Freie Universität Berlin

# HOW TO GET RID OF THE NOISE IN THE CORPUS:

## Cleaning Large Samples of Digital Newspaper Texts

Cathleen Kantner and Amelie Kutter with the collaboration of Andreas Hildebrandt and Mark Püttcher

December 2010

**Berliner Arbeitspapier zur Europäischen Integration Nr. 17**

***Berlin Working Paper on European Integration No. 17***

This paper is part of a series of method papers on the use of advanced text analytical methods in the political and social sciences. The papers where presented at the workshop "How to Analyse Millions of Texts and Still Be Home for Tea? Innovations in Textual Analysis in the Social Sciences" conducted at Freie Universität Berlin in May 2010. The workshop was organised by Amelie Kutter and Cathleen Kantner and it was co-financed by the FP6 Integrated Project "RECON – Reconstituting Democracy in Europe" and the Jean Monnet Centre of Excellence "The EU and its Citizens" at Freie Universität Berlin.

Arbeitsstelle Europäische
Jean Monnet Lehrstuhl Integration

## Content

Die Berliner Arbeitspapiere zur Europäischen Integration werden von Prof. Dr. Tanja A. Börzel, Jean Monnet Lehrstuhl für Europäische Integration und Leiterin der Arbeitsstelle Europäische Integration am Otto-Suhr-Institut für Politikwissenschaft der Freien Universität Berlin, veröffentlicht. Die Arbeitspapiere sind auf der gemeinsamen Internetseite von Jean Monnet Lehrstuhl und Arbeitsstelle verfügbar: http://www.fu-berlin.de/europa

The Berlin Working Papers on European Integration are published by Prof. Dr. Tanja A. Börzel, Jean Monnet Chair for European Integration and Director of the Centre for European Integration of the Otto Suhr-Institute for Political Science at the Freie Universität Berlin. The Working Papers are available on the joint website of the Jean Monnet Chair and the Centre: http://www.fu-berlin.de/europa

## Die Autoren / The Authors

Cathleen Kantner is Professor of International Relations and European Studies at the Institute for Social Sciences of the University of Stuttgart. Her current research focuses on public debates about European security and defence issues with particular reference to shared normative beliefs. Previously, she has been at the Otto Suhr Institute for Political Science of the Freie Universität Berlin, the Humboldt University Berlin, the Bundeswehr Institute for Social Sciences and the Robert Schuman Centre for Advanced Studies at the European University Institute in Florence.

Amelie Kutter received her MA (Diplom) in political science from Otto-Suhr Institute at Free University Berlin in 2001 and has since then worked as lecturer and researcher at TU Dresden, European University Viadrina Frankfurt (Oder), Freie Universität Berlin, with research stays at Institut d'études politiques de Paris (CERI, Sciences Po), Institute for Advanced Studies at Lancaster University, and at the Centre of International Relations in Warsaw. She was member of EU-funded projects on EU enlargement and European public spheres and received scholarships from the German Federal Ministry of Education and Research and the German Academic Exchange Service for her PhD-related work.

Andreas Hildebrandt (Department of Linguistics, University Potsdam) and Mark Püttcher (Freie Universität Berlin) worked as student assistants in the research project “In search of a new role in world politics. The common European foreign, security and defence policies (CFSP/ESDP) in the light of identity-debates in the member states’ mass media”, directed by Thomas Risse and Cathleen Kantner at Freie Universität Berlin.

## Abstract

Large digital text samples are promising sources for text-analytical research in the social sciences. However, they may turn out to be very troublesome when not cleaned of the ‘noise’ of doublets and sampling errors that induce biases and distort the reliability of content-analytical results. This paper claims that these problems can be remedied by making innovative use of computational and corpus-linguistic procedures. Automatic pairwise document comparison based on a vector space model will bring doublets to light, while sampling errors can be discerned with the help of textmining procedures that measure the ‘keyness’ of a document, i.e. the degree to which it contains or does not contain keywords representing the research topic.

**Keywords**: content analysis, corpus, textmining, semantic fields, tf-idf, keyness, newspaper, military intervention

## 1. Introduction

The availability of digital text archives has given rise to and aided numerous text analytical research projects in the social sciences. Fascinated by the higher degree of representativeness which large-n content analysis promises to produce, as well as with the options of statistical analysis, researchers have embarked on sampling and annotating large collections of digital texts. Web-based newspaper archives, in particular, suggest themselves as source for the investigation of patterns of publicised opinion and political communication on a broader scale.[1] By means of simple queries in publicly accessible web archives such as LexisNexis or Factiva, one can easily retrieve vast collections of texts on a particular issue from a particular source.

However, the promises of easy access and sheer magnitude may well turn out to be a double-edged sword when the implemented query terms are semantically ambiguous and when the source archive is full of omissions and doublets. As a result, even a properly balanced corpus of texts, e.g. balanced with regard to time-span, sources, languages, or wording related to the issue under investigation, may be incomplete and contain a lot of ‘noise’ (Gabrielatos 2007: 6). Incompleteness of newspaper corpora results from changing copyright conditions due to which certain types of news articles are not accessible or disappear from the web archive during the sampling period (e.g. articles published by free-lance authors). ‘Noise’ emerges because the query terms implemented in the search engines of the web-archive may be semantically ambiguous and produce sampling errors, i.e. retrieve documents from the web that are not relevant to the research question. In addition, news web archives often unsystematically assemble full and truncated texts, doublets, or quasi-doublets such as re-occurring publications, different editions, draft and final or print and online versions. Incompleteness can be balanced using further archives, and it can be documented and considered during the analysis. The ‘noise’ of sampling errors and doublets, however, if not identified and shut off from the remaining corpus, may well undermine the validity and reliability of the findings.

This paper suggests how to markedly reduce noise in the corpus with the help of computer-aided methods. It claims that, by using insights of computational and corpus linguistics for corpus cleaning, researchers will be able to considerably advance the

[1] In large-scale investigations on transnational political communication, such sources were used, for example, by: Baker/McEnery (2005); Kantner (2006); Kantner (2009); Kantner et al. (2008); Koenig et al. (2006); Kutter (2007); Liebert (2007); Renfordt (2009); Trenz (2004).

reliability and validity of content analyses in the social sciences that draw on large-n text samples. Problems of bias induced by erroneously sampled documents can be solved by keyword analyses that draw on linguistic theories of semantic fields and text-mining tools. Problems of redundancy in the sample – doublets and quasi-doublets – can be tackled by means of automatic pairwise document comparison based on a multidimensional vector space model.[2]

These suggestions are based on the experience of constructing and cleaning a large-scale multilingual corpus of newspaper articles on war and military intervention published in six European countries and the USA between January 1990 and March 2006.[3] After cleaning, this corpus comprised 489,508 articles and about 393,268,540 words. It was sampled with the help of a standardized multi-term query from LexisNexis and a newspaper's internal electronic archive. By means of Boolean operators, this query combined keywords related to war and military intervention (e.g. war on terrorism, peace-keeping, stabilisation mission, etc.), on the one hand, and names of states and regions, on the other, where violent conflicts and/or internationally-led interventions had taken place during the period under investigation (e.g. East Timor, Karabakh, Kuwait etc.). This sampling strategy proved broad enough to capture all relevant texts on the research topic contained in the source archives; however, it also comprised irrelevant texts, e.g. texts that referred to the many derived meanings of 'war', but not to its primary meaning. Hence, to balance 'precision' and 'recall' in the corpus (Cowdhury cited in Gabrielatos 2007: 6), we had to adopt a sophisticated cleaning procedure.

[2] Many thanks to the participants of the JMCE&RECON Workshop on Computer-Aided Methods of Textual Analysis (FU Berlin, May 27-28th, 2010) and to Marianne van de Steeg, in particular, for commenting on a first version of this article.

[3] The corpus includes articles published on war and military intervention during the years 1990-2006 from: the New York Times, The Washington Post (USA); The Guardian, The Times (UK); The Irish Times (Ireland); NRC Handelsblad, De Volkskrant (Netherlands); Le Monde, Le Figaro, Les Echos (France), Süddeutsche Zeitung, Frankfurter Allgemeine Zeitung (Germany); Der Standard, Die Presse (Austria). The corpus was used for a longitudinal analysis of media content on military intervention, the role of international actors, European foreign, security and defence policies, international law, and foreign policy identities; carried out within the frames of the project "In search of a new role in world politics. The common European foreign, security and defence policies (CFSP/ESDP) in the light of identity-debates in the member states' mass media", directed by Prof. Dr. Thomas Risse and PD Dr. Cathleen Kantner at Free University Berlin. It was funded by the German Research Foundation (DFG; contract no. RI 798/8)' and by the Sixth Framework Programme of the European Commission as part of the Integrated Project 'RECON – Reconstituting Democracy in Europe', (contract no. CIT4-CT-2006-028698) in the years 2005-2010. We wish to thank FAZIT foundation for sponsoring access to the archive of the Frankfurter Allgemeine Zeitung.

Before cleaning, the articles were parsed and imported into a MySQL database and automatically annotated with an identification number and meta data on the heading, the body, and the length of the text; the country of origin; the source; the publication date; and the author and section (if available). This infrastructure enabled various computational operations necessary for data management and for the procedures described in this paper.

In the next section, we will address the issues faced regarding sampling errors to then go on to discuss the problem of doublet identification and removal. The conclusions will give a summation of the added value of the procedures developed in the project and the research desiderates for the cleaning of large text samples used for content analysis in the social sciences.

## 2. Identifying and Removing Sampling Errors

Sampling errors are documents retrieved from the web archives and included into the raw text sample of a research project, despite the fact that they do not deal with the topic or primary text genre of that research project. In the case of our project on wars and military interventions in the press, many articles were sampled that contained relevant keywords of the query[4], but in fact did not refer to wars and military interventions, or that dealt with war and military interventions, but not in text genres that are usually associated with political communication and political journalism. The former type of articles dealt, for instance, with touristic information on countries in which civil wars took place some time ago (e.g. Croatia), legal disputes containing military metaphors, or sports events (e.g. in a soccer game one player comes from Somalia and in the 20th minute there is an intervention by the referee). The latter type of sampling errors comprised advertisements, calendars of events, and book or movie reviews that mentioned international conflicts. As the focus of the project was, however, on public-political news and debate, the formal criterion was adopted that these texts genres had to be excluded.[5]

[4] The query included two groups of keywords: the names of *crisis countries* and keywords related to war and military intervention (examples are given in the introduction above). If an article contained at least one search-word from both keyword groups, it was automatically retrieved and entered the raw sample.

[5] This decision was difficult, because cultural communication is often an important field for political communication too. There was also some discussion about an article from the sports section which narrated the biography of a famous African marathon runner who fled from his

Manual identification of sampling errors was not manageable given the large amount of data and not reliable given the many human readers involved. Moreover, meta-data on newspaper sections that, according to the formal criterion, were unlikely to contain political news and debate (e.g. 'sports', 'TV programmes', 'book reviews' etc.) could not be used as an automatic filter, because information on sections was inconsistent in the source archives and captured only for some newspapers. In turn, automatically filtering and discarding all articles that used sports-specific terms or culture, would have excluded many pieces of political journalism, such as reports on soldiers in multinational missions playing soccer or table tennis in their spare time or references to peace-keepers repairing sporting facilities in a war-torn village.

A way out of this dilemma is to develop a measure of 'keyness' that can be applied as filter in automatic text-mining procedures. 'Keyness', in the corpus-linguistic literature, denotes a statistical value of relative frequency in occurrence. Those words are considered 'key', i.e. particularly representative for the topic, contents and (generic-linguistic) features of a specific corpus, which occur consistently more frequently in the specific corpus than in reference corpora such as another specific corpus or a general reference corpus of a particular language such as the British National Corpus (Culpeper 2009; Scott 2008).[6] With regard to the particular problem of sampling errors, this general definition of keyness, of course, did not apply. What was needed was a measure that would identify *keyness of single documents within the corpus*, i.e. the frequency of 'key-words' representing the topic war and military intervention relative to the frequency of words that did not belong to that topic, but to sports, tourism and recreation, and cultural news. This measure was developed (a) by identifying the semantic field of war and military intervention, on the one hand, and the semantic fields of sports and cultural reviews, on the other; and (b) by calculating their relative density in the newspaper articles.

country during the civil war and lost much of his family. Such articles are certainly extraordinarily important to raise ordinary people's attention to 'forgotten conflicts'; however, since we were not able to systematically include the broader cultural discourse into our study, we systematically excluded it in order to not put into question the comparability of the data for different newspapers.

[6] Consistent relative frequency is given when the statistical probability as computed by an appropriate procedure (e.g. comparison of wordlists in the specific and the reference corpus) is smaller than or equal to a p value specified by the researcher (e.g. through the chi-square procedure or other statistical tests). Definition and statistical calculation of keyness is strongly associated to the software WordSmith (Mike 2010).

## 2.1 Identifying the key semantic field and erroneous semantic fields

A semantic field is a conceptual field of meaning of a natural language.[7] It consists

> "of basic key-words, which command an army of others. The semantic area may be regarded as a network of hundreds of associations, each word of which is capable of being the centre of a web of associations radiating in all directions. A word like man might have as many as fifty such associations – chap, fellow, guy, gentleman, etc." (Mackey 1965: 76).

A semantic field can be understood as "a closely knit and articulated lexical sphere where the significance of each unit is determined by its neighbours" (Ullmann 1951: 157). The units of a semantic field are words and collocations that are constitutive for the meaning of the keyword (e.g. man) we started with. A semantic field includes, however, much more than synonyms (e.g. male human, adolescent), more general categories (e.g. human) and sub-categories (types of 'man' e.g. according to age (boy, adult) or social roles (father, worker, professor) of the key-concept. In addition, it also encompasses all the words and collocations speakers typically use to describe and explain the key-category in other words and in relation to other concepts (e.g. according to social functions: provider for family, defender of family; according to gender clichés: strong, successful, virile). Therefore, a semantic field is much more open and broader than a dictionary entry.

In order to identify the key semantic field of war and military intervention and the irrelevant fields 'cultural events/reviews', 'travelling/tourism' and 'sports', we adopted a hermeneutic procedure that established the most significant semantic categories per field (Culpeper 2009). It combined automatic-inductive keyword analysis and consultation of abstract knowledge. The keyword-analysis was based on the 'concept extraction function' of the textmining software SPSS Clementine. This function produces a wordlist that ranks all words occurring in the corpus according to frequency.[8] The most often used words are usually terms like 'and', 'this' and so on.

[7] For a detailed history of the concept of 'semantic field' and the distinction of three distinct language-philosophical approaches to this concept, see Nerlich and Clarke (2000).

[8] For large newspaper samples such as the US newspapers, only every third article was included in this procedure, due to the limited processing powers of our computers. The tool 'SPSS Clementine' is designed for mining, statistical analysis and visualisation of characteristics of character strings, partly supported by a lexicon; cf. http://www.spss.com/software/modeling/modeler-pro/ (June 5th, 2010). However, the operation of included functions such as 'concept extraction' proved to be entirely intransparent, which is why we switched to WordSmith for wordlist-generation and self-designed scripts for the mining of character-strings later on.

Their display can be inhibited. The displayed 'domain list' of meaningful terms then needs to be checked and classified by human readers, frequently looking again into the respective texts for the meanings actually expressed. We manually checked all those words used in at least two percent of the articles in the respective papers. The most significant words were put on the list to cover the semantic field. We proceeded paper-wise in order to account for possible differences in wording between left-liberal and conservative papers, but eventually merged these lists per country. This systematic inductive method ensured that the indicator operated on terms which were actually used in the coverage. Additionally consulted knowledge related to the initial sampling query – the included keywords related to war and military intervention –, dictionary entries and experiences with typical sampling errors made during coding pre-tests. Drawing on this information, human readers decided whether or not to include a particular term in a term-list representing the respective semantic field.[9]

To construct the term-list of the key semantic field on war and military intervention, we started with setting the keywords of the sampling query in all possible grammatical forms, as lemmatisation in SPSS Clementine was not reliable. These terms were supplemented by terms which occurred disproportionately often in the news articles under examination, as displayed by the 'concept extraction function' of SPSS Clementine. For the semantic field 'war and intervention', the following terms were – in the context of our very specific sample[10], and to give one example – indicative in the British newspapers:

> armed force, army, attack, battle, ceasefire, danger, ethnic cleansing, enemy, enemies, force, fight, genocide, murder, missile, security, safety, soldier, terror, troop, victim, victory, violence, weapon, casualties, casualty, death, fear, massacre, civilian, bombardment, uno, nato, united nations, refugee, protection, world war, rebel, military, human rights, shooting

For the semantic field 'sports' we created lists encompassing all kinds of sports[11] and words often occurring in our specific corpus, such as:

[9] The logic behind our procedure is similar to – and inspired by – the functions of modern corpus-linguistic software. However, the software that was available to us was not specific and transparent enough to be employed for the very specific concepts we wanted to cover. So we used our initial software – especially SPSS Clementine – only for the preparatory steps of this procedure and then again for text extraction.

[10] Note that the terms might be completely different in other samples, such a newspaper sample on migration, a sample of the complete content of a newspaper, or a sample of official documents.

[11] As electronic dictionaries were not available in all the languages, Wikipedia lists had to suffice (Source for Germany: http://de.wikipedia.org/wiki/Liste_der_Sportarten (November 3rd, 2006).

champions league, championship, test game, friendly game, friendly match, qualifying match, uefa-cup, ui-cup, world cup

The same was done for the semantic field ‘travelling/tourism’. The term ‘travelling’ itself, however, had to be excluded from the list, since it is widely used in coverage on political themes: “Foreign secretary Rice is travelling to Lebanon tomorrow”. There was no encyclopaedic entry on ‘cultural events/reviews’ so the list for this semantic field was developed solely inductively. It was checked several times whether the terms indeed identified those sampling errors searched for. A slash was often used in articles listing different cultural events so ‘/’ was also added to the list as a good indicator.

## 2.2 Measuring the keyness of single documents

On the basis of the term-lists representing the key semantic field and the erroneous semantic fields, we were able to identify the keyness of individual documents, i.e. the density with which terms related to the key semantic field occurred in the document compared to terms belonging to the irrelevant semantic fields. Each term-list was transferred into a set of character strings combined by the Boolean operator OR. These sets were subsequently used for textmining and calculation in SPSS Clementine which followed a three-step procedure. First, a measure of ‘density’ was defined which considered not only the overall frequency of a term from the term-list ‘war and intervention’ in the individual documents, but weighed it in accordance with the position of the term in the text: Three points were awarded if a term was used in the title, one point for all other occurrences. The formula used for density was thus:

*D = (frequency of term weighted by points for position / number of characters per article) * 10 000.*

For all articles from the sample for which the value of the quotient equalled zero, it could be assumed that the articles did not deal with war and intervention. They were deleted from the text sample. Second, articles on sports (e.g. the childhood of a sportsman during civil war), tourism in former war regions and cultural events (e.g. reviews of books on crises) that did not include information on news paper sections could be identified and deleted. However, to avoid erroneous deletion of articles, all articles with more than four points indicating that ‘war and intervention’ wording played an important role in the article were kept and checked in a third analytical step.

Third, using the keywords from the four semantic fields, the density of each semantic field in each remaining article was calculated and depicted in a chart with SPSS Clementine. To identify the articles for deletion, the keyword density in the fields 'cultural events', 'travelling' and 'sports' were added together and compared to the density of keywords for 'war and intervention'. This method had the advantage of also identifying articles with a mixture of two fields; e.g. sports activities during summer vacation. These articles might have displayed a low density for each semantic field individually. By adding coverage of the 'erroneous' topics together, these sampling errors were kept from slipping through the filter. Mathematically, the following function (F) was used:

$F\ (D) = D_P\ (war)\ /\ D_P\ (travelling + cultural\ events + sports)$[12]

With this formula, we arrived at a relative measure to decide whether an article dealt more with 'war and interventions' than with other topics. For all articles, the quotient was sorted in ascending order to identify blocks of articles with similar characteristics. Looking at these blocks, certain groups of sampling errors could be eliminated. For example, longer French articles with a quotient between 0 and 1.5 were deleted. Articles in this interval did not deal with war and intervention but with other issues. This was applicable to 10 201 articles from the initial total of 70 807 longer French articles.

The margins set were comparatively demanding. Checks confirmed that most sampling errors could be traced and deleted. Very short newsflashes and articles in which related policy issues (e.g. migration, asylum policy, developmental policy issues etc.) are debated while humanitarian military intervention is picked up as a side issue,[13] may therefore sometimes not have survived the cleaning procedure.

[12] $D_P$ = density quotient of one article with regard to the specified semantic field (in brackets) weighted according to the point system.

[13] In previous research projects on transnational European political communication, we found that newspaper articles which deal with the issues of interest as side issue as well as references without further in-depth discussion can be an important indicator for how present a certain issue is as a general point of reference and shared background knowledge (Kantner 2006; Trenz 2004).

## 3. Identifying and removing doublets

Doublets are hard to identify through manual check. Firstly, if the text sample exceeds a certain size, readers will not remember whether they already read a specific text or whether it simply resembles a previously read text. When several analysts are working synchronously, as was the case in the project on war and military intervention in the news, they may not realise that they are annotating, in fact, identical texts. Secondly, two texts may share most contents identified by the analyst, but still each represent an original text, because the contents differ semantically and stylistically, because the second version extends the first, or because intentional quasi-identical re-publication can be considered as a deliberate stressing of a message, which therefore ought not to be deleted from the text corpus ('quasi-doublets'), as it is a marker of the intensity of the debate, which we were interested in. Thirdly, either due to the lax storage practices of the source archives or due to adjustment and repetition of the web query, various versions of articles may enter the data set that differ only with regard to a single word or meta-data categorisation (e.g. regarding sections or authorship).

These characteristics specific to media texts and digital news archives easily escape the reader's eye. They also escape an automatic check for identical documents and often come to light only during the interpretative analysis. Hence, shutting down the 'noise' of doublets requires prior identification of similar versions of an article. Only after having identified these versions and characteristics of their similarity can one develop criteria for deciding upon whether they are 'doublets', which should be discarded, or 'quasi-doublets', which, for one reason or an other, should be kept.

### 3.1 Identifying similar versions of articles

A means of automatically identifying similarities between two documents is to let computers compare the constituent characters of the two texts by implementing a vector space model that yields a numerical similarity value (Manning, Raghavan, and Schütze 2007: 121ff). We applied this procedure to each newspaper separately, because otherwise we would have falsely identified documents as 'doublets' or 'quasi-doublets' that were published in more than one newspapers on the same day. Similar articles printed on the same day in different newspapers are very likely based on prefabricated media messages professionally provided by political institutions or news agencies, e.g. communiqués of international actors or news agency reports. The application of pairwise comparison to *all* characters of all the documents contained in the newspaper corpus on war and military intervention was, however, not feasible given the limited processing power of computers.[14] Therefore, the first step was to create an abstract representation of each document which was less complex and required less calculation effort. An abstract representation that proved effective in practice, i.e. which reduced calculation effort considerably and which nevertheless was significant enough for later comparison, was achieved by (a) setting the entire document set of a newspaper in lower-case letters and (b) extracting from each single document the set of the fifteen least frequent alphabetical characters used in the document set. This method neutralised misleading differences in spelling, punctuation, and capitalisation and reduced the number of characters to be compared, thereby decreasing calculation effort considerably (see Example 1).

**Example 1: Abstract document representation**

> "The first step of the procedure is to identify the 15 least frequent alphabetical characters of the respective lower-case document set (15 turned out to work fine). Based on these characters, abstract representations of all documents are generated by removing all other characters."

Take a look at this paragraph and assume that all characters are lower-case. The 15 least frequent alphabetical characters contained in this paragraph are: k q g v w y m b p u d f l i and h.

Accordingly, the abstract representation of the previous paragraph is:

> "hfipfhpduiidifyhlfqulphbilhfhpivlwdumuduwkfibdhhbpiflldumgdbymvigllhh"

The next step towards pairwise document comparison is to create a vector space, in which each abstract document representation constitutes a vector and, therewith, a measure for assessing the similarity between documents. As dimensions for our vector space, we chose the *5-grams* occurring in the respective set of abstract document representations. A *5-gram* is a unit of 5 adjacent characters; the more general term of

[14] For 100 documents, 4 950 pairwise comparisons have to be calculated, 200 documents already require 19 900 comparisons. As we had up to 100 000 articles to compare for some papers, reduction was essential.

an *n-gram* denotes a unit of *n* adjacent characters with *n* being any desired natural number. We chose *n = 5* because this produced significant results in the subsequent procedure of comparison. From the document representation in Example 1, for instance, sixty-five *5-grams* can be extracted (see Example 2).[15]

**Example 2: n-grams**

From the document representation in Example 1, 65 5-grams can be extracted:

**hfipf**hpduiidifyhlfqulphbilhfhpivlwdumuduwkfibdhhbpiflldumgdbymvigllhh
h**fipfh**pduiidifyhlfqulphbilhfhpivlwdumuduwkfibdhhbpiflldumgdbymvigllhh
hf**ipfhp**duiidifyhlfqulphbilhfhpivlwdumuduwkfibdhhbpiflldumgdbymvigllhh
...
hfipfhpduiidifyhlfqulphbilhfhpivlwdumuduwkfibdhhbpiflldumgdbym**vigll**hh
hfipfhpduiidifyhlfqulphbilhfhpivlwdumuduwkfibdhhbpiflldumgdbymv**igllh**h
hfipfhpduiidifyhlfqulphbilhfhpivlwdumuduwkfibdhhbpiflldumgdbymvi**gllhh**

Out of the possible $15^5$ = 759,375 5-grams for character strings containing 15 characters, 65 occur exactly once in this example.

Each possible n-gram is a 'type' while each actual occurrence is an 'instance' of a type. A multidimensional vector space is constituted by all those types that have instances in the document set of a particular newspaper as the vector space's dimensions. More abstractly, the n-gram types can be denoted as 'features'. Within the vector space, each document can be displayed as a vector (a point) with a certain value for each dimension. The value corresponds to the number of instances of the respective n-gram type in the document (0 or more). The values jointly determine the vector's (and thus the document's) position in the vector space. The similarity between two documents can be calculated by various means of comparison between their vectors. We decided to use the cosine of the angle between the position vectors of two documents as similarity measure because it allows abstracting from the length of a document, using only the proportions of its features: The smaller the angle, the higher the similarity between the two compared documents. The minimal angle amounts to 0°, the maximal to 90°, thus cosine values between 0 and 1 are possible. The value 0 indicates that the two documents are completely distinct with regard to the considered features; while the

[15] The 'n-gram' conventionally applied in computational linguistics is composed of three characters. For instance, a document represented by the characters 'abcabc' contains the 3-grams: 'abc' (twice); 'bca' (once); and 'cab' (once).

value 1 indicates that documents are identical. All features occurring less than twice (single features cannot occur in two different documents) and more than twelve times (features that occur too frequently lack expressiveness) were discarded. All discarded features were deleted from the document representations, which further decreased the necessary calculation effort.

Up to this point, the value for each feature of each document had been the frequency of occurrence in the document, but a more sophisticated type of value was needed to increase expressiveness. The tf-idf-measure (term frequency-inverse document frequency) is defined thus:

*tf-idf = tf * ( |D| / df)*[16]

It expresses the relevance a particular feature (term) has for a document by setting frequency of occurrence in the document in relation with frequency of occurrence in the entire document set (Manning et al. 2007: 117 ff.). The more often the feature occurs in the document and the lower the number of documents sharing the feature, the higher the tf-idf value will be. All feature values were converted to tf-idf. The resulting document representations were compared pairwise using again the cosine of the angle between the position vectors as similarity measure. By this means, we retrieved all documents that were very similar. However, this method was limited to pairwise document comparison and disregarded the fact that multiple documents can easily be each others' duplicates. Hence, all pairs had to be compared with other pairs containing one of both documents. By this method, we identified sets of duplicates ('similarity sets'), sorted by the above similarity measure between 0 and 1.[17]

## 3.2 Distinguishing doublets and quasi-doublets

The similarity sets and the similarity measure were used for manually distinguishing doublets from quasi-doublets. A random sample check of the similarity sets revealed various types of quasi-doublets that were neither identical nor entirely different, but deviated with regard to document length, heading, word order, style, spelling etc. to

[16] Note: tf-idf is the value of the tf-idf-measure, i.e. the term frequency (tf; the number of occurrences of the term in the respective document) multiplied by the number of documents in the set (|D|) divided by the document frequency (df; the number of documents in the set containing the term).

[17] The findings could be displayed with SPSS Clementine, cf. http://www.spss.com/software/modeling/modeler-pro/ (June 5th, 2010).

varying extents. We further identified a margin of similarity for each newspaper separately. However, some articles neither clearly tended towards 'doublets' nor towards 'different articles' (usually between 0.8 and 0.4 on the similarity scale). This was particularly problematic: what should we do about these articles? How to decide whether they were worth keeping or better deleted from the corpus?

A corpus linguistic research project would probably have kept them all because they extend the size of the text corpus and allow the researcher to boost statistical data on language use. Also journalism scholars would have kept them to find out whether and what type of recycling or editing practices had yielded these different versions and whether they originated in the online or print edition of the newspaper. For our content analytical and comparative purposes, however, i.e. for a proper representation of all contents provided by the selected newspapers on war and military intervention and for cross-publication comparison, we had to find a measure that was equally applicable to all the newspapers regardless of from which particular web source the articles had been retrieved and regardless of whether the archives included (non-identifiable) online editions or not.

We decided to only keep the longer version of two articles included in a 'problematic' similarity set – drawing on meta-data on document length – to make sure that the surplus of information was not lost. Moreover, we classified as doublets those articles out of 'problematic similarity sets' which were re-published on the same day, consulting the meta-data on the publication date. This was motivated by the following consideration: re-publications on the same day are likely to be online versions and print versions or morning and evening editions of the same news, whereas re-publication after more than one day is likely to be new 'news' again. By discarding 'problematic' re-publications of the same day, we arrived at similar selection for all the newspapers under investigation. By means of this procedure, altogether 25,736 redundant articles were identified and deleted from the corpus.[18] Hence, the combination of automatic document comparison and the manual check of 'problematic cases' helped to build a cleaned corpus in a transparent manner.

[18] This is, however, not the total of all identified doublets. For most of the newspapers, the above procedure had been preceded by an automatic check for identical articles out of which only one exemplary was kept. This procedure had already reduced the document set of some newspapers to only fifty percent of the initial size. However, this measure turned out to be not fine-graded enough to capture the above problematic cases of similarity.

## 4. Conclusions

This paper expounded two problems that social scientist face when operating with large-n samples of texts retrieved from digital archives: redundancy caused by doublets and bias caused by erroneously sampled documents that may contain the words of the query implemented in the search engines of the digital archive, but do not relate to the topic or genre focussed on in a particular research project. The paper suggested two ways of avoiding these problems and developed procedures for the identification and removal of doublets and sampling errors that innovatively draw on corpus- and computational-linguistic concepts and methods.

Doublets can be identified by combining (a) abstract document representation with (b) automatic pairwise document comparison and (c) manual check of the so-identified ‘problematic cases’. Sampling errors can be identified by calculating the lack of keyness of the individual documents, which is expressed either in the complete absence of keywords representative of the topic investigated or in their under-representation when compared to keywords representative of topics identified as irrelevant to the research in question. Calculation and identification of document-keyness is based on (a) the deductive-inductive identification of semantic fields comprising the keywords of the relevant and the irrelevant topics, (b) their mining in the single documents and relative (weighed) frequency calculation, and (c) manual check of ‘problematic cases’.

Both procedures, to our knowledge, are new in the field of computer-aided textual analysis in the social sciences. Consequently, they need improvement, in particular, through the use of more exact measures and tools for textmining developed in corpus- and computational linguistics. Moreover, our procedures revealed related problems that have, so far, received little attention in social scientists’ approaches to textual analysis: the idiosyncrasies of particular web archives such as LexisNexis and of text genres (news texts) and resulting problems of separating ‘relevant’ and ‘irrelevant’ documents; the (better) design of queries implemented in the search engines of web archives; and the poor inter-operability between different software packages currently available that deal with textmining, statistical calculation or manual annotation. We largely compensated for the latter problem by using our tailored database as a platform, a substantial time-investment.

Finally, the paper showed that computer-aided procedures of corpus cleaning relied to a large extent on intense conceptual work, both regarding the mathematic-informational design of the procedures and the capture of phenomena of matter. They also involved considerable hermeneutic effort, based on careful reading of many texts and profound lexical and grammatical knowledge of the four languages involved. The paper intended to draw attention to the fact that the use of corpus-linguistics and computational linguistics for textual analysis in the social sciences is not a matter of automatisation only, but a truly interdisciplinary endeavour, which requires mutual adaptation of concepts and approaches – a problem we will elaborate in a separate paper on the synergy of content and corpus analysis.

## 5. References:

*Baker, P./McEnery., T.* 2005: A corpus-based approach to discourses of refugees and asylum seekers in UN and newspaper texts, in: Journal of Language and Politics 4/2, 197-226.

*Culpeper, J.* 2009: Keyness: Words, parts-of-speech and semantic categories in the character-talk of Shakespeare's Romeo and Juliet, in: International Journal of Corpus Linguistics 14/1, 29-59.

*Gabrielatos, C.* 2007: Selecting query terms to build a specialised corpus from a restricted-access database, in: ICAME Journal 31, 5-43.

*Kantner, C.* 2006: Die thematische Verschränkung nationaler Öffentlichkeiten in Europa und die Qualität transnationaler politischer Kommunikation, in: Imhof, K./Blum, R./Bonfadelli, H./Jarren, O. (eds.): Demokratie in der Mediengesellschaft, Wiesbaden, 145-160.

*Kantner, C.* 2009: Transnational Identity-Discourse in the Mass Media. Humanitarian Military Interventions and the Emergence of a European Identity (1990-2006). Habilitation Thesis, Freie Universität Berlin.

*Kantner, C./Kutter, A./Renfordt, S.* 2008: The Perception of the EU as an Emerging Security Actor in Media Debates on Humanitarian and Military Interventions (1990-2006). RECON Online Working Paper 2008/19, Oslo.

*Koenig, T./Mihelj, S./Downey, J./Bek, M.G.* 2006: Media framings of the issue of Turkish accession to the EU, in: Innovation: The European Journal of Social Sciences 19/2, 149-169.

*Kutter, A.* 2007: Petitioner or partner? Constructions of European integration in Polish print media debates on the EU Constitutional Treaty, in: Fairclough, N./Cortese, G./Ardizzone, P. (eds.): Discourse and Contemporary Social Change, Bern.

*Liebert, U.* 2007: Introduction: Structuring Political Conflict about Europe: National Media in Transnational Discourse Analysis, in: Perspectives on European Politics and Society 8/3, 236-260.

*Mackey, W.F.* 1965: Language Teaching Analysis. Bloomington, IN.

*Manning, C.D./Raghavan, P./Schütze, H.* 2007: Introduction to Information Retrieval. Cambridge.

*Mike, C.* 2010: Mining a corpus of biographical texts using keywords, in: Lit Linguist Computing 25/1, 23-35.

*Nerlich, B./Clarke, D.D.* 2000: Semantic fields and frames: Historical explorations of the interface between language, action, and cognition, in: Journal of Pragmatics 32/2, 125-150.

*Renfordt, S.* 2009: The Emerging International Law Script in the Media: Evidence from a Longitudinal, Cross-National Analysis of Western Mass Debates about Military Interventions, 1990 to 2005. PhD dissertation, Freie Universität Berlin.

*Scott, M.* 2008: Wordsmith Tools. Liverpool.

*Trenz, H-J.* 2004: Media coverage on European governance. Exploring the European public sphere in national quality newspapers, in: European Journal of Communication 19/3, 291-319.

*Ullmann, S.* 1951: The Principles of Semantics. Glasgow.