both methods the representational approach is followed. The network evaluation approach has its roots in evaluative assertion analysis (Osgood *et al.*, 1956), which starts from the position that every language has three kinds of words:

- Common meaning terms: words that have a common evaluation among ‘reasonably sophisticated users of the language'. The common meaning of words such as ‘peace’ is always positive; whereas that of words like ‘enemy’ is always negative in connotation.
- Attitude objects: these have no fixed evaluative meaning. A word like ‘car’ is likely to be evaluated differently by different people.
- Verbal connectors: words that indicate the association (‘it is…') or dissociation (‘it is not…') of attitude objects with common meaning terms or with other attitude objects.

By investigating how attitude objects are associated or dissociated, one can investigate how these attitude objects are valued in a text. For this it is necessary to parse texts into clauses, in which the three word-types can be found.

The network evaluation approach has been used in particular to investigate how newspapers report on issues in which governments are involved (Kleinnijenhuis *et al.*, 1997). In order to do so, two specific themes are needed. The user can encode a statement as a positive (is good) or negative (is bad) evaluation of a theme by relating it to the abstract theme ‘Ideal'. The statement ‘the man is friendly’ is reformulated into ‘the man has a good relationship with the Ideal (of the statement's source)'. By connecting a theme to the theme ‘Real', the user can encode a statement as an affirmation that a theme's referent exists (is) or does not exist (is not). The statement ‘unrest is rampant’ is changed to ‘Reality shows a high level of unrest'.

Cognitive mapping involves extracting relations from texts and then representing the ‘mental models’ or ‘cognitive maps’ that individual sources had in their memory at the time the relations were expressed. Within a cognitive map, the meaning of a theme is the aggregate set of relations it has to all other themes that make up a conceptual network. Mental models are dynamic structures that are constructed and expanded as individuals make inferences and gather information. They contain specific information about particular items and also general (or social) knowledge. A transcript of an individual's speech is a reflection of the individual's mental model at a particular point in time. Accordingly, such texts may be thought of as a sampling of information from the individual's memory.

The map comparison method (Carley, 1993) affords not only graphic descriptions of individuals’ mental models, but also comparisons among models maintained by various social groups. Carley (1994) showed how the method is used in four different fields that are related to culture and how it differs from thematic content analysis. In one of the studies she portrays the development over time of the theme ‘Robot', as used in science fiction.

A new field of research in which mapping is used is presented in Popping and Wittek (2015). They look at negotiations. The position that parties take in negotiations can be represented as a cognitive map of a game theoretic model. The authors explained the voting behaviour in the Dutch parliament over a certain time with respect to motions. They could explain 60 per cent of this behaviour; one third of this amount was due to the positions taken by parliament and government during negotiations.

Manual Versus Machine Coding

Trade-offs between manual and machine coding are often presented in methods literature. A number of attributes are nearly always mentioned: manual coding is slow, and therefore only used for small data sets and it does not use dictionaries; native coders can code complex sentence structures and can interpret all kinds of ambiguous texts; the coding is not replicable and is expensive as coders and trainers have to be paid; machine coding is fast and suited for large data sets; it is possible to modify dictionaries; simple sentence structures are to be coded, containing literal (manifest), present-time text; and as soon as the dictionary has been developed, there are hardly any costs.

The question of what is necessary for the purpose of your study is hardly posed. In other words, sometimes new technological affordances threaten accepted methodological standards.

Some Other Remarks

A lot of software for quantitative text analysis has been developed by researchers themselves and is usually written in the context of a specific study. This generally implies the software is not for general use, but if others want to use it, this is fine. Documentation is usually poor and the software is as it comes. This implies there has not always been a complete control for imaginable bugs, and errors are not captured. In software for qualitative research, this is generally taken good care of because private companies are responsible for the software. Popping (2015) formulated a number of questions a user should ask before choosing the software that will be used. He also referred to an often suggested disadvantage of manual coding – that code assignments are not reliable. Every researcher has to learn that coders need training, at least part of the data should be coded twice, and intercoder reliability has to be computed. The quality of manually coded data is often higher than that of machine coded data, certainly when complex texts are used.

Texts on which a study will be based are found more and more on the Internet (direct or via organizations like LexisNexis), and eventually as a blog or Tweet. These texts will be downloaded and formatted in such a way that they can be entered into software for text analysis. This all is a question of text mining (Lee *et al.*, 2010).

Prospective users should question whether the software used has a facility to weed out false positives – i.e. it must be possible to do away with texts that are selected (based on the keywords) but do not fulfil the requirements for inclusion in the dataset to be used. An estimation of possible false negatives is also needed – i.e. texts that are not selected but that should have been selected. The data should actually constitute the population of texts that can be used. From this population a representative sample can be drawn.

Computer Programs

Table 19.1 shows an overview of a number of computer programs that are available today. Most programs to be used when instrumental coding applies demand a dictionary. In the table, programs for machine learning are not listed. This overview is not exhaustive.

Table 19.1Some computer programs for content analysis and their URL

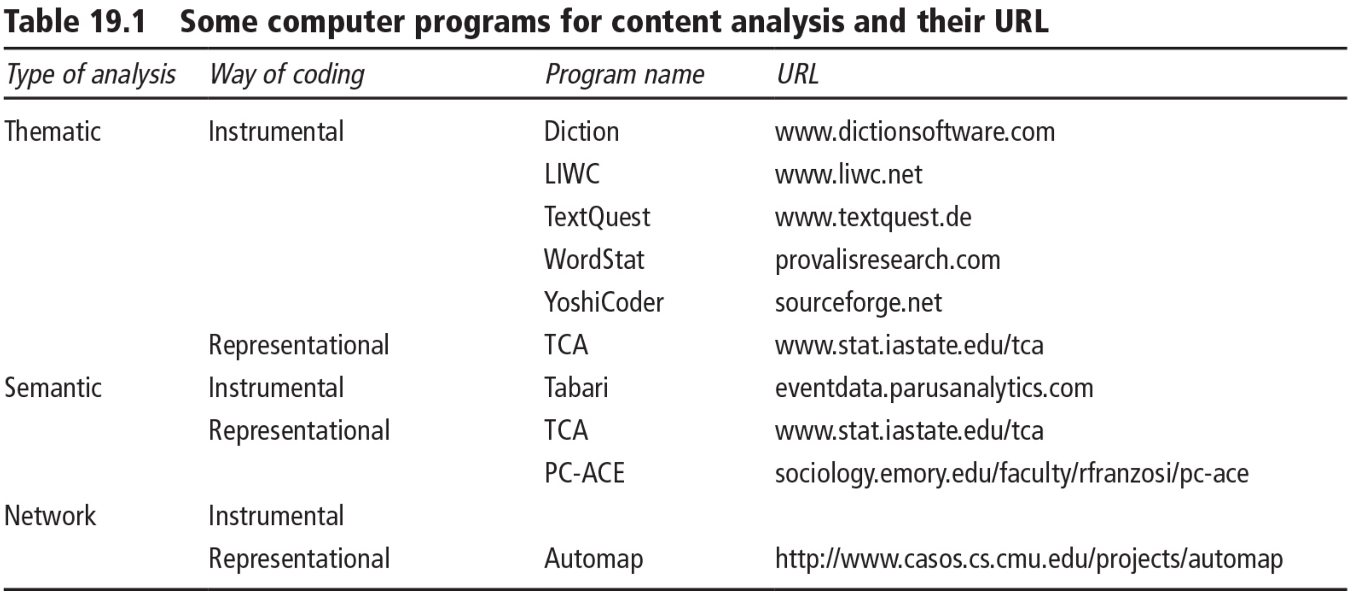


Table 19.1 Some computer programs for content analysis and their URL

Type of analysis	*Way of coding*	*Program name*	*URL*
Thematic	Instrumental	Diction	www.dictionsoftware.com
		LIWC	www.liwc.net
		TextQuest	www.textquest.de
		WordStat	provalisresearch.com
		YoshiCoder	sourceforge.net
	Representational	TCA	www.stat.iastate.edu/tca
Semantic	Instrumental	Tabari	eventdata.parusanalytics.com
	Representational	TCA	www.stat.iastate.edu/tca
		PC-ACE	sociology.emory.edu/faculty/rfranzosi/pc-ace
Network	Instrumental		
	Representational	Automap	http://www.casos.cs.cmu.edu/projects/automap

Well-known programs for qualitative research include Atlas.ti, NVivo, MaxQDA, QDA Miner.

Conclusion

Recent developments in the field of quantitative content analysis have been sketched in broad terms. A lot of attention has been given to problems that seem to be overlooked. On the one hand, language is very complex and ambiguous; this should be taken into account. Coder training and explanation of choices is a must, as is the software for managing the coding process. On the other hand, a good research question makes demands. Software can perform analyses on enormous amounts of texts in a very short time. This might be helpful, but it is not the criterion for good research.

References

Berelson, B. (1952). *Content Analysis in Communication Research*. New York, NY: Free Press.

Blei, D.M., **Ng, A.Y.**, **Jordan, M.I.** and **Lafferty, J.** (2003). ‘Latent Dirichlet allocation', *Journal of Machine Learning Research*, 3 (4–5): 993–1022.

Carley, K. (1993). *‘Coding choices for textual analysis: a comparison of content analysis and map analysis'*. In: **P.V. Marsden** (ed.), *Sociological Methodology*. Cambridge, MA: Basil Blackwell, pp. 75–126.

Carley, K. (1994). ‘Extracting culture through textual analysis', *Poetics*, 22 (4): 291–312.

Franzosi, R., **Doyle, S.**, **McClelland, L.E.**, **Putnam Rankin, C.** and **Vicari, S.** (2013). ‘Quantitative narrative analysis software options compared: PC-ACE and CAQDAS (ATLAS.ti, MAXqda, and NVivo)', *Quality & Quantity*, 47 (6): 3219–47.

Geminas, K. (2013). ‘What to do (and not to do) with the Comparative Manifestos Project data', *Political Studies*, 61 (1): 3–23.

Gottschalk, L.A. and **Bechtel, R.** (1989). ‘Artificial intelligence and the computerization of the content analysis of natural language', *Artificial Intelligence in Medicine*, 1 (1): 131–7.

Grimmer, J. and **Stewart, B.M.** (2013). ‘Text as data: the promise and pitfalls of automatic content analysis methods for political texts', *Political Analysis*, 21 (3): 267–97.

Kleinnijenhuis, J., **De Ridder, J.A.** and **Rietberg, E.M.** (1997). *‘Reasoning in economic discourse: an application of the network approach to the Dutch press'*. In **C.W. Roberts** (ed.), *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts*. Mahwah, NJ: Lawrence Erlbaum Associates, pp. 191–207.

Laver, M., **Benoit, K.** and **Garry, J.** (2003). ‘Extracting policy positions from political texts using words as data', *American Political Science Review*, 97 (2): 311–31.

Lee, S., **Song, J.** and **Kim, Y.** (2010). ‘An empirical comparison of four text mining methods', *Journal of Computer Information Systems*, 51 (1): 1–10.

Leites, N., **Bernaut, E.** and **Garthoff, R.L.** (1951). ‘Politburo images of Stalin', *World Politics*, 3 (3): 317–39.

Mikhaylov, S., **Laver, M.** and **Benoit. K.** (2012). ‘Coder reliability and misclassification in the human coding of party manifestos', *Political Analysis*, 20 (1): 78–91.

Miller, M.M. (1997). ‘Frame mapping and analysis of news coverage of contentious issues', *Social Science Computer Review*, 15 (4): 367–78.

Namenwirth, J.Z. (1969). ‘Marks of distinction: an analysis of British mass and prestige newspaper editorials', *American Journal of Sociology*, 74 (4): 343–60.

Navigli, R. (2009). ‘Word sense disambiguation: a survey', ACM *Computing Surveys*, 41 (2): 1–69.

Osgood, C.E., Saporta, S. and **Nunnally, J.C.** (1956). ‘Evaluative assertion analysis', *Litera,* 3: 47–102.

Pennebaker, J.W., **Mehl, M.R.** and **Niederhoffer, K.G.** (2003). ‘Psychological aspects of natural language use: our words, our selves', *Annual Review of Psychology*, 54: 547–77.

Pennings, P. and **Keman, H.** (2002). ‘Towards a new methodology of estimating party policy positions', *Quality & Quantity*, 36 (1): 55–79.

Popping, R. (2015). ‘Analyzing open-ended questions by means of text analysis procedures', *Bulletin de Méthodologie Sociologique*, 128: 23–39.

Popping, R. and **Roberts, C.W.** (2009). ‘Coding issues in semantic text analysis', *Field Methods*, 21 (3): 244–64.

Popping, R. and **Wittek, R.** (2015). ‘Success and failure of parliamentary motions. A social dilemma approach', *PLoS ONE*, 10 (8): e0133510.

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

Roberts, C.W. (1997). ‘A generic semantic grammar for quantitative text analysis: applications to East and West Berlin radio news content from 1979', *Sociological Methodology*, 27: 89–129.

Roberts, C.W. (2008). *‘The' Fifth Modality: On Languages that Shape our Motivations and Cultures*. Leiden, the Netherlands: Brill.

Roberts, C.W., **Popping, R.** and **Pan, Y.** (2009). ‘Modalities of democratic transformation: forms of public discourse within Hungary's largest newspaper, 1990–1997', *International Sociology*, 24 (4): 498–525.

Roberts, C.W., **Zuell, C.**, **Landmann, J.** and **Wang, Y.** (2010). ‘Modality analysis: a semantic grammar for imputations of intentionality in texts', *Quality & Quantity,* 44 (2): 239–57.

Schonhardt-Bailey, C. (2005). ‘Measuring ideas more effectively: an analysis of Bush and Kerry's national security speeches', *PS: Political Science and Politics*, 38 (4): 701–11.

Schrodt, P.A. (2012). ‘Precedents, progress and prospects in political event data', *International Interactions*, 38 (4): 546–69.

Shapiro, G. (1997). *‘The future of coders: human judgments in a world of sophisticated software'*. In **C.W. Roberts** (ed.), *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts*. Mahwah, NJ: Lawrence Erlbaum, pp. 225–38.

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

Tambayong, L. and **Carley, K.M.** (2013). ‘Network text analysis in computer-intensive rapid ethnography retrieval: an example from political networks of Sudan', *Journal of Social Structure*, 13 (2): 1–24.

Van Atteveldt, W., **Kleinnijenhuis, J.** and **Ruigrok, N.S.** (2008). ‘Parsing semantic networks, and political authority using syntactic analysis to extract semantic relations from Dutch newspaper articles', *Political Analysis*, 16 (4): 428–46.

http://dx.doi.org/10.4135/9781473957992.n19