

Reporting Societal Events to Facilitate the Interpretation of Survey Results

Cornelia Zuell & Juliane Landmann

Abstract

Societal events are events like elections, significant changes in laws, demonstrations, but also extreme weather conditions. All such events can have an effect on a society and, consequently, influence the attitudes of its population. When conducting a survey, the impact of an event must be considered and, whenever possible, controlled especially in survey projects in which different countries participate. However, manual identification of significant events that occur during the data collection phase is a very error-prone and time-consuming task. Therefore, we have developed a procedure to identify events using a combination of two different approaches of the (quantitative) computer-assisted content analysis: the reference text technique and the statistical association approach. On the basis of distinctive features of word usage in a so-called reference text corpus and in newspaper texts of a specific time period in which events should be located, words are selected and classified by means of an exploratory factor analysis. We will demonstrate this procedure by means of an example in which we will identify events automatically in Germany as well as in Great Britain. First, we will describe the composition of a reference text corpus. Thereafter, we will provide details for the calculation of relative differences between the relative frequencies for each word in the reference corpus and in the newspaper text. We will use the words with the highest relative differences as keywords for further analysis. Based on the co-occurrence of these keywords in each newspaper article, we will conduct a factor analysis to identify the events.

Keywords

Computer-assisted content analysis, event reporting, survey context variables

1. Introduction

Since the beginning of the first data collection period of the ESS, the events occurring during the data collection phase have been documented in all countries of the ESS (Stoop 2002, 2004, 2006). The assumption is that respondent behaviour or answers to some questions are influenced by significant events in various areas. Hence, the impact of an event must be considered and, whenever possible, controlled across the countries in the ESS. Each national coordinator in the ESS is asked to provide major national events that could influence the answers to questions or respondent behaviour in general. The way how events are to be identified is specified in the guidelines for national coordinators¹. In 2002, survey

coordinators were asked to send an overview of events each month. In further rounds, weekly reports have been requested.

Unfortunately, manual documentation of events is a very time-consuming and error-prone task in such cross-border projects. The ESS event data collection lacks consistency, in the sense that the definition of a major event seems to be handled quite differently in each ESS country. The number of reported events in round 3 of the ESS varies from 7 (in Norway) to 197 (in Spain) for the whole data collection period. The number of reported events varies from 1 to 15 per week (<http://www.scp.nl/ess/eventnet/>). Major events that drew front-page headlines for many days (for example, the climate change report or the execution of Saddam Hussein) were mentioned in only some of the countries.

Therefore, we will propose an approach to identify major events using computer-assisted content analysis.

The most frequently used approach of computer-assisted content analysis is the dictionary-based approach. This approach requires an a priori developed dictionary defining all possible events. The dictionary is used to code texts according to the specified categories. The coding results can be used to identify the most frequently reported and a priori categorised events.

Another approach, the Statistical Association Approach, is based on consideration of co-occurrences of words. The co-occurrence of words in a text unit defines a matrix of similarities between words and this matrix can be further analysed by classification methods.

In the following, we will discuss the applicability of the two approaches and we will propose a procedure to identify the major events reported in newspapers during a specific time period combining the statistical association approach with a reference text technique.

2. The Dictionary-based Approach

Initially, we preferred the dictionary-based approach as the most often used approach in computer-assisted content analysis. The basis for this approach is a user-defined dictionary containing the definition of categories in form of word lists. Based on our knowledge about the dictionary-based approach, we do not recommend this approach to identify events for several reasons. The following two main reasons can be outlined as:

Time-consuming Development

- The development of dictionaries is very time-consuming. Philipp Schrodtt (2001:2-7) mentions in his paper that it took nearly four years to develop and evaluate a dictionary to code international events in English texts.
- The dictionary has to be developed and validated for every language spoken in the countries participating in the ESS or all newspaper texts have to be translated to English before coding.

A Priori Development

- The dictionary has to be developed a priori which means you have to know which events can possibly occur because you have to define categories in the dictionary.
- The dictionaries have to be updated every time a new coding phase starts because, for example, new events can occur and politicians change. Word lists to define new events have to be added as soon as these events happen.

3. Combination of a reference text technique with the Statistical Association Analysis

Recognising these problems, we decided to test a second approach of computer-assisted content analysis to identify major events. Some time ago, we discussed the statistical association approach as an alternative to the dictionary-based approach (Landmann & Zuell 2004). One result of this test was that the statistical association approach offers possibilities for an explorative analysis and the enormous time-consuming text pre-processing phase can be significantly reduced by lemmatisation and parsing routines. Regarding the aim of identifying events, the crucial advantage is that one does not need a priori defined categories, which means that such an approach could be very appropriate for finding events without too much previous knowledge about the text itself.

One major problem of this kind of analysis is how to differentiate between words which are indicators for events and words which are so-called meaningless words. Our assumptions are that a) major events are reported frequently in 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 much more in the texts of a specific time period than in a larger text sample of general language usage. For our research question we use newspaper texts because newspapers are the medium in which societal events are reported typically.

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

4. The Procedure

In the following we will describe our procedure in a more detailed fashion. In general, the procedure can be described in four basic steps, starting with the composition of the reference text corpus, selecting the event text corpus, moving to the calculation of word frequencies and differences between the word frequencies, and concluding with the application of a statistical association analysis based on the word frequencies of selected words to identify the events.

In the first step, we determined the reference text corpus. For our tests we selected “The Guardian” and “The Times” as representatives of British coverage. For Germany, we selected “Süddeutsche Zeitung” and “Welt”. The decision for a specific newspaper does not seem really important for our analysis because we are looking only for outstanding events. We suppose that these events are reported in all newspapers as well as in television and broadcast independently from the political or cultural tendency of the medium. Here we emphasise that we are not interested in how something is reported but in what is reported. We collected all articles published in a two-year period (December 2004 to November 2006) in the sections of national and international news as well as all articles published on the first page of each newspaper edition of the selected newspapers. Finally, the reference text corpus for Great Britain consists of 102.949 articles composed of 38.994.210 words and the corpus for Germany consists of 56.845 articles composed of 16.942.771 words. We consider the words in these corpora as normal use of vocabulary in newspaper articles. The corpora also include the articles of the time period to be analysed. All texts were automatically lemmatised by the

programme TreeTagger (<http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/>). Lemmatisation refers to the matching of all different forms of a word regardless of whether its root is the same, e.g. 'say' as well as 'said' share the same lemma. We assume that using the lemmata instead of the words will lead to clearer event groups.

In the second step, we prepared the event text corpus, containing newspaper texts over a specific time period where we expected to detect the major events. We illustrate the functionality of the procedure which we outlined in the previous section with an example. We were interested in events occurring at the beginning of the data collection phase of the ESS in September 2006. Our interest was focused on major events in Great Britain and Germany during the first week of September (week 36). As event text corpus, we selected the specific texts published in the above mentioned newspapers in the specific week. With these texts we expected to identify the major events which occurred in this week.

These two steps are the most relevant steps in our procedure because they establish the base for isolating words as indicators for events.

In the third step, we calculated word frequencies for all words of the reference text corpus and for all words of the event text corpus. Additionally, the relative frequency of each word was calculated as the proportion of the total number of words in the text.

Afterwards, we calculated the differences between the relative frequencies of the words of the reference text corpus and the frequencies of the words in the event text corpus. For our purposes we have to use relative differences because for words with higher relative frequencies higher differences are expected. To avoid this effect we calculated the relative differences as the portion of the total number of words in the reference text corpus.

Words were then sorted by relative differences and the words with the highest relative differences were used as event words in the further analysis following our assumption that these differences are indicators for specific events.

In our example we set up two more restrictions. Firstly, we removed all words that have very low frequencies, in our case smaller than 25, because of our assumption that major events are reported frequently in newspapers in a specific time period and can be identified by frequently used words. Moreover, we determined that we consider only words in our analysis which are found at least in two percent of all articles of the event text corpus. These restrictions are necessary because the importance and relevance of an event can be determined by the frequency of reporting. Additionally, we limited the number of words selected for the factor analysis to 30 words with the highest deviation from the reference text corpus and which, additionally, meet all other conditions. The specific cut-off points are somewhat arbitrary but can be based on some familiarity with the data and the decision on how many events should be handled as major events. Looser restrictions result in more events to be considered in the following analysis. The selected words can be found in table 1.

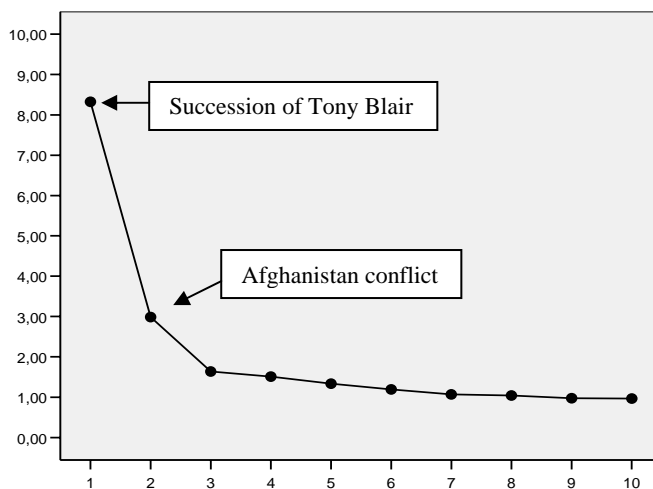
We applied an exploratory factor analysis to identify the latent semantic fields of the event words within the text under examination and to identify reported events. The goal is the representation of the latent semantic fields of the correlations of word frequencies. The factor analysis allows us to replace many more or less correlated variables by few independent factors without crucial information loss.

Table 1: Selected Words

Great Britain	Afghanistan, Blair, Blairite, Blairites, Brown, Campbell, Festival, Gordon, Johnson, Kabul, Laden, Milburn, NATO, Nimrod, Operation, PM, Science, Steve, War, Watson, Wright, aide, chancellor, code, departure, festival, leadership, quit, resignation, timetable
Germany	Kampusch, Seeblockade, Natascha, Lebensmittelkontrolle, Gammelfleisch, Schnappauf, Datei, Fleisch, Anforderung, Gesundheitsfonds, Bonn, Marine, Beirut, Seehofer, Schiff, Ware, Blockade, Labour, Küste, libanesisch, Libanon, Überwachung, Skandal, Terror, Taliban, kontrollieren, Anti, Gesundheitsreform, Fond, CIA (Kampusch, sea blockade, Natascha, checking foodstuff, rotten meat, Schnappauf, file, meat, requirement, "health pool", Bonn, Navy, Beirut, Seehofer, ship, goods, blockade, Labour, coast, Lebanese, Lebanon, control, Scandal, terror, Taleban, control, anti, "health reform", pool, CIA)

Based on word frequencies, correlations were calculated and subsequently we performed an exploratory factor analysis using principle component as extracting method, and Varimax rotation. Our strict rules for selections of the event words (e.g. the variables for the analysis) and our relatively strong expectations concerning the factor patterns to be revealed in the text help us to interpret the factor results.

To limit the number of dimensions, we used the screeplots of Eigenvalues. For Great Britain the screeplot (fig. 1) indicates that the first two factors can be interpreted as indicators for events. These two factors (see table 2) explain the most amount of variance.

Figure 1: Screeplot of the Eigenvalues (Great Britain)

The first factor comprises the words "Blairite, Watson, Brown, Milburn, chancellor, Gordon, leadership, aide, Blair" and represents the event "succession of Tony Blair". Blair announced his resignation and ignited a discussion about his successor. Blair was not willing to give Brown the public endorsement he wanted as his successor.

The second factor is explained by the words "Nimrod, NATO, Kabul, Afghanistan, operation", which are indicators for the event "Afghanistan conflict". The continued discussion of the reinforcement of the troops of the NATO and the engagement of British

soldiers as well as the increasing number of killed soldiers are the main topics concerning this event.

Table 2: Great Britain: Factor loadings (Varimax rotated) of the selected factors

	1	2
Nimrod	-,026	,539
Blairites	,428	-,047
Blairite	,812	-,009
Watson	,670	-,010
NATO	-,025	,859
Kabul	-,043	,801
Brown	,878	-,031
departure	,460	-,046
timetable	,166	-,050
Milburn	,542	-,020
Afghanistan	-,031	,852
Laden	,005	-,013
chancellor	,919	-,007
Johnson	,349	-,040
PM	,360	-,038
Science	-,024	-,058
Wright	,073	,054
Gordon	,779	-,049
resignation	,434	-,041
Operation	-,033	,728
quit	,463	-,040
Steve	-,081	-,102
leadership	,723	-,061
aide	,592	-,031
Festival	-,030	-,060
Campbell	,097	-,062
Blair	,840	-,034
festival	-,060	-,069
War	-,088	-,106
code	,011	-,058

For Germany, too, the screeplot (fig. 2) indicates that two factors can be interpreted.

The first factor comprises the words "Seeblockade, Anforderung, Beirut, Blockade, libanesisch, Libanon" ("sea blockade, request, Beirut, blockade, Lebanese, and Lebanon") and represents the event "mission of German Navy in Lebanon conflict". The German Navy should secure the sea frontier in the Lebanon conflict as part of the NATO to support freedom. But the government of Lebanon formulated some conditions before allowing foreign troops in the country. Therefore, the German mission was delayed.

The second factor is explained by the words "Lebensmittelkontrolle, Gammelfleisch, Schnappauf, Fleisch, Seehofer, Ware, and Skandal" ("checking foodstuff, Rotten Meat, Schnappauf, Meat, Seehofer, Goods, Scandal), which are indicators for the event "Rotten Meat Scandal". In Germany, rotten meat was found in a wholesaler store in Bavaria. This resulted in a discussion between the German Minister of Agriculture Seehofer and the

Bavarian Minister for Environment, Health and Consumer Protection Schnappauf about the responsibility for and consequences of this scandal.

Figure 2: Screeplot of the Eigenvalues (Germany)

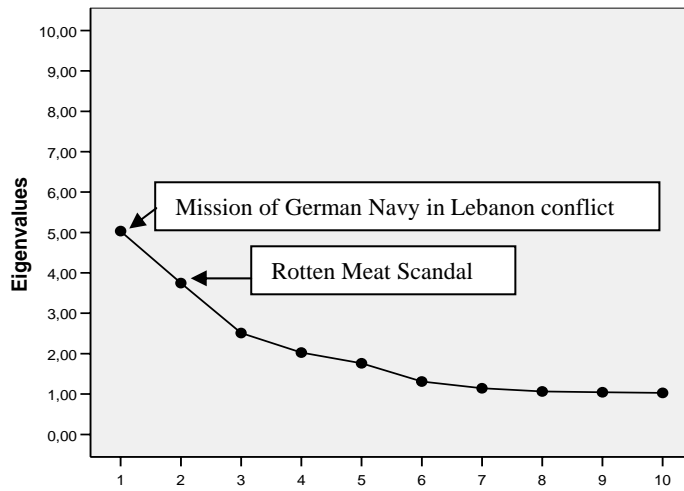


Table 3: Germany: Factor loadings (Varimax rotated) for the selected factors

	1	2
Kampusch	-,020	-,039
Seeblockade	,802	-,031
Natascha	-,033	-,045
Lebensmittelkontrolle	-,062	,709
Gammelfleisch	-,032	,763
Schnappauf	-,043	,589
Datei	-,017	-,004
Fleisch	-,029	,678
Anforderung	,575	-,029
Gesundheitsfonds	-,036	-,048
Bonn	-,041	-,088
Marine	,284	-,068
Beirut	,904	-,043
Seehofer	-,061	,774
Schiff	,239	-,057
Ware	-,005	,635
Blockade	,714	-,053
Labour	-,007	,031
Küste	,323	-,065
libanesisch	,820	-,059
Libanon	,788	-,084
Überwachung	,137	,145
Skandal	-,098	,636
Terror	-,092	-,148
Taliban	-,085	-,238
kontrollieren	,179	,238
Anti	-,049	-,067
Gesundheitsreform	-,056	-,080
Fond	-,032	-,040
CIA	-,064	-,164

5. Discussion

In conclusion, one can say that our procedure leads to good results when searching for major events in newspaper articles without too much pre-processing work done by humans. The advantage of such an approach is that

- it identifies events uniformly (in contrast to manual coding as described above),
- no knowledge about events is necessary in advance (no a priori categorization as necessary with a dictionary-based approach), and
- the number of events selected can be regulated by setting the analysis parameters more or less restrictively (number of words, frequency of words, etc.)

The procedure offers a systematic way to create the event data base for all countries participating in the ESS. Nevertheless, the short description of the different selected factors based on (automatically selected) newspaper articles remains an important task. Moreover, the decision about the presumed effects of an event on respondent behaviour remains: The decision of the effect of an event to be expected on respondent behaviour cannot be made with content analysis and not even by coders. In our opinion it has to be done by researchers, for example by those who developed the questionnaire and/or those who analyse the data.

Additionally, a further question remains. Texts were automatically lemmatised for our project. The lemmatisation routine works rule-based. This results in some words not combined to one root (for example, Blair, Blairite, Blairites). At the project start we assumed lemmatisation would be very important to get clearer event groups. After our tests, we propose to prove the necessity of lemmatisation. In our experiences the words with different word forms are always combined in one factor. For working without lemmatisation it could be helpful to change some of the analysis parameters (word frequencies, number of event words). Although programmes for lemmatisation for different languages are available, it is a lot of work to apply them to the amount of texts necessary for the reference text corpora. Further test are planned and necessary.

References

- Landmann, Juliane, Zuell, Cornelia (2004): Computerunterstützte Inhaltsanalyse ohne Diktionär? Ein Praxistest. ZUMA-Nachrichten 54.
http://www.gesis.org/Publikationen/Zeitschriften/ZUMA_Nachrichten/documents/pdfs/54/09_Landmann.pdf
- Schrodt, Philip A., Deborah J. Gerner (2001): Analyzing International Event Data.
<http://web.ku.edu/keds/papers.dir/automated.html>
- Stoop, Ineke (2002): Context and Event data. Guidelines for National Coordinators.
http://www.scp.nl/users/stoop/ess_events/guidelines_events.htm
- Stoop, Ineke (2004): Event data collection Round 2. Guidelines for ESS National Coordinators
http://naticent02.uuhost.uk.uu.net/questionnaire/context_event/event_reporting_guidelines_round_2.pdf
- Stoop, Ineke (2006): Event Reporting Guidelines.
http://naticent02.uuhost.uk.uu.net/ess_docs/R3/Fieldwork/r3_event_reporting.pdf

About the Authors

Cornelia Zuell and **Juliane Landmann** are members of the Text Analysis group at the Center for Survey Research and Methodology (ZUMA), Mannheim, Germany. Cornelia Zuell's interests include computer-assisted text analysis methodology, text analysis software and statistical issues. Juliane Landmann earned her doctorate in Social Science at the University of Mannheim. Her interests include computer-assisted text analysis methodology and issues of organized interests. Both can be contacted at GESIS-ZUMA, P.O. Box 122155, D-68072 Mannheim, Germany; tel. +49 621 1246147; fax +49 1246100; e-mail zuell@zuma-mannheim.de.

