Using computer-assisted text analysis to identify media reported events

Cornelia Zuell

GESIS – Leibniz Institute for the Social Sciences – D-68072 Mannheim – Germany

Abstract

Events like elections can influence answers to survey questions. Thus, the documentation of major events occurring during the data collection period is important. We have developed a procedure combining two approaches of computer-assisted text analysis to identify events: the reference text technique and the statistical association approach. We composed a reference text corpus based on newspaper articles of a two years' time period. The words in this text corpus are considered as general language usage. To identify events of a one week time period newspaper articles from this week are used. Word frequencies and differences between the relative frequencies of each word in both text corpora are calculated. The words with the highest relative differences serve as keywords. Based on the co-occurrence of these keywords, an explorative factor analysis is conducted. The resulting factors can be used as indicators of major events. Finally, the events are visualized means of a multiple correspondence analysis.

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

1. Introduction

The European Social Survey [1] (ESS) is an academically-driven social survey designed to chart and explain the interaction between Europe's changing institutions and the attitudes, beliefs and behaviour patterns of its diverse populations. Now in its fourth round, the survey covers over 30 nations and employs the most rigorous methodologies. A repeat cross-sectional survey, it has been funded through the European Commission's Framework Programmes, the European Science Foundation and national funding bodies in each country.

Since the beginning of the first data collection period of the ESS, the societal events occurring during the data collection phase have been documented in all countries of the ESS (Stoop, 2007). The assumption is that respondent behaviour or answers to some questions are influenced by significant events in various areas (see, for example, Das and Bezemer, 2005). Hence, the impact of an event must be considered and, whenever possible, controlled across the countries participating 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 coordinators (Stoop, 2002, 2004, 2006). 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

[1] http://www.europeansocialsurvey.org/.

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 6 (in Norway) to 197 (in Spain) for the whole data collection period. The number of reported events varies from 1 to 15 per week. 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.

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 et al. (1994: 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 were looking for all occurrences which can influence public opinion and therefore respondent behaviour (political activities as well as natural catastrophes). Therefore, we define events as all significant or major occurrences reported in mass media which are significant, interesting, exciting or unusual. Such events can be political or economical activities as well as catastrophes (natural catastrophes, disasters, or accidents). To give some examples, political/economical events are events like terror attacks, international conflicts, elections, bank scandals, insolvencies, or strikes. Catastrophes are, for example, floods, earthquakes, plain crashes, etc. To qualify as a major event, the following must apply: The news are reported for at least several days in mass media and give rise to widespread public discussion and/or to a substantive increase in media use. Additionally, major events attract wide and long-lasting attention.

Because of the problems with manual event reporting we decided to use computer-assisted text 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 Method

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. Schrodt and Gerner (2001: 2-7) mention in their 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.

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 and 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 the research we used 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.

In the following we will describe the procedure in a more detailed fashion (see also Landmann and Zuell, 2008). In general, the procedure can be described in five 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 (see Fig. 1).

In the first step, we determined the reference text corpus. For the 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 the analysis because we were 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. The Lexis Nexis data base (http://www.lexis-nexis.com) was used to obtain machine-readable articles. 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.

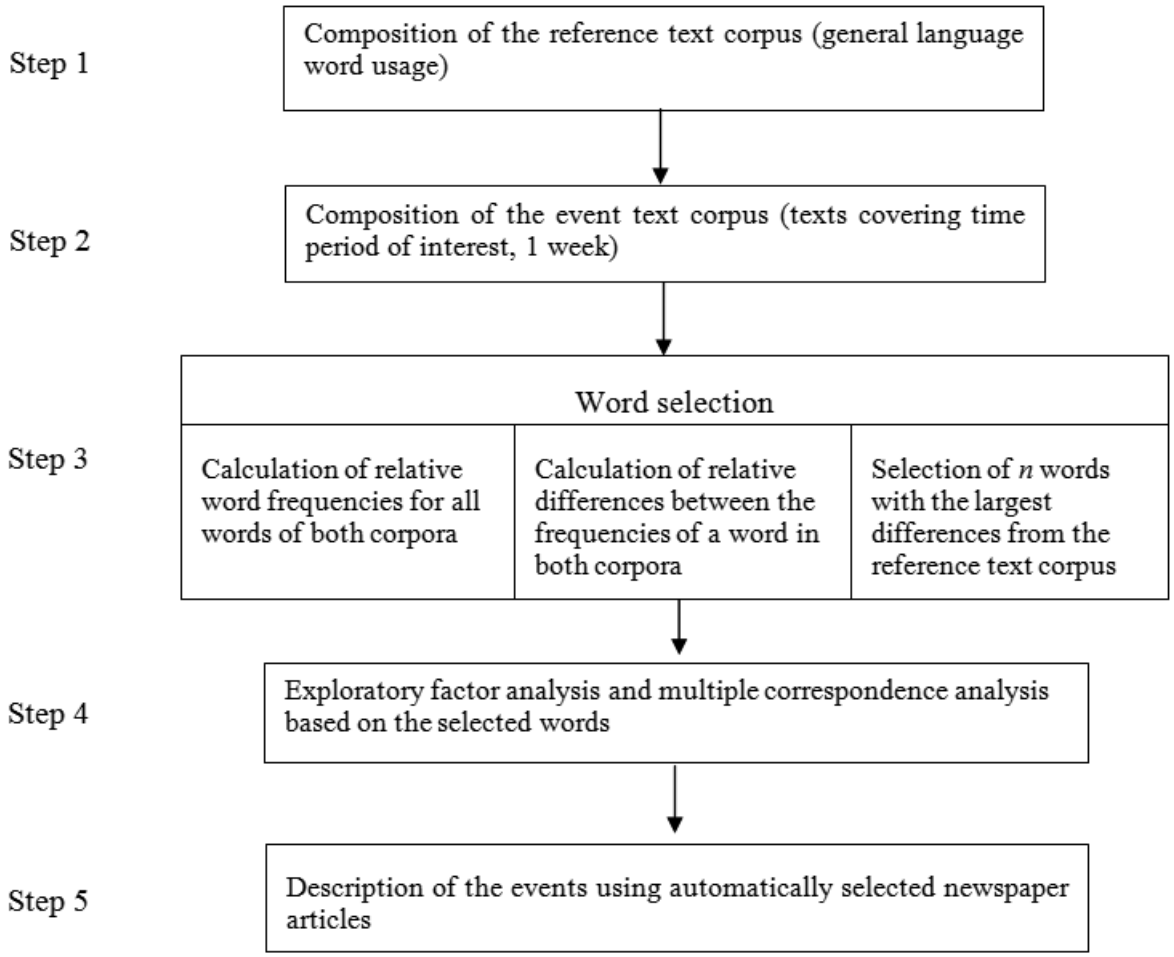

Figure 1: Workflow of the Event Data Procedure

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 with an example in which we were interested in events occurring at week 41 (October 2006). Our interest in the following example was focused on major events in Great Britain and Germany during this week. 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 first two steps are the most relevant steps in the 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. Then we calculated the

differences between the relative frequencies of the words of the reference text corpus and the relative 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 the 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.

In the fourth step 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. Additionally, we used multiple correspondence analysis (MCA) to visualise the results.

3. Some Results

The words listed in Tab. 1 have the greatest differences in relative word frequencies between the reference text corpus of general language usage and the event text corpus. These words are handled as indicators for major events.

Great Britain	Anna, Paisley, master, Basra, Pyongyang, missile, Blunkett, Reid, novel, Dannatt, River, nuclear, Jong, Straw, possession, Kim, apartment, presence, Korea, bridge, prise, Korean, ethnic, sanction, Muslim, extract, test, North, footage, veil
Germany	Anna, Atomtest, Barroso, Bremer, EADS, Hamburger, Hochschule, Il, Jong, Journalistin, Jugendamt, Kevin, Kim, Mord, Nordkorea, Pjöngjang, Politkowskaja, Präsidentschaft, Putin, Putins, Sanktion, Südkorea, Test, Tschetschenien, Uni, Video, Völkermord, Wladimir, nordkoreanisch, sogenannt

Table 1: Selected words for Great Britain and Germany (week 41)

Based on the frequencies of the selected words per newspaper article, 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. 2) indicates that the first two factors can be interpreted as indicators for events. These two factors explain the most amount of variance.

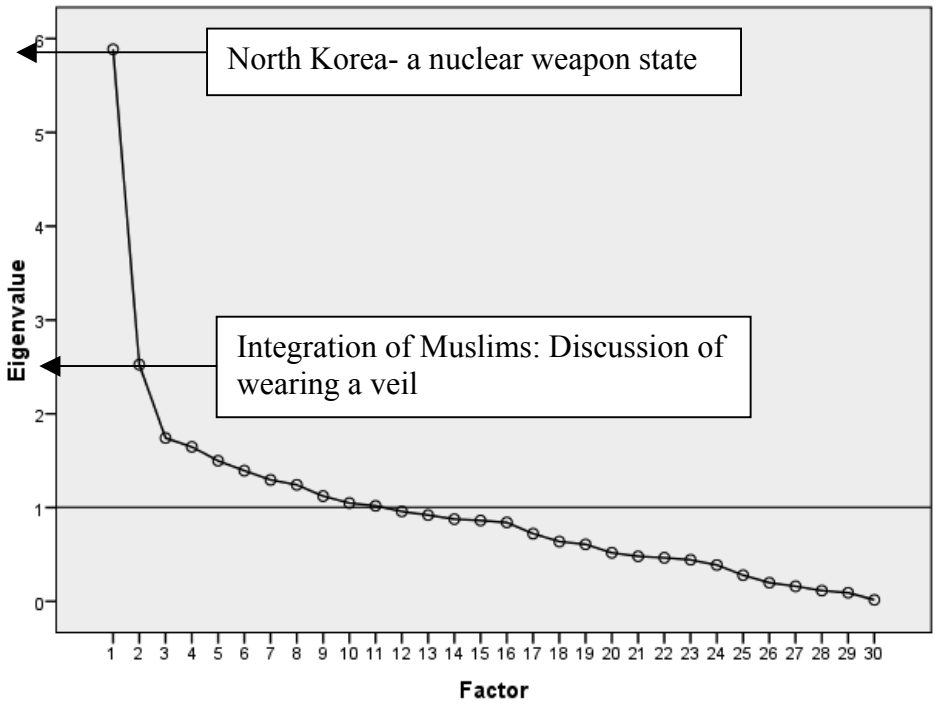

Figure 2: Screeplot of the Eigenvalues (Great Britain)

The first factor comprises the words "Pyongyang, Korea, Korean, North, nuclear, sanction, missile, test" and represents the event "North Korea – a nuclear weapon state" (see Tab. 2 for the factor loadings). After North Korea claiming its first nuclear weapons explosion, the UN condemned the test and began negotiations on imposing tougher sanctions against Kim Jong-il's reclusive state.

The second factor is explained by the words "Veil, Straw, Muslim", which are indicators for the event "Integration of Muslims: Discussion about wearing a veil". In Britain most people want Muslims to try harder to integrate. The acceptance of Britain's Muslim community runs alongside fears about the development of a divided society. The British Politician Straw was expressing his opinion, which is that veils "sucks". He described the veil as a "visible statement of separation and difference" between Muslims and non-Muslims.

To visualize the association between the words and the resulting events we used a multiple correspondence analysis (MCA). Instead of word frequencies per newspaper article we used dummy variables "word used/not used" per article. The first two dimensions represent again the two major events (see Fig. 3).

As for Great Britain, for Germany, the screeplot (Fig. 4) indicates that two factors can be interpreted.

The first factor comprises the words "Atomtest, Pjöngjang, Nordkorea, nordkoreanisch, Südkorea, Sanktion (nuclear test, Pyongyang, North Korea, North Korean, South Korea, sanction)" and represents the event "North Korea as a nuclear weapon state". This major event is the same as reported in Great Britain.

The second factor is explained by the words "Politkowskaja, Anna, Journalistin, Tschetschenien, Mord (Politkowskaja, Anna, journalist, Chechnya, murder)", which are indicators for the event "Murder of Russian journalist". The Russian journalist Anna Politkowskaja was murdered. She was known as one of the hardest detractors of Putin and was honored by many awards for courageous journalism in the past. In the Russian newspaper Nowaja Gaseta should appear an article about torture in Chechnya by her, but the text has not reached the editorial office. It was assumed that is was a political murder. Worldwide there was a claim to solve this case.

Country	*Factor and items*	*Loadings*
Great Britain	**North Korea - a nuclear weapon state**	
	Pyongyang	.764
	Korea	.857
	Korean	.587
	North	.830
	nuclear	.895
	sanction	.592
	missile	.684
	test	.799
	Integration of Muslims	
	Veil	.905
	Straw	.896
	Muslim	.922
Germany	**North Korea - a nuclear weapon state**	
	Atomtest	.854
	Pjoengjang	.825
	Nordkorea	.848
	Südkorea	.668
	nordkoreanisch	.711
	Sanktion	.765
	Murder of Russian journalist	
	Politkoskaja	.940
	Anna	.894
	Journalistin	.903
	Tschetschenien	.570
	Mord	.767

Table 2: Factor loadings (>.5) of the selected words

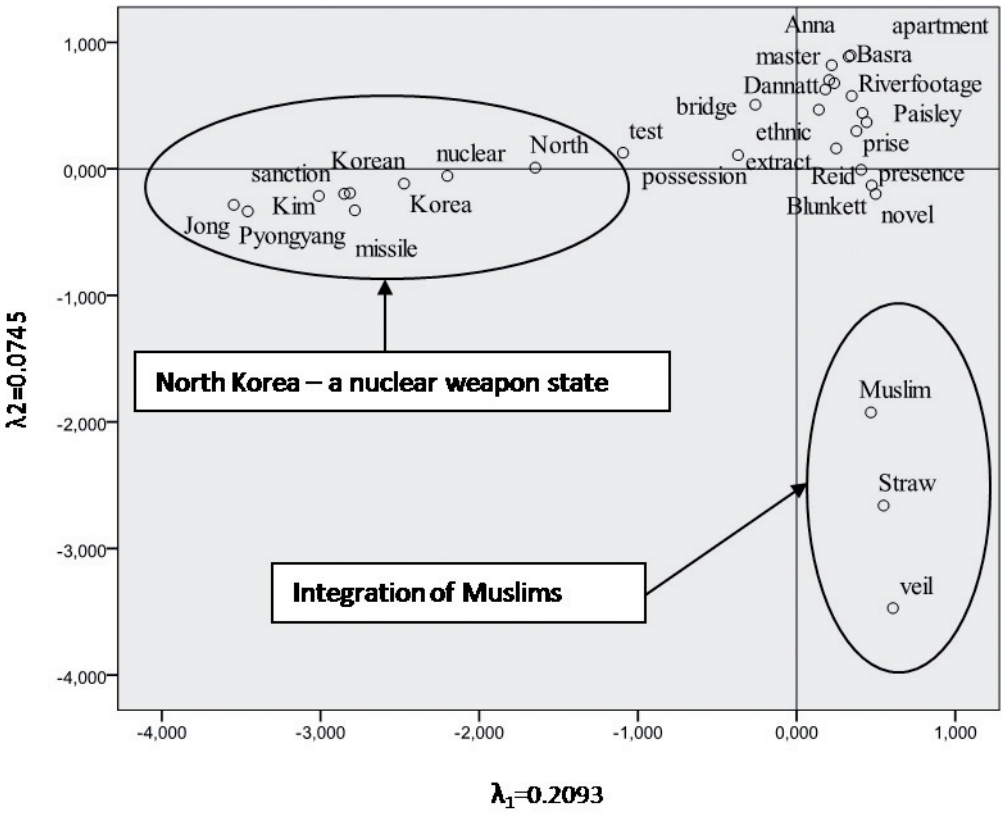

Figure 3: Great Britain: The first two dimensions of the MCA

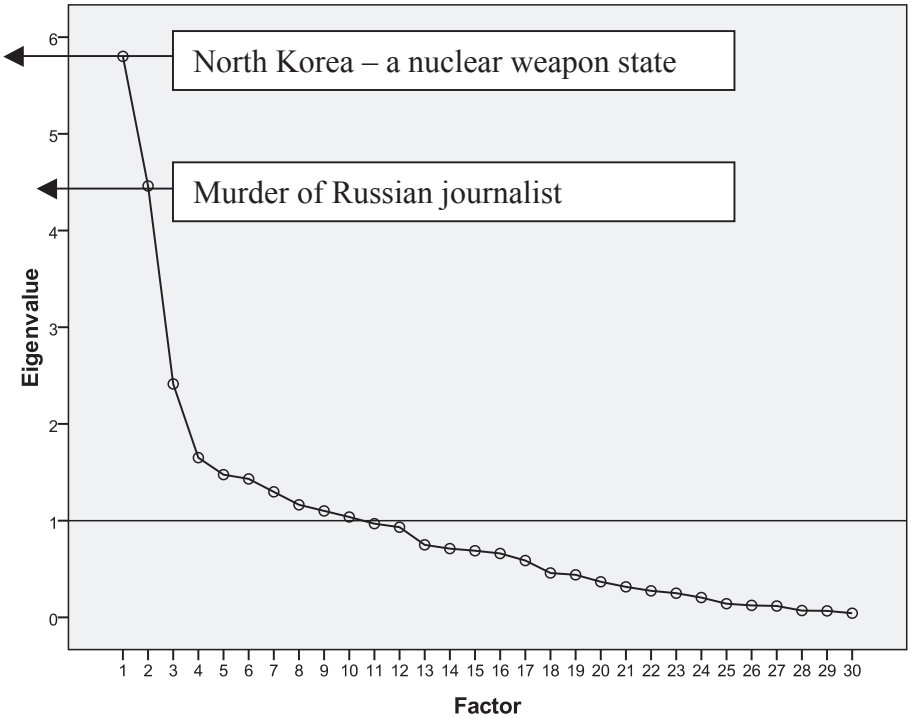

Figure 4: Screeplot of the Eigenvalues (Germany)

The results of the MCA for the 30 German words can be found in Fig. 5.

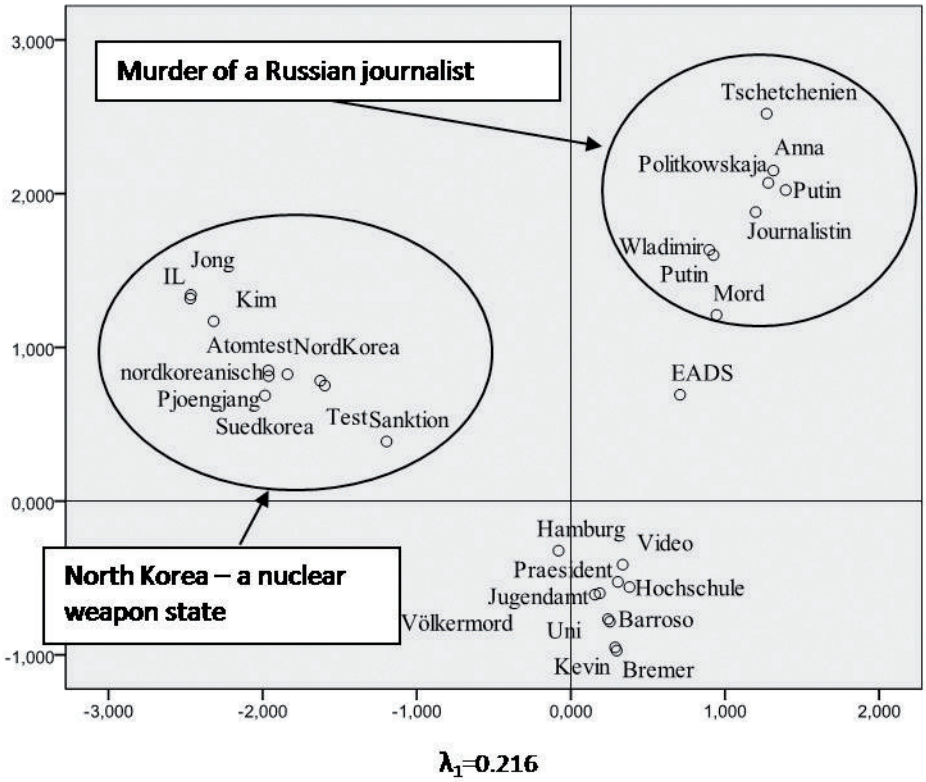

Figure 5: Germany: The first two dimensions of the MCA

4. Some Technical Considerations

4.1. Languages

I tested the procedure for Spanish newspapers in a reduced fashion, too. For Spanish event reporting we used only one newspaper ("El País") and four weeks of event reporting. Although we did not present the results here, they are convincing for Spanish texts and can be used in a similar way as for English and German event reporting.

The results of our tests suggest that the procedure will work for all word based languages (in contrast to, for example, syllable based languages). This means it will work for all languages used in countries participating in the ESS.

4.2. Newspaper Selection

Actually, the decision for a specific newspaper is not really critical (see above). The ESS reporting only looks for outstanding events, and these events will be reported in all newspapers, as well as in television and broadcast, irrespective of the political or cultural tendency of the medium. Additionally, as said before, we are not interested in *how* something is reported but in *what* is reported.

In our first tests based on event reporting 2002 we used one newspaper per country (Germany and Great Britain). As newspaper for Great Britain coverage we used "The Guardian". We

searched for monthly reported events following the rules for event reporting of the first round. In August 2002 one of the factors represented the article "*The Guardian* Edinburgh International TV Festival." This factor, however, could be attributed to the specific newspaper: "*The Guardian"* had a special interest in highlighting its own festival and arousing public interest. Therefore, this event was not a major event and should be disregarded. This somewhat qualifies our assumption that the selection of the newspaper plays no significant role. For all further testing we decided to use two different newspapers per country to avoid such newspaper specific effects.

4.3. Selection of Newspaper Articles

We collected all articles published on the first page of each newspaper edition as well as all articles published in *the se*ctions "International Politics" and "National Politics". The use of only the first page of each newspaper is not suitable because these pages differentiate completely in the newspapers. For example, "The Times" publishes one (major) article on the first page. Other newspapers have only short notes on the first page and link articles on further pages of the newspaper. Therefore, the use of different sections is absolutely necessary.

4.4. Lemmatisation

As mentioned above we used word lemmatisation because we assumed that using the lemmata instead of the words will lead to clearer event groups. The results based on German, English as well as Spanish texts show that lemmatisation only has a very small impact on the results.

Week	*With Lemmatisation*	*Without Lemmatisation*
36	Mission of German Navy in Lebanon conflict Rotten Meat Scandal	Mission of German Navy in Lebanon conflict Rotten Meat Scandal
37	Pope in Germany Mission of German Navy in Lebanon conflict	Pope in Germany Mission of German Navy in Lebanon decided UN try to force Germany to send the Navy to the Lebanon
38	German State Elections	German State Elections
39	Displacement of the Opera Idomeo in Berlin Transrapid Train Crash Islam Conference	Transrapid Train Crash Displacement of the Opera Idomeo in Berlin Islam Conference
40	Election in Austria Airbus-crisis "Health Fund"	Election in Austria Airbus-crisis "Health Fund"
41	North Korea Nuclear Test Murder of Russian Journalist	North Korea Nuclear Test Murder of Russian Journalist

Table 3: Events in Germany with and without lemmatised texts

Two types of differences may occur:

1. Different sequence of events, e.g. event 1 gets event 2 and vice versa (for example, see Tab. 3, week 39).

2. Events may be split into two events with different special aspects (for example, see Tab. 3, week 37 "Lebanon conflict"). The first Lebanon event emphasizes the decision of sending German soldiers to the Lebanese coast. The second Lebanon event points out that the United Nations try to convince or even try to force Germany to participate in the mission and to support the plan of saving the Lebanese coast.

The main reasons for these differences are that some event words have lower frequencies without lemmatisation. To give an example, the word Republican (lemmatised) may have a frequency of 51, without lemmatisation the word is split in Republican (frequency 34) and Republicans (frequency 17). This causes that the word will be not regarded as an event word. Additionally, without lemmatisation some words are regarded twice in the event word list because they occur in two different forms with high frequencies. This causes a disregard to other words.

Our recommendation for the event reporting of the ESS is therefore that lemmatisation is not absolutely necessary for German, English and Spain. It is helpful to use it if available. Nevertheless, the results without lemmatisation are good enough for the event reporting purposes of the ESS. Using other languages, the necessity for lemmatisation has to be considered. Lemmatisation may be important if a language allows very frequently very different word forms of substantives.

5. Conclusion

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.

References

Das E., Bushman B., and Bezemer M. (2005). *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.

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

Landmann J. and Zuell C. (2004). Computerunterstützte Inhaltsanalyse ohne Diktionär? Ein Praxistest. *ZUMA-Nachrichten*, 54, 117-140.

Landmann J. and Zuell C. (2008). Identifying Events Using Computer-Assisted Text Analysis. *Social Science Computer Review*, 26(4): 483-497.

Schrodt Ph.A. and Gerner D.J. (2001). Analyzing International Event Data. http://web.ku.edu/~keds/papers.dir/automated.html

Stoop I. (2002). Context and Event data. Guidelines for National Coordinators. http://www.europeansocialsurvey.org/index.php?option=com_content&view=article&id=88:contextual-and-event-data&catid=116:questionnaire&Itemid=133.

Stoop I. (2004). Event data collection Round 2. Guidelines for ESS National Coordinators http://www.europeansocialsurvey.org/index.php?option=com_docman&task=doc_download&gid=37&itemid=80

Stoop I. (2006): Event Reporting Guidelines. http://www.europeansocialsurvey.org/index.php?option=com_content&view=article&id=88:contextual-and-event-data&catid=116:questionnaire& Itemid=133

Stoop I. (2007). If it bleeds, it leads: the Impact of Media-Reported Events. In Jowell J., Roberts C., Fitzgerald R. and Eva G., editors, *Measuring Attitudes Cross-Nationally. Lessons from the European Social Survey*. Sage.