SSOAR

Open Access Repository
www.ssoar.info

Identifying events using computer-assisted text analysis

Landmann, Juliane; Züll, Cornelia

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:
GESIS - Leibniz-Institut für Sozialwissenschaften

Dieser Beitrag ist mit Zustimmung des Rechteinhabers aufgrund einer (DFG geförderten) Allianz- bzw. Nationallizenz frei zugänglich. / This publication is with permission of the rights owner freely accessible due to an Alliance licence and a national licence (funded by the DFG, German Research Foundation) respectively.

Empfohlene Zitierung / Suggested Citation:
Landmann, J., & Züll, C. (2008). Identifying events using computer-assisted text analysis. *Social Science Computer Review*, *26*(4), 483-497. https://doi.org/10.1177/0894439307313703

Nutzungsbedingungen:
Dieser Text wird unter einer Deposit-Lizenz (Keine Weiterverbreitung - keine Bearbeitung) zur Verfügung gestellt. Gewährt wird ein nicht exklusives, nicht übertragbares, persönliches und beschränktes Recht auf Nutzung dieses Dokuments. Dieses Dokument ist ausschließlich für den persönlichen, nicht-kommerziellen Gebrauch bestimmt. Auf sämtlichen Kopien dieses Dokuments müssen alle Urheberrechtshinweise und sonstigen Hinweise auf gesetzlichen Schutz beibehalten werden. Sie dürfen dieses Dokument nicht in irgendeiner Weise abändern, noch dürfen Sie dieses Dokument für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, aufführen, vertreiben oder anderweitig nutzen.
Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

Terms of use:
This document is made available under Deposit Licence (No Redistribution - no modifications). We grant a non-exclusive, non-transferable, individual and limited right to using this document. This document is solely intended for your personal, non-commercial use. All of the copies of this documents must retain all copyright information and other information regarding legal protection. You are not allowed to alter this document in any way, to copy it for public or commercial purposes, to exhibit the document in public, to perform, distribute or otherwise use the document in public.
By using this particular document, you accept the above-stated conditions of use.

gesis Leibniz-Institut für Sozialwissenschaften

Mitglied der Leibniz-Gemeinschaft

Diese Version ist zitierbar unter / This version is citable under:
https://nbn-resolving.org/urn:nbn:de:0168-ssoar-443686

Social Science Computer Review
Volume 26 Number 4
November 2008 483-497
© 2008 Sage Publications
10.1177/0894439307313703
http://ssc.sagepub.com
hosted at
http://online.sagepub.com

Identifying Events Using Computer-Assisted Text Analysis

Juliane Landmann
Cornelia Zuell
ZUMA-GESIS, Mannheim, Germany

Events such as elections, significant changes in laws, but also extreme weather conditions, may affect societal values and, consequently, public opinion. Accordingly, a central assumption for public opinion surveys is that respondents' behavior is influenced by significant events. It is therefore necessary to consider the impact of potential events when designing a survey and, whenever possible, to control for these. To support the documentation of such societal events, the authors have developed a procedure to identify events using computer-assisted text analysis. Event words are selected and grouped by means of exploratory factor analysis based on a comparison of a large text corpus that forms the reference for a smaller text corpus consisting of media items on significant events. As a result, the factors represent significant events during a specific time period.

Keywords: *computer-assisted text analysis; statistical association approach; reference text corpus; event reporting; newspaper articles*

Events in Survey Research

Solid knowledge about events such as elections, significant changes in laws, demonstrations, extreme weather conditions, and so forth is very important in all scientific fields that rely on survey-generated data. Theoretically, all such events potentially affect a society and, consequently, may in some way influence the attitudes of its population.

Despite increased interest in event data, there is no single universally accepted definition of an event. Mostly, events are defined as activities by international actors. For example, Gerner, Schrodt, Francisco, and Weddle (1994, p. 95) define an event as

> an interaction, associated with a specific point in time that can be described in a natural language sentence that has as its subject and object an element of a set of actors and as its verb an element of a set of actions, the contents of which are transitive verbs.

Such definitions of an event are too restrictive for our purposes because we are looking for all occurrences that can influence public opinion and therefore respondent behavior (political activities as well as natural catastrophes). Therefore, we define events as all significant or major occurrences reported in mass media that are significant, interesting, exciting, or unusual. Such events can be political or economic activities as well as catastrophes (natural catastrophes, disasters, or accidents). To give some examples, political and economic

Authors' Note: Please address correspondence to Juliane Landmann at `j.landmann@web.de` or Cornelia Zuell at `cornelia.zuell@gesis.org`

events are events such as international conflicts, elections, discussions on unemployment, bank scandals, insolvencies, or strikes. Catastrophes are, for example, floods, earthquakes, train crashes, and so forth. To qualify as a major event, the following must apply: The news is reported for at least several days in mass media and gives rise to widespread public discussion and/or to a substantive increase in media use. In addition, major events attract wide and long-lasting attention.

As already indicated, the theoretical assumption at the individual level is that response behavior in survey questionnaires is influenced by such significant events. So for example if respondents give socially desirable or politically correct answers (Kaase, 1999, p. 32) to questions affected by an event, this can affect the quality of survey results and result in data bias. Every user of survey-generated data should have access to information about events that occurred during or just prior to data collection so that he or she is able to properly handle the survey-generated data. Considering the commonly cited typology of sources of survey errors, that is, sampling error, coverage error, nonresponse error, and measurement error (Braun, 2003; Groves, 1989), by focusing on critical events during data collection our approach can be classified as a contribution to reducing measurement error (Braun, 2003, p. 141). The media-reported event overview used in the European Social Survey (ESS) 2002/2003 is a promising approach in which information about events that occurred during data collection in the different ESS countries are relatively systematically collected and made available to the general public. Our endeavor is therefore based on this approach.

The ESS (`http://www.europeansocialsurvey.org/`) is designed as a cross-national survey in which 22 countries took part in Round 1 in 2002/2003. The general aim of the ESS is to develop and conduct a systematic study of changing values, attitudes, attributes, and behavior patterns within European politics. Designed to feed into key European policy debates, the ESS intends to measure and explain how people's social values, cultural norms, and behavior patterns are distributed, the way in which they differ within and between nations, and the direction and speed with which they change. Data collection is scheduled to take place every 2 years as face-to-face interviews lasting approximately 1 hr. The first data collection phase took place in September 2002. Accordingly, the first data collection phase was supplemented with an overview of media-reported events that occurred during the fieldwork. To assemble an ESS event database, consisting of a collection of identified events in the ESS countries, each national coordinator was asked to provide major national events that could generally influence the answers to questions or respondent behavior in the ESS survey. To help them classify events as major events in a somewhat comparable manner, the national coordinators were provided with a not-very-detailed classification scheme specifying the kinds of events to be reported (Stoop, 2002a, p. 2ff), for example, international conflicts, elections and plebiscite, resignation of politically significant persons, fall of cabinet, disasters such as crashing financial markets, explosions, or extreme weather conditions.

Each national coordinator was asked to specify a title for the event they entered in the database, such as "new immigration law," an appropriate keyword, names of coverage sources, headlines, and the approximate time frame, and to deliver a short event description and an assessment of scope or impact. Furthermore, they were required to give additional information on how this event could affect the fieldwork of the ESS (Stoop, 2002b, p. 2).

The resulting event collection is available on the World Wide Web (`http://www.europeansocialsurvey.org/`). These reports are heterogeneous in two ways:

First, the definition of a major event seems to differ significantly between national coordinators, and second, the assignment of events into a priori categories is also inconsistent.

To illustrate this, we use all available reports for January 2003 about the first publicly noticed tensions between the United States and Iraq, which were then followed by a war against Saddam Hussein's Iraq in March and April 2003. Reports of 12 countries are available for the selected month. Eight out of 12 national coordinators identified these first tensions between the United States and Iraq as a major event as defined above. The remaining four national coordinators did not identify these tensions as a major event at all.

In the same month, the differences in deciding what a major event is can explain the striking numerical discrepancy concerning reported events between Switzerland and the Netherlands. The Swiss national coordinator noted, "In the month of January, no major event which could affect the fieldwork can be reported" (`http://www.scp.nl/users/stoop/ess_events/ch01.htm`), whereas the Dutch national coordinator identified a total of 29 events that could affect the fieldwork (`http://www.scp.nl/users/stoop/ess_events/nl01.htm`).

The assignment of events into the a priori defined category system is not quite convincing either, because the national coordinators did not interpret the categories in the same way. For example, in the January 2003 reports of the first publicly noticed tensions between the United States and Iraq, only two national coordinators used the category system in the predefined way and labeled this major event as "international conflict." All other national coordinators seem to have been imprecise in their categorization decisions.

To sum, the media-reported events during the first data collection phase of the ESS are problematic due to the inconsistent definition of major events as well as the inconsistent event assignment into predefined categories. To deliver a suitable collection of major political, social, and other potentially critical events in the ESS countries during the data collection phase, these problems need to be solved. One way of supporting the event data collection process is to implement techniques of computer-assisted text analysis.

Identifying Events Using Computer-Assisted Text Analysis

Techniques of computer-assisted text analysis, such as dictionary-based approaches, statistical association approaches (Krippendorff, 2004, p. 283ff), and techniques working with text corpus comparison, seem to be particularly appropriate to identify major events in a systematic way. A short description of the basic ideas behind these different approaches to computer-assisted text analysis follows. After that, we discuss whether and how these approaches can actually contribute to solving the problem of uniform selection and categorization of events.

The most frequently used approach in computer-assisted text analysis is the dictionary-based approach. This approach requires an a priori developed category system. The so-called dictionary contains categories that are described by lists of words with shared meanings. Typically, the categories and word lists in such a dictionary are tailored to a specific theory. For example, words such as *trouble*, *anger*, *pleasure*, or *affection* can be regarded as indicators for the category "emotion." Using the dictionary approach to code a text, a word is given a certain code if this word is included in the corresponding category

list. Examples of the application of the dictionary-based approach can be found, for instance, in Gerner et al. (1994) or in Laver and Garry (2000).

The statistical association approach is based on the dream of artificial intelligence by researchers in the 1960s who hoped to generate abstracts of written documents automatically. Klaus Krippendorff (2004) describes the basic idea as follows:

> Statistical abstracting assumed that the important words in a text are identifiable by their relative frequency; that their meanings are a function of their proximity to each other, which calls for the observance of co-occurrences; and that sentences that contain statistically prominent co-occurrences can then serve as representations of a text as a whole. (p. 289)

The dream failed to come true, but it has stimulated the development of programs that do not require a priori defined categories. The statistical association approach typically consists of a four-step process. First, word frequencies are calculated. Second, words for the following analyses are selected, and third, proximities are calculated based on the co-occurrences of words. In the last step, classification procedures such as cluster analysis or multidimensional scaling procedures are applied. The programs differ in the process of selecting words, in the calculation of the relations between words, and in the classification procedure. Examples of applications of the statistical association approach can be found in articles by Sherblom, Reinsch, and Beswick (2001), Eisner (1992), and Doerfel (1994).

Another approach in the field of computer-assisted text analysis uses a technique that works with reference-text corpora. The idea behind this approach is to identify text characteristics of a source text by comparing it to a reference-text corpus. Characteristics of the source text are determined by identifying differences in relative word frequencies between the source text and the reference-text corpus. With this kind of analysis, the reference-text corpus usually represents the typical usage of the vocabulary in texts of a specific type. Examples of applications of the reference-text techniques can be found in Laver, Benoit, and Garry (2003) and in Kabanoff, Murphy, Brown, and Conroy (2001).

At first glance, the dictionary-based approach seems to be the most convenient approach to identify events. The event identification using this approach seems to solve both problems of event identification mentioned above. Providing a well-defined dictionary whose use is mandatory could solve the problem with the nonuniform definition of major events. Because the coding process is automated, the problem regarding the nonuniform usage of the predefined categories would also be eliminated. Unfortunately, the application of this approach in a cross-national context would be enormously expensive, both in terms of time and money, because of the inevitable necessity of building a dictionary for every language spoken in the countries participating in the ESS. To illustrate the effort it takes to develop dictionaries, Philip A. Schrodt (Schrodt & Gerner, 2001, p. 2.7) describes that it took nearly 4 years to develop and evaluate a dictionary to code international events in English texts. Moreover, the success of the implementation of the dictionary-based approach is uncertain because the dictionaries have to be constructed in advance. No one knows in advance which (new) events could possibly occur during the data collection phase, and therefore dictionaries need to be updated whenever a new event or a new political actor emerges. Recognizing these problems, we focus on the remaining two approaches: the statistical

association approach and the reference-text corpora technique. Unfortunately, contrary to dictionary-based methods, neither approach implements a categorization system, and therefore events cannot be automatically coded according to predefined categories.

The idea of identifying events using the statistical association approach seems very promising because no prior knowledge about possible events is necessary. The challenge, however, is how to optimally select relevant words (Landmann & Zuell, 2004). One possibility is to choose the words with the highest frequencies as analysis words for which the co-occurrences are subsequently calculated. Because the most frequently used words in texts are typically meaningless (e.g., articles, functional words), one has to define lists of stop words that are excluded from the analysis. The definition of such stop-word lists for many different languages is, like the dictionary construction, very time-consuming. A viable alternative would be to identify the relevant words by using the reference-text technique. This text-analyzing technique allows us to select words by comparing a source text with a reference-text corpus. These words are then used in subsequent analyses.

Based on these two techniques, the reference-text approach and the statistical association approach, we developed a specialized event identification procedure. This procedure will now be described in detail. First, we explain the technique in a step-by-step fashion. Then we present details on our first effort to automatically identify events during 1 month of the data collection phase in Round 1 of the ESS.

The Procedure

The goal of the proposed procedure is the consistent identification of events occurring during a specific time period.

Using the reference-text corpus technique, we first identify the words that describe the most important events of a selected time period. We will refer to these words as "event words" because these words constitute the basis for identifying major events. Our assumptions for the selection of these event words are that

(a) major events are reported frequently in newspapers during a specific time period and can be identified by frequently used words and
(b) the words used to describe the events are distinguishable from other words because they occur more frequently in the newspaper texts of a specific time period than in a text corpus of general word usage.

Based on these assumptions, we compare a reference-text corpus composed of newspaper texts covering a longer time period with a corpus of newspaper texts from a specific time period of interest (the so-called event-text corpus). Our assumption is that the reference-text corpus represents the typical vocabulary usage in newspapers in general and the event-text corpus contains specific event words for the selected time period.

A word frequency list consisting of all words in the reference-text corpus is generated, and these frequencies are compared to the word frequencies generated from the event-text corpus. The difference between these frequencies is used to identify words that occur

Table 1
The Four Steps of the Procedure

Steps of the procedure of automatic event identification	
Step 1	Composition of the reference-text corpus that represents general language usage
Step 2	Composition of the event-text corpus: texts covering time period of interest
Step 3	a. Calculation of word frequencies and of relative frequencies for all words in both corpora
	b. Calculation of differences between the relative word frequencies of specific words in the reference-text corpus vs. event-text
	c. Selection of the *n* words with the largest differences between general language usage and the event texts
Step 4	Exploratory factor analysis based on the selected words

significantly more frequently in the event-text corpus than in the reference-text corpus. Following our assumption that these words are indicators for major events reported in the specific time period, words with the largest deviation in frequencies are used in the subsequent statistical association analysis.

Table 1 shows the steps involved in the procedure of automatic event identification. A detailed description follows below.

Step 1

First, the reference-text corpus is determined. This reference-text corpus represents the general word usage. Specifying the reference-text corpus is a crucial aspect of our analysis: If inappropriate reference texts are selected, differences in word frequencies between reference-text and event-text corpora may actually be due to this inadequate selection rather than the actual occurrence of an event. For example, if we used poetry as a reference-text corpus, all names of politicians or words with a politically oriented background would be strongly represented in our event-text corpus but not in the reference-text corpus. Although one could say something about the characteristics of the two text corpora, it would not be possible to identify events. A further requirement for the reference-text corpus is that it should include as many words as possible. The more comprehensive the word universe, the clearer the differences between the event-text corpus and its reference-text corpus, thus leading to more significant differences. We used newspaper texts to create our reference-text corpus because newspapers are the medium in which events are typically reported. The resulting text corpus represents the general word usage in newspaper texts.

Step 2

Second, we prepare the event-text corpus. It is made up of newspaper texts published during the specific time period of interest. The event-text corpus is included in the reference-text corpus by definition and to avoid division by zero in the subsequent calculation.

Step 3

Third, we calculate word frequencies and relative frequencies for all words in the reference text corpus. PR_i is the relative frequency of word i in the reference-text corpus and is defined as

$$PR_i = \frac{FR_i}{\sum_n FR},$$

where FR_i is the absolute frequency of word i in the reference-text corpus.

Afterwards, we calculate word frequencies and relative word frequencies (PE_i) for all words of our event-text corpus. Differences between the relative frequencies of the words in the reference-text corpus and the frequencies of the words in the event-text corpus are also calculated. For our purposes, it is necessary to use relative differences because for words with higher relative frequencies we expect larger differences between the two word frequency counts. D_i is the relative difference between the relative frequency of the word i in the reference-text corpus and the relative frequency of the same word in the event-text corpus. It is defined as

$$D_i = \frac{PE_i - PR_i}{PR_i}.$$

The words with the largest relative differences are those that occur much more frequently in the event-text corpus than in the general vocabulary of newspapers. Our procedure sorts words by their relative differences, and words with the largest relative differences are used as event words in the further analysis. However, it is necessary to find a cutoff point to determine how many of these large-deviation words should actually be included. If one opts for few words, the number of identifiable events is limited. In contrast, if one includes many words, one can expect to identify a larger number of events.

According to our definition of an event as significant occurrences reported in mass media over several days, we specified two more prerequisites for words to be considered as event words: The word must occur frequently in the event-text corpus and the word must appear in a specific proportion of all newspaper articles (the logic being that a salient event is covered by a large number of articles).

These two prerequisites, together with the target number of selected event words, determine the event-search process. Restrictive conditions yield few, but obviously relevant, events. Less restrictive conditions identify more, albeit potentially unimportant, events.

Step 4

We use the statistical association approach to group the words. We calculate word frequencies for all selected words in the event-text corpus. The calculation of word co-occurrences is based on these frequencies. Newspaper articles originally included in the event-text corpus that do not contain a single selected event word are excluded from further analysis. Finally, we reduce the complexity of the numerous event words via an exploratory factor analysis yielding factors describing events.

An Example

In this section, we illustrate the functionality of the procedure outlined in the previous section. Our example focuses on major events in Great Britain occurring at the beginning of the data collection phase of the ESS in August 2002.

To compose the reference-text corpus, we chose *The Guardian* as a representative of British media coverage. Actually, the decision of a specific newspaper is not really critical. We are only looking for outstanding events, and presumably these will be reported in all newspapers, as well as in television and broadcast, irrespective of the political or cultural tendency of the medium. Please note that we are not interested in *how* something is reported but in *what* is reported. We collected all articles published between November 2000 and October 2002 in *The Guardian* sections Home and Foreign as well as all articles published on the first page of each newspaper edition. The Lexis Nexis database[1] was used to obtain machine-readable articles. The resulting final reference-text corpus consists of 35,793 articles composed of 14,779,142 words that can be considered as normal vocabulary use in newspaper articles. Our event-text corpus consists of specific texts published in *The Guardian* in August 2002. It consists of 1,630 articles and 679,377 words. All words were lemmatized using the Treetagger[2] routine. Absolute and relative word frequencies were calculated for all corpus words.

After preparing the two text corpora and the word frequency lists, we calculated the difference between the relative frequency of a word in the reference-text corpus, the relative frequency of the same word in the event-text corpus, and relative differences as described above. The words with the largest differences are shown in Table 2.

Words were then sorted by their relative differences, and words with the largest relative difference (event words) were inputted in the statistical association analysis. In this example, we added two more restrictions: First, we removed all words with very low frequencies, in this case smaller than 50, based on our assumption that major events are reported frequently in newspapers and can be identified by frequently used words. Second, we included only words that were found in at least 2% of all articles in the event-text corpus. These restrictions reflect the fact that the importance and relevance of an event can be determined by the frequency of reporting. In addition, we limit the number of words selected for the factor analysis to the 30 words with the largest deviation from the reference-text corpus, given that all other conditions are met. The specific cutoff points are somewhat arbitrary, but the decision was based on our familiarity with the data and a judgment call on the number of target major events.

The following words were selected:

Soham, *Holly*, *Jessica*, *Chapman*, *Wells*, *Johannesburg*, *Cambridgeshire*, *sustainable*, *Baghdad*, *delegation*, *earth*, *disappearance*, *Iraq*, *bird*, *girl*, *league*, *invasion*, *Earth*, *grade*, *Neil*, *festival*, *crop*, *Saddam*, *Edinburgh*, *summit*, *cancer*, *Iraqi*, *college*, *exam*, and *Hussein*.

A visual inspection discerns certain events in the word list, for example, the conflict in Iraq or the death of two schoolgirls. A more comprehensive and thorough exploratory factor analysis is applied to identify the latent semantic fields of the event words and to identify reported events. The factor analysis allows us to replace many more or less correlated

Table 2
Frequencies and Relative Differences of the 30 Words Used in the Analysis

Word	FR	PR	FE	PE	D
Soham	226	.0000153	179	.0002666	16.4345
Holly	255	.0000173	176	.0002621	14.1928
Jessica	289	.0000196	172	.0002562	12.1007
Chapman	201	.0000136	88	.0001311	8.6372
Wells	314	.0000212	135	.0002011	8.4639
Johannesburg	237	.0000160	98	.0001460	8.1021
Cambridgeshire	291	.0000197	117	.0001743	7.8503
sustainable	294	.0000199	75	.0001117	4.6154
Baghdad	743	.0000503	129	.0001921	2.8218
delegation	478	.0000323	73	.0001087	2.3617
earth	978	.0000662	148	.0002204	2.3311
disappearance	391	.0000265	58	.0000864	2.2652
Iraq	4580	.0003099	678	.0010098	2.2586
bird	733	.0000496	107	.0001594	2.2132
girl	3373	.0002282	490	.0007298	2.1977
league	538	.0000364	75	.0001117	2.0686
invasion	646	.0000437	90	.0001340	2.0667
Earth	415	.0000281	57	.0000849	2.0234
grade	843	.0000570	110	.0001638	1.8723
Neil	559	.0000378	72	.0001072	1.8352
festival	862	.0000583	109	.0001623	1.7835
crop	890	.0000602	112	.0001668	1.7701
Saddam	1,760	.0001191	220	.0003277	1.7515
Edinburgh	911	.0000616	113	.0001683	1.7304
summit	2,203	.0001490	273	.0004066	1.7278
cancer	2,156	.0001459	262	.0003902	1.6750
Iraqi	1,654	.0001119	200	.0002979	1.6617
college	1,153	.0000780	134	.0001996	1.5582
exam	983	.0000665	113	.0001683	1.5304
Hussein	878	.0000594	99	.0001474	1.4820

Note: FR = frequency in reference text; PR = relative frequency in reference text; FE = frequency in event text; PE = relative frequency in event text; D = relative difference.

variables (our event words) by few independent factors without crucial information loss. A similar use of factor analysis can be found in Simon and Xenos (2004), who used factor analysis to find category definitions for coding purposes.

Based on the frequencies of the event words in each newspaper article, correlations were calculated. Subsequently, we performed an exploratory factor analysis using principle component as extracting method and varimax rotation. Given our strict rules for the selection of the event words (our variables in the factor analysis), the interpretation of the factor results is relatively straightforward.

To limit the number of dimensions, we adopt the Kaiser (1974) rule accepting all factors with Eigenvalues greater than 1 and only looked at factor loadings greater than 0.5 (cf. Table 3).

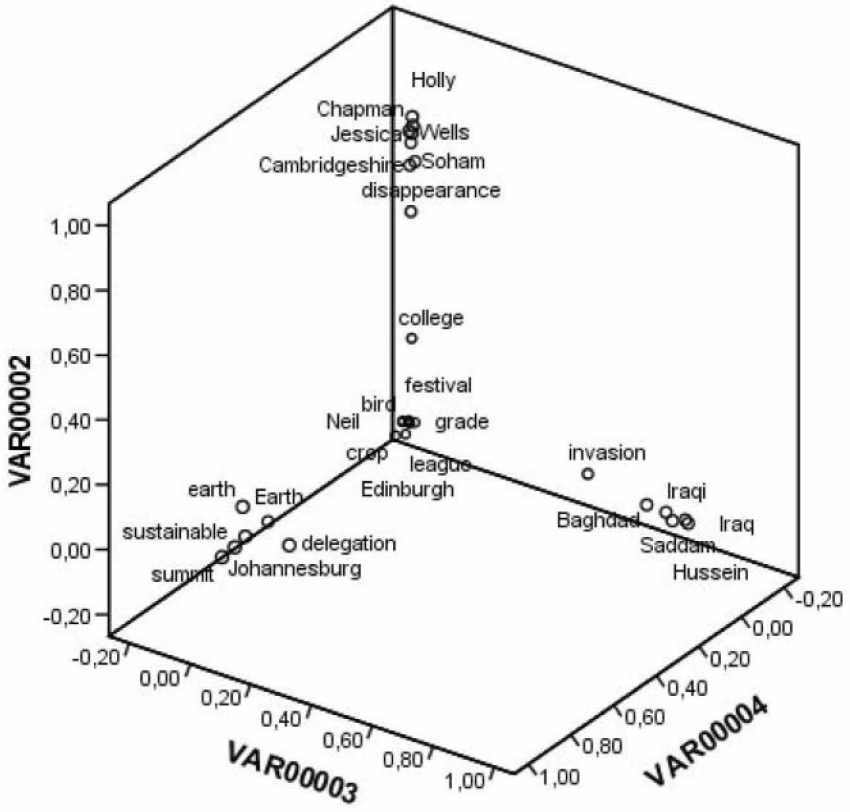
Table 3
Factors and Factor Loadings (Varimax Rotated)

Factor and Item	Loading
Kidnapping and murder of two girls	
Soham	.761
Holly	.882
Jessica	.909
Chapman	.864
Wells	.866
Cambridgeshire	.762
disappearance	.622
girl	.800
Potential war against Iraq	
Baghdad	.740
Iraq	.844
invasion	.518
Saddam	.876
Iraqi	.799
Hussein	.830
Earth Summit	
Johannesburg	.785
sustainable	.720
delegation	.704
earth	.721
Earth_Summit	.582
summit	.838
Publication of General Certificate of Secondary Education and A-level exam results	
grade	.908
exam	.907
Edinburgh International TV Festival	
festival	.825
Edinburgh	.822

Following the Kaiser criterion, the analysis results in nine different factors. The first three factors explain the most variance. The first factor comprises the words *Soham*, *Holly*, *Jessica*, *Chapman*, *Wells*, *disappearance*, *girl* and represents the event "Kidnapping and murder of two girls." Two 11-year-old girls went missing in Great Britain at the beginning of August, and after more than 2 weeks of blanket media coverage and numerous false hints, their bodies were found. A local school caretaker and his girlfriend were arrested and the caretaker charged with their murder. This story completely dominated the British media in August, and there was considerable public grief when the two bodies were found. There has since been criticism of the behavior of some members of the press and the public. A debate about the best way to ensure children's safety continues.

The second factor is characterized by the words *Baghdad*, *Iraq*, *invasion*, *Saddam*, *Iraqi*, and *Hussein*, which are indicators for the event "Potential war against Iraq." The continued discussion of an invasion in Iraq and the justification of a war are the main topics concerning this event.

Figure 1
The First Three Factors Illustrated in a Three-Dimensional Scatterplot

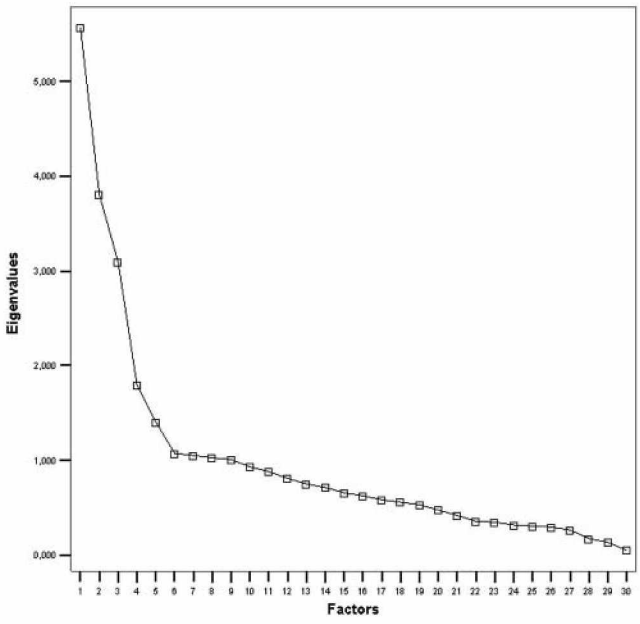


The third factor can be labeled "Earth Summit." The Earth Summit in Johannesburg had a bad start in Great Britain, with a controversy about who should attend. The government's environment minister was initially told he should not go (because of concern about the number of delegates) but was later reinstated. This was given considerable media coverage. Apart from this, media coverage of events building up to the summit focused upon the United States's relative lack of enthusiasm for the event, the ironic contrast between its extravagance and the poverty of the surrounding area, and whether the summit's size precluded any meaningful conclusions being reached. The three main factors are shown in a three-dimensional scatterplot (see Figure 1).

An inspection of the screeplot of Eigenvalues (see Figure 2) is instrumental in deciding on the number of factors to be interpreted as events. In our case, the screeplot indicates that five factors should be interpreted instead of the three described above.

Factor four is explained by the words *grade* and *exam* and describes the event "Publication of GCSE [General Certificate of Secondary Education] (taken by 16-year-old

Figure 2
Screeplot of the Eigenvalues



school pupils) and A level (taken by 18-year-olds) exam results." Annual exam results were published in the month of August. A discussion started about whether exams are getting easier (provoked by the fact that the pass rate over the past few years has increased) and about a growing gap between girls and boys in GCSE results (with girls doing markedly better), as well as about the declining number of children learning foreign languages.

Factor five (*festival* and *Edinburgh*) represents "*The Guardian* Edinburgh International TV Festival." This factor, however, can be attributed to the specific newspaper: *The Guardian* had a special interest in highlighting its own festival and arousing public interest. Therefore, this event is not a major event and should be disregarded. This somewhat qualifies our assumption that the selection of the newspaper plays no significant role for finding major events.

Conclusion

Our aim was to easily and clearly identify major events using computer-assisted text analysis. After selecting event words by means of a text corpus comparison, we used an exploratory factor analysis to identify the events. As the results presented here demonstrate, such a technique is plausible and convincing. The rules provided to identify major events

are consistent and clear. Three parameters were used to control how many events are treated as major events: the number of words selected for the analysis, the required minimum frequency of words, and the number of articles in which an event word must occur. The recommended approach has two major advantages: It is applicable without any specific knowledge of events, and it is applicable for different countries and different languages. The only special requirements are country-specific reference-text corpora and country-specific event-text corpora. In contrast to the event reporting currently used by the ESS, the proposed procedure is independent of individual or country-specific definitions of what major events are and thus offers a quasi-standardized procedure that is appropriate for comparative survey research.

One precondition of our approach is the more general definition of a major event: Major events in our understanding are reported frequently on the front page of quality daily papers and can be identified by frequently used words. We are not interested in how events are reported or who the actors are but in what is reported. Subject to these preconditions, co-occurrences of words provide good results in identifying major events. An alternative approach to identify events systematically is a textual analysis based on a categorical coding scheme and conducted by human coders. Similar to a dictionary-based approach, textual units have to be coded in abstract categories. But compared to our approach, there are two disadvantages: The first one is the necessity to develop coding schemes for all countries and languages in advance and to continuously update them whenever events occur that have not been previously anticipated. The second disadvantage is even more crucial. This type of coding is time-, labor-, and thus cost-intensive. These two drawbacks of manual coding provide substantial arguments for our automatic procedure of computer-assisted text analytical event identification.

The intention of the ESS is to provide researchers with a substantial pool of major political, social, and other potential important events that happened in ESS countries during the data collection period. We propose to first identify events by our automatic approach and then manually classify the resulting events into predefined categories. An automatic coding of these events is possible in general, but given the small number of major events, the development of dictionaries would be more costly than the manual coding.

The broad interest in agenda-setting research in communication and political science is mostly focused on voting behavior and the influence of a particular event on voting (see, e.g., Brody, 2000). In contrast, social science survey research has somewhat neglected the influence of events on respondents' behavior. Two examples for research on that topic are papers by Das, Bushman, and Bezemer (2005) and Winneg (2006). Both papers investigate the influence of one specific event (in both papers a terrorist attack) on respondents' behavior. The event database is a unique opportunity for social research to launch more systematic research in this area.

Beyond the identification of events, the approach proposed in this article provides a general framework for identifying specific topics in texts. All that is necessary is to generate adequate corpora for the research subject. The approach, for example, could be applied if researchers are interested in the themes that dominate the discussion in an organization during a specific time period, such as the themes that are discussed in political parties before and after elections.

Notes

1. The Lexis Nexis database can be accessed online (`http://www.lexis-nexis.com/`) and offers (among other things) many different newspapers as full texts.

2. Lemmatization refers to the matching of all different forms of a word regardless of whether its root is the same, for example, both *say* and *said* have the same lemma. Treetagger was developed by the Institute for Natural Language Processing, University of Stuttgart, Germany, and can be used to parse and lemmatize texts written in different languages (German, English, French, Italian, Spanish, Greek; `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger`)

References

Braun, M. (2003). Errors in comparative survey research: An overview. In J. A. Harkness, F. J. R. Van de Vijver, & P. Ph. Mohler (Eds.), *Cross-cultural survey methods* (pp. 137-142). Hoboken, NJ: John Wiley.

Brody, R. A. (2000). International crisis and public support for President Clinton. United Nations' arms inspection in Iraq. *The Public*, *3*, 55-66.

Das, E., Bushman, B., & Bezemer, M. (2005, July). *The impact of terrorist acts on Dutch society: The case of the Van Gogh murder.* Paper presented at the First Conference of European Association Survey Research, Barcelona, Spain.

Doerfel, M. L. (1994, November). *The 1992 presidential debates: A new approach to content analysis.* Paper presented at the annual meeting of the Speech Communication Association, New Orleans, LA.

Eisner, M. (1992). Semantische Assoziation und Dissoziation von politischen Leitbegriffen. Ein neues textanalytisches Verfahren zur Identifikation von semantischen Assoziationsfeldern und einige Anwendungsbeispiele [Sematic association and dissociation of political guiding terms. A new approach of text analysis to identify semantic association fields and some application examples.]. In C. Zuell & P. Ph. Mohler (Eds.), *Textanalyse. Anwendungen der computerunterstuetzten Inhaltsanalyse* (pp. 185-212). Opladen, Germany: Westdeutscher Verlag.

Gerner, D. J., Schrodt, P. A., Francisco, R. A., & Weddle, J. L. (1994). Machine coding of event data using regional and international sources. *International Studies Quarterly*, *38*, 91-119.

Groves, R. M. (1989). *Survey errors and survey costs*. New York: John Wiley.

Kaase, M. (1999). *Qualitaetskriterien der Umfrageforschung. Denkschrift* [Quality criteria of survey research]. Berlin, Germany: Akademie Verlag.

Kabanoff, B., Murphy, W., Brown, S., & Conroy, D. (2001). The DICTION of Howard and Beazley. What can computerised content analysis tell us about the language of our political leaders? *Australian Journal of Communication*, *28*, 85-103.

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrica*, *39*, 31-36.

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology.* Thousand Oaks, CA: Sage.

Landmann, J., & Zuell, C. (2004). Computerunterstuetzte Inhaltsanalyse ohne Diktionaer? Ein Praxistest [Computer-assisted content analysis without dictionary? An application test]. *ZUMA-Nachrichten*, *54*, 117-140.

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

Laver, M., & Garry, J. (2000). Estimating policy positions from political texts. *American Journal of Political Science*, *44*, 619-634.

Schrodt, P. A., & Gerner, D. J. (2001). *Analyzing international event data*. Retrieved October 10, 2007, from `http://www.ku.edu/~keds/papers.dir/AIED.C2.pdf`.

Sherblom, J. C., Reinsch, N. L., & Beswick, R. W. (2001). Intersubjective semantic meanings emergent in a work group: A neural network content analysis of voice mail. In M. D. West (Ed.), *Applications of computer content analysis* (pp. 33-50). Westport, CT: Ablex.

Simon, A. F., & Xenos, M. (2004). Dimensional reduction of word frequency data as a substitute for intersubjective content analysis. *Political Analysis*, *12*, 63-75.

Stoop, I. (2002a). *Context and event data. Guidelines for national coordinators*. Retrieved October 10, 2007, from http://www.scp.nl/users/stoop/ess_events/guidelines_events.htm

Stoop, I. (2002b). *Context and event data. Progress report*. Retrieved October 10, 2007, from http://www.scp.nl/users/stoop/ess_events/events_context_interim_report.pdf

Winneg, K. (2006, May). *Impact of new exposure on beliefs about the likelihood of terrorist attacks*. Paper presented at the annual conference of the American Association for Public Opinion Research, Montréal, Canada.

Juliane Landmann is a member of the Text Analysis Group at ZUMA-GESIS and has earned her doctorate in social science at the University of Mannheim. Her interests include computer-assisted text analysis methodology and issues of organized interests.

Cornelia Zuell is a member of the Text Analysis Group at ZUMA-GESIS. Her interests include computer-assisted text analysis methodology, text analysis software, and statistical issues.