FQS FORUM: QUALITATIVE SOCIAL RESEARCH SOZIALFORSCHUNG

Volume 14, No. 2, Art. 23
May 2013

Opening up to Big Data: Computer-Assisted Analysis of Textual Data in Social Sciences

Gregor Wiedemann

Key words: qualitative data analysis; quantitative text analysis; text mining; computer-assisted text analysis; CAQDAS; mixed methods; corpus linguistics; lexicometrics; digital humanities; eHumanities; discourse analysis

Abstract: Two developments in computational text analysis may change the way qualitative data analysis in social sciences is performed: 1. the availability of digital text worth to investigate is growing rapidly, and 2. the improvement of algorithmic information extraction approaches, also called text mining, allows for further bridging the gap between qualitative and quantitative text analysis. The key factor hereby is the inclusion of context into computational linguistic models which extends conventional computational content analysis towards the extraction of meaning. To clarify methodological differences of various computer-assisted text analysis approaches the article suggests a typology from the perspective of a qualitative researcher. This typology shows compatibilities between manual qualitative data analysis methods and computational, rather quantitative approaches for large scale mixed method text analysis designs.

Table of Contents

1. Introduction: Qualitative Data Analysis in a Digital World

Since computer technology became available widespread at universities during the second half of the last century, social science and humanities researchers used it for analyzing huge amounts of textual data. Surprisingly, after 60 years of experience with computer-assisted automatic text analysis and an amazing development in information technology, this is still not a common approach in the social sciences. Nonetheless, an overview of recent developments provided in this article will show that the deployment of (semi-) automatic text analysis technologies is spreading also to fields beyond communication and media studies. At the same time, the underlying algorithmic approaches have made reasonable progress, creating vast opportunities for new mixed method large-scale text analyses. [1]

For some years now computer-assisted text analysis is much more than just counting words. In particular, the combination of statistical and pattern-based approaches of text analysis, referred to as "text mining" may be applied to support established qualitative data analysis designs. In March 2012 *TIME* magazine reported that text mining might be "the next big thing" (BELSKY, 2012). That does not mean it is a very new research area within computer studies. But there is truly much unlocked potential applying recently developed approaches to the tons of digital texts available these days—for economic use cases (as TIME highlights) as well as for various other social science disciplines. [2]

This article introduces an attempt to systematize the existing approaches of computer-assisted text analysis from the perspective of a qualitative researcher. The suggested typology is based not only on the capabilities contemporary computer algorithms provide, but also on their notion of context. The perception of context is essential in a two-fold manner: From a qualitative researcher's perspective it forms the basis for what may be referred to as meaning; and from the computer linguists perspective it is the decisive source to overcome the simple counting of character strings towards more complex models of human language and cognition. Hence, the dealing with context in analysis may act as decisive bridge between qualitative and quantitative research designs. While "real understanding" by computers may remain wishful thinking, nowadays text mining algorithms increasingly include contextual information into their analyses, thus making reasonable progress towards the automatic extraction of "meaning" from text. If open to those new approaches, qualitative social research may profit from that development initiated by technically open-minded scholars more than half a century ago. [3]

1.1 Qualitative analysis and the "digital humanities"

One of the early starters was the Italian theologist Roberto BUSA, who became famous as "pioneer of the digital humanities" for his project "Index Thomasticus" (BONZIO, 2011). Started in 1949—with a sponsorship by IBM—this project digitalized and indexed the complete work of THOMAS AQUINAS and made it publicly available for further research (BUSA, 2004). Another milestone was the software THE GENERAL INQUIRER, developed in the 1960s by communication scientists for the purpose of computer-assisted content analysis of newspapers. It made use of frequency counts of keyword sets to classify documents into given categories. But due to a lack of theoretical foundation and commitment to deductive research designs, emerging qualitative social research remained skeptical about those computer-assisted methods for a long time (KELLE, 2008, p.486). It took until the late 1980s, when personal computers entered the desktops of qualitative researchers, that the first programs called CAQDAS (computer assisted qualitative data analysis software) were created. Since then, a growing variety of software packages, like MAXQDA, ATLAS.ti or NVivo, with relatively sophisticated functionalities, became available, which is making life much easier for qualitative text analysts. Nonetheless, almost all of those software packages have remained "truly qualitative" for a long time by just replicating manual research procedures of coding and memo writing formerly conducted with pens and highlighters, scissors, and glue (KUCKARTZ, 2007, p.16). This once justified methodological skepticism against computational analysis of qualitative data might be a reason for qualitative social research lagging behind in a recent development labeled by the popular catchword "digital humanities" (DH) or "eHumanities." [4]

For some years now the digitalization of the humanities has grown in big steps. Annual conferences are held, institutes and centers for DH are founded and designated chairs have been set up. With CLARIN (Common Language Resources and Technology Infrastructure) the European Union funds a long-term international project (165 million Euros for a period of 10 years) to leverage digital language resources and corresponding technologies. Interestingly, although mission statements of the transnational project and its national counterparts (for Germany CLARIN-D) speak of humanities and social science as their target groups[1], few social scientists have engaged in the project so far. Instead, user communities of philologists, anthropologists, historians and, of course, linguists are dominating the process. In Germany, for example, no single working group for social sciences in CLARIN-D yet exists. This is surprising given the fact that textual data is the primary form of empirical data most qualitatively-oriented social scientists use—even before the linguistic turn hit the discipline. [5]

The branch of qualitative social research devoted to understanding instead of explaining avoided mass data—reasonable in the light of its self-conception as a counterpart to the positivist-quantitative paradigm and scarce analysis resources. But it left a widening gap since the availability of digital textual data, algorithmic complexity and computational capacity is growing exponentially during the last decades. Two humanist scholars highlighted this development in their recent work. Since 2000, the Italian literary scholar Franco MORETTI has developed the idea of "distant reading." To study actual "world literature," which he argues is more than the typical Western canon of some hundred novels, one cannot "close read" all books of interest. Instead, he suggests making use of statistical analysis and graphical visualizations of hundreds of thousands of texts to compare styles and topics from different languages and parts of the world (MORETTI, 2000, 2007). Referring to the Google Books Library Project the American classical philologist Gregory CRANE asked in a famous journal article: "What do you do with a Million Books?" (2006). As possible answer he describes three fundamental applications: digitalization, machine translation and information extraction, to make the information buried in dusty library shelves available to a broader audience. So, how should social scientists respond to these developments? [6]

[1] "CLARIN-D: a web and centres-based research infrastructure for the social sciences and humanities" (http://de.clarin.eu/en/home-en.html [Accessed: January 12, 2013]).

1.2 Digital text and social science research

It is obvious that the growing amount of digital text is of special interest for the social sciences as well. There is not only an ongoing stream of online published newspaper articles, but also corresponding user discussions, Internet forums, blogs and microblogs as well as social networks generate tremendous amounts of text impossible to close read, but worth further investigation. Yet, not only current and future social developments are captured by digital texts. Libraries and publishers worldwide spend a lot of effort retro-digitalizing printed copies of handwritings, newspapers, journals and books. The project Chronicling America by the Library of Congress, for example, scanned and OCR-ed[2] more than one million pages of American newspapers between 1836 and 1922. The Digital Public Library of America strives for making digitally available millions of items like photographs, manuscripts or books from numerous American libraries, archives and museums. German newspaper publishers like the *Frankfurter Allgemeine Zeitung*, *DIE ZEIT* or *DER SPIEGEL* also made all of their volumes published since their founding digitally available. [7]

Interesting as this data for social scientists may be, it becomes clear that single researchers cannot read through all of these materials. Sampling data requires a fair amount of previous knowledge on the topics of interest, which makes especially projects with a long investigation time frame prone to bias. Technologies and methodologies supporting researchers to cope with these mass data problems become more and more important. This is also one outcome of the KWALON Experiment *FQS* conducted in April 2010. For this experiment, different developer teams of software for qualitative data analysis (QDA) were asked to answer the same research questions by analyzing a given corpus of more than one hundred documents on the financial crisis 2008-2009 (e.g. newspaper articles and blog posts) with their product (EVERS, SILVER, MRUCK & PEETERS, 2011). Only one team could include all the textual data in its analysis, because they did not use an approach replicating manual steps of qualitative analysis methods. Instead, they implemented a semi-automatic tool, which combined the automatic retrieval of key words within the text corpus, with a supervised, data-driven dictionary learning algorithm. In an iterated coding process, they "manually" annotated text snippets suggested by the computer, and they simultaneously trained the retrieval algorithm generating the suggestions. This procedure enabled them to process much more data than all other teams, making pre-selections on the corpus unnecessary. However, they only conducted more or less an exploratory analysis which was not able to dig deep into the data (LEJEUNE, 2011). [8]

This article will show that applying and developing further those (semi-) supervised text mining approaches is one big chance to cope better with the trade-off between shallowness and broadness of automatic analyses. Hence these techniques may gain further acceptance within the social science research community by supplementing traditional methods of qualitative research and thus,

[2] OCR – Optical Character Recognition is a technique for the conversion of scanned images of printed text or handwritings into machine-readable character strings.

also address well-known problems of reliability, validity and credibility of their results. In the following section I shortly reflect on methodological aspects of the use of software discussed in qualitative social science research. After that, a typology of four generic concepts is suggested, systematizing how computer-assisted text analyses have already been applied in social science research. In the final section, I give some methodological thoughts on how today's (semi-) automatic text analysis and qualitative research methods may be productively integrated. [9]

2. Computer-Assisted Text Analysis between Quality and Quantity

In the German as well as in the Anglo-Saxon social research community still a deep divide between quantitative and qualitative oriented methods is prominent. This divide can be traced back to several roots, for example the Weberian differentiation between explaining versus understanding as main objectives of scientific activity or the conflict between positivist versus post-positivist research paradigms. During the 1970/80s the latter led to the emergence of several qualitative text analysis methodologies seeking to generate a deep comprehension of a rather small number of cases. Shortcomings of both, qualitative and quantitative approaches for text analysis may be cushioned through integration of the paradigms in mixed method research designs. Analogous to this divide two general tasks in the application of computer-assisted text analysis (CATA) may be distinguished: data processing and data management. [10]

Data processing of large document sets for the purpose of quantitative content analysis framed the early perception of software usage for text analysis. During that time, using computers for qualitative data analysis appeared somehow as retrogression to protagonists of truly qualitative approaches, especially because of their awareness of the history of content analysis. Advantages of CAQDAS programs for *data management* in qualitative analyses (e.g. for documents sets and code categories) have been accepted only gradually since the late 1980s. On the one hand a misunderstanding was widespread, that CAQDAS may be used to analyze text like SPSS is used to analyze numerical data (KELLE, 2011, p.30). Qualitative researchers intended to avoid a reductionist positivist epistemology, which they associated with such methods. On the other hand, it was not seen as an advantage to increase the number of cases in qualitative research designs through the use of computer software. To generate insight into their subject matter researchers should not concentrate on as much cases as possible, but on as much distinct cases as possible. From that point of view using software bears the risk of exchanging creativity and opportunities of serendipity for mechanical processing of some code plans on large document collections (KUCKARTZ, 2007, p.28). Fortunately, the overall dispute for and against software use in qualitative research nowadays is more or less settled. Advantages of CAQDAS for data management are widely accepted throughout the research community. But there is still a lively debate on how software influences the research process —for example through its predetermination of knowledge entities like code hierarchies or linkage possibilities, and under what circumstances quantification may be applied to it. [11]

Interestingly, functions to evaluate quantitative aspects of empirical textual data (like MAXDictio in MAXQDA), have been integrated in all recent versions of the leading analysis software packages. But studies on the usage of CAQDAS indicate that qualitative researchers usually confine themselves to the basic features (KUCKARTZ, 2007, p.28). And for good reason they are reluctant to naively mixing qualitative and quantitative methodological standards of both paradigms—for example, not to draw general conclusions from the distribution of codes annotated in a handful of interviews, if the interviewees have not been selected by representative criteria (SCHÖNFELDER, 2011, §15). Quality criteria well established for quantitative (survey) studies like validity, reliability and objectivity do not translate well for the manifold approaches of qualitative research. The ongoing debate on quality of qualitative research (see for example the *FQS* debate on Quality of Qualitative Research and the respective articles) generally concludes that those criteria have to be reformulated differently. Possible aspects are a systematic method design, traceability of the research process, documentation of intermediate results, permanent self reflection and triangulation (FLICK, 2007). Nonetheless, detractors of qualitative research often see these rather "soft" criteria as a shortcoming of these approaches compared to "hard science" based on numbers. [12]

To overcome shortcomings of both, the qualitative and the quantitative research paradigm, new "mixed method" designs are gradually introduced in QDA. Although the methodological perspectives of quantitative content analysis and qualitative methods, like for instance grounded theory methodology (GTM), are almost oppositional, application of CATA may be fruitful not only as a tool for exploration and heuristics. However, Udo KUCKARTZ states: "Concerning the analysis of qualitative data, techniques of computer-assisted quantitative content analysis are up to now widely ignored" (2010, p.219; my translation). This perspective suggests that qualitative and quantitative approaches of text analysis should not be considered as competing, but as complementing techniques. They enable us to answer different questions on the same subject matter. While a qualitative view may help us to understand what categories of interest in the data exist and how they are constructed, quantitative analysis may tell us something about the relevance, variety and development of those categories. Hence, I fully agree with KUCKARTZ advertising the advantages a quantitative perspective on text may contribute to an understanding—especially to integrate micro studies on text with a macro perspective. [13]

But a closer look reveals that KUCKARTZ's statement as well as many other "mixed method" descriptions lack of a fair consideration of current quantitative text analysis approaches. Their focus on computational content analysis (CCA) and simple "term based analysis functions" (p.218) reflects a limited comprehension of contemporary CATA approaches. Conventional content analysis (CA) is spurned for reason in the QDA community. Already in 1952, Siegfried KRACAUER criticized quantifying content analysis for its limitations:

reduced accuracy due to neglect of qualitative exploration, and preclusion of judicious appraisal of bias emerging from qualitative aspects of its categories. As a result, qualitative approaches to content analysis were strengthened in later decades (see Section 3). In contrast CCA, adhered to the quantitative paradigm and restricted by computational and algorithmic capacities of that time, largely failed to address this critique up to now. [14]

Interestingly, two recent developments of computer-assisted text analysis may severely change the circumstances which in the past have had been serious obstacles for a fruitful integration of qualitative and quantitative QDA. Firstly, the availability and processability of full-text archives (e.g. all articles of a specific newspaper from 1985-2005 or all twitter postings from November 2012) enables researchers to properly combine methodological standards of both paradigms. Instead of a potentially biased manual selection of a small sample (n < 100) from the population of all documents, a statistical representative subset (n > 1,000) may be drawn, or even the full corpus (n > 100,000) may be analyzed. Secondly, the epistemological gap between how qualitative researchers perceive their object of research compared to what computer algorithms are able to identify is constantly narrowing. The key factor hereby is the algorithmic extraction of "meaning" which is approached by the inclusion of "context" into the applied computational linguistic models of analysis. [15]

3. From Context to Meaning: A Typology of Computer-Assisted Text Analyses

In the literature on computer-assisted text analysis, several typologies of existing approaches can be found. The aim of this exercise usually is to draw clear distinctions between capabilities and purposes of software technologies and to give guidance for possible research designs. By the very nature of the matter it is obvious that these typologies have short half-life periods due to the ongoing technological progress. KRIPPENDORFF for example suggests in a famous text book on content analysis the differentiation of three types of computer-assisted text analysis: 1. retrieval functions for character strings on raw text, 2. computational content analysis and 3. CAQDAS. Although published recently in its 3rd edition (2013), it largely ignores the developments of the last decade by not covering approaches of statistical/linguistic text mining. Another distinction, analogue to the first two types mentioned, dates back to the Annenberg conference on content analysis at the end of the 1960s. There CA methods were divided into exploration of term frequencies and concordances without theoretical guidance on the one hand and hypothesis guided categorizations with dictionaries on the other (STONE, 1997). In contrast, newer approaches that consider current algorithmic capabilities differentiate into 1. dictionary based CCA, 2. parsing approaches to CCA and 3. contextual similarity measures (LOWE, 2003). The latest suggestion from SCHARKOW (2012) distinguishes three dimensions of computational text analysis: 1. unsupervised vs. supervised approaches; and within the supervised ones 2. statistical vs. linguistic and 3. deductive vs. inductive approaches (see Section 3.4). Unquestionably, this classification covers important characteristics of CATA software currently in use.

But one could easily find more dimensions to distinguish like intra-textual vs. trans-textual or subsumptive vs. extractive approaches. In the following sections, I will explain characteristics and differences of these different approaches. [16]

To not to make it too complicated and having in mind that not content analysts but researchers from a qualitative, more reconstructivist background should be intrigued to use computer-assisted text analysis, I highlight one specific dimension. This typology marks what progress has been made from observation of document surfaces on a simple term level to more complex semantic structures seeking to extract "meaning" from document collections. I argue that if we imagine a one-dimensional space between deep understanding, e.g. qualitative data analysis through hermeneutic or reconstructive methods, and superficial observation, e.g. quantitative analysis by just counting frequencies of terms (or character strings) in digital text, nowadays approaches of text mining lie somewhat in between both ends of this spectrum. The more they enable us to extract "meaning" by keeping their capacity to be applied to mass textual data, the more they may truly contribute to the integration of qualitative and quantitative text analysis. [17]

What KRACAUER (1952) criticized in the mid-20th century was the methodological neglect of substantial meaning in quantitative content analysis. Content analysis, especially its computer-assisted version, observed the occurrence of specific sets of terms within its analysis objects, but systematically ignored its contexts. To generate understanding out of the analysis objects in favor to gain new insights, counting words did not prove as adequate to satisfy deeper research interests. In this respect, upcoming methods of qualitative content analysis were not conceptualized to substitute its quantitative counterparts, but to provide a systematic method for scientific rule-based interpretation. One essential characteristic of these methods is the embedded inspection and interpretation of the material of analysis within its communication contexts (MAYRING, 2010, p.48). Thus, the systematic inclusion and interpretation of contexts in analysis procedures is essential to advance from superficial counts of character strings in text corpora to the extraction of meaning from text. [18]

Since the linguistic turn took effect in social science (BERGMANN, 1952), it is widely accepted that structures of meaning are never fully fixed or closed. Instead, they underlie a permanent evolvement through every speech act which leaves its traces within the communicative network of texts of a society. Hence, meaning can be inferred only through the joint observation of the differential relations of linguistic structures in actual language use and it always stays preliminary knowledge (TEUBERT, 2006). For CATA this can be translated into the observation of networks of simple lexical or more complex linguistic units within digitalized speech. The underlying assumption is that structures of meaning evolve from the interplay of these units, measurable for example in large text collections. [19]

Luckily, identifying structures in digital data is one major strength of computers. Nonetheless, we have to narrow down what we consider as context for our analysis. Usually, in computer-assisted text analysis we are constrained to linguistic contexts of sentences, paragraphs, documents or document collections to infer our knowledge from. Hence, situational contexts (e.g. biographical information about the author of a text) may be integrated in CATA only indirectly, mostly through assumptions about condensation of such information in specific language structures. To a certain degree, today's text mining algorithms also may integrate text-external, structured knowledge through specific statistical models (see Section 3.4). One can see easily the advantages context-aware CATA approaches have over such which just recognize isolated patterns in digital text:

> "The elementary 'statistics of text' show that reference to frequency counts alone is a bad idea. If we consider a contingency table showing the number of times a given word c occurs one word to the left of a target word t, the variation in the frequency of co-occurrence will be driven by the marginal frequency of the target word as well as by its true level of association with c. It is the association between c and t that is important in quantifying a context, not just the number of times they share one" (BRIER & HOPP, 2011, p.106). [20]

Taking into account those basic linguistic principles appears as one necessary but not trivial condition for CATA methods to be truly beneficial for more "qualitatively" oriented research. It is essential for their capacities to identify patterns of language use in an inductive or abductive manner which may be of value within research designs primarily guided to deepening comprehension. [21]

One further important distinction is if these methods solely rely on the observation of overt variables or if they dig down into "latent meaning" by applying various statistics on the textual material. "Latent meaning" may be computed as non-observable variables by statistical dimension reduction on observable data. Those methods detach the analysis from the retrieval of fixed linguistic patterns like single key terms to complex semantic relations. This leads us to a typology of four types of CATA approaches:

1. method independent software that provides tools for *manual coding* processes (CAQDAS) making allowance for linguistic and situational contexts;
2. hypothesis-driven computational content analysis (CCA) yielding automatically annotated texts through *observation* of term occurrences while largely ignoring contexts;
3. data-driven lexicometrics and corpus linguistic methods allowing inductive *exploration* of language patterns by measuring overt contexts of linguistic symbols;
4. text mining approaches which strive for *extraction* of "meaning" through application of complex statistical models calculating latent contexts of linguistic symbols. [22]

In the following section I will explain characteristics of these types more detailed and give examples of social science studies applying those kinds of methods. [23]

3.1 CAQDAS: Context-comprehensive manual coding

As SCHÖNFELDER states it, "qualitative analysis at its very core can be condensed to a close and repeated review of data, categorizing, interpreting and writing" (2011, §29). To support these manual tasks of qualitative data analysis, software packages like Ethnograph, MAXQDA, NVivo or ATLAS.ti have been developed since the 1980s. They provide functions for data/document management, development of code hierarchies, annotation of text segments with codes, writing memos, exploring data and text retrieval as well as visual representations of data annotations. The major characteristic of this class of CAQDAS is that

> "none of these steps can be conducted with an algorithm alone. In other words, at each step the role of the computer remains restricted to an intelligent archiving ('code-and-retrieve') system, the analysis itself is always done by a human interpreter" (KELLE, 1997, §5.7). [24]

Mostly CAQDAS packages are relatively flexible concerning the research methods they are used with. Early versions developed, usually had concrete methodologies in mind which should be mapped onto a program-guided process. Data representations and analysis functionalities in ATLAS.ti for example were mainly replicating concepts known from grounded theory methodology (MÜHLMEYER-MENTZEL, 2011). Later on, while the packages matured and integrated more and more functions, they lost their strict relations to specific qualitative methods. Although differences are marginal, debates on which software suits which method best[3] persist in the qualitative research community (e.g. KUŞ SAILLARD, 2011). Nonetheless the use of CAQDAS is nowadays widely accepted. Anxious debates from the 1980s and early 1990s, if or how computers affect qualitative research negatively *per se*, have been settled. Already mid of the 1990s a study by FIELDING and LEE suggested

> "that users tend to cease the use of a specific software rather than adopt their own analysis strategy to that specific software. There seem to be good reasons to assume that researchers are primarily guided by their research objectives and analysis strategies, and not by the software they use" (KELLE, 1997, §2.9). [25]

The KWALON experiment (see Section 1.2) largely confirmed that assumption. The experiment sought to investigate the influence of CAQDAS on research results in a laboratory research design (same data, same questions, but different software packages and research teams). Regarding the results FRIESE (2011) concluded that the influence of software on the research process is more limited when the user has fundamental knowledge of the method he/she applies. Conversely, if the user has little methodological expertise, he/she is more prone

[3] The University of Surrey provides an useful overview of CAQDAS packages on its website.

to predefined concepts the software advertises. To deal with those pitfalls, scholars interested in qualitative research may be trained in using CAQDAS packages more regularly (MÜHLMEYER-MENTZEL & SCHÜRMANN, 2011). [26]

Taking context of analysis objects into account when using CAQDAS is not determined by the program, but by the applied method. Due to its focus on support of various manual analysis steps it is flexible in methodological regard. Situational contexts like historic circumstances during times of origin of the investigated texts may be easily integrated into the analysis structure through memo functions or linkages with other texts. Linguistic contexts of the entities of interest are part of the analysis simply because of the qualitative nature of the research process itself. However, this kind of CATA limits the researcher to a narrow corpus. Although CAQDAS may increase transparency and traceability of the research process, as well as possibilities for teamwork in research groups, it does not dissolve problems of quality assurance of qualitative research directly related to the rather small number of cases investigated. Analyzing larger, more representative amounts of text to generate more valid results and dealing with the problem of reliability in the codification process is the objective of the other types of CATA, strongly incorporating a quantitative perspective on the qualitative data. [27]

3.2 Computational content analysis: Context-neglecting automatic coding

Quantitative approaches of content analysis have a long history, especially in media studies. As a classic deductive research design CA aims at a data-reducing description of mass textual data by assigning categories on textual entities like newspaper articles, speeches, press releases etc. The set of categories, the code hierarchy, usually is developed by domain experts on the basis of pre-existing knowledge and utilized for hypothesis testing of assumptions on the quantitative development of code frequencies in the data. Categories may be assigned on several dimensions, like occasion of a topic (e.g. mentioning ethical, social or environmental standards in business reports), its share of an analyzed text (once mentioned, higher share or full article) or its valuation and intensity (e.g. overall/mainly pro, contra or neutral). Codebooks explain these categories in detail and give examples to enable trained coders to conduct the data collection of the study "manually" by close reading. Following a rather nomothetic research paradigm, CA is described by KRIPPENDORFF as "a research technique for making replicable and valid inferences from texts [...] to the contexts of their use" (2013, p.24). Thus, replicability should be achieved, among other things, through inter- and intracoder-reliability—two metrics which calculate the matches of code assignments between several coders or the same coder on repeated coding processes. [28]

Automatic CCA has to operationalize its categories in a different way. Already in 1955, a big conference on CA marked two main trends in the evolvement of the method: 1. the shift from analysis of contents to broader contexts and conditions of communication which led to more qualitative CA, and 2. counting of symbol frequencies and co-occurrences instead of counting subject matters (p.19). The latter strand paved the way for the overly successful CCA software THE

GENERAL INQUIRER during the 1960s. While neglecting implicit meaning, thus concentrating on linguistic surfaces, CCA simply observed character string occurrences and their combinations in digital textual data. Researchers therefore create lists of terms, called dictionaries, describing the categories of interest. Computers then process hundreds of thousands of documents looking for those category-defining terms and in case of detection, assign the given label to them. The process can be fine tuned by expanding or narrowing the dictionary, applying pattern rules (observation of one, several or all category-defining terms, 1...n times). Counting the labels in the end allows making assertions on the quantitative development of the overall subject-matter. Thus, developing the dictionaries became the main task of the research process in a CCA designs. [29]

In social science research the method is applicable when large corpora of qualitative data need to be analyzed. ZÜLL and MOHLER (2001) for example used the method to summarize open questions of a survey study on the perception of aspects of life in the former GDR. Another big research project evaluated tens of thousands of forum postings of a public campaign on bio ethics in Germany (TAMAYO KORTE, WALDSCHMIDT, DALMAN-EKEN & KLEIN, 2007). The project is interesting insofar it embeds CCA in a framework of discourse analysis. The development of the categories of interest was conducted in an abductive manner: from observed lexical units underlying discourse and knowledge structures were inferred inductively. These structures, operationalized as dictionaries in MAXDictio, then again are tested as hypothesis against the empirical data. The project shows that CCA is not constrained to a pure nomothetic research paradigm. [30]

Nonetheless, because of serious methodical concessions CCA is comprised with several obstacles. Researchers need a detailed comprehension of their subject matter to construct dictionaries which deliver valid results. If not developed abductively, their categories need to "coincide well with those of the author" of the analyzed document (LOWE, 2003, p.11). In fact, a lot of effort has been made during last decades by exponents of CCA to develop generic dictionaries applicable to various research projects. The project Linguistic Inquiry and Word Count, for example, provides dictionaries for linguistic and psychological processes like swear words, positive emotions or religion related vocabulary. But, having the above-mentioned constraint in mind, experience has demonstrated that these general dictionaries alone are of little use for generating insights in qualitative data analysis. Although often freely available, dictionaries were almost never re-used outside the research projects for which they were developed originally (SCHARKOW, 2012, p.79). Furthermore, studies comparing different versions of the same translated texts from one language into the other had shown that vocabulary lists of single terms are not necessarily a good indicator for similar content (KRIPPENDORFF, 2013, p.239). The deterministic algorithmic processing of text guarantees best reliability (identical input generates identical output), but poor validity due to incomplete dictionaries, synonyms, homonyms, misspellings and neglect of dynamic language developments. Hence, CCA bears the risk to "end up claiming unwarranted generalizations tied to single words, one word at a time" (p.264). The systematic omission of contexts limits the method to "very superficial meanings" with a tendency to "follow in the footsteps of behaviourist assumptions" (ibid.). [31]

3.3 Lexicometrics/corpus linguistics: Context-observing content exploration

As a critical reaction to nomothetic, deductive and behaviorist views on social research with linguistic data, notably in France the emergence of (post-) structuralism had sustainable impact on computational text analysis. In 1969 the historian Michel PÊCHEUX published his work "Analyse automatique du discours" (AAD) which attracted much attention in the Francophone world, but remained largely ignored in the English speaking world due to the fact that till 1995 no translation existed (HELSLOOT & HAK, 2007, §3). While the technical capacities of computational textual analysis did not allow realizing his ideas during that time, AAD was conceptualized as a theoretical work. PÊCHEUX generally accepted the need of analyzing large volumes of text for empirical research, but rejected the methods of CCA, because of the ideological distortions by naively applying dictionary categories onto the data:

> "Given the volume of material to be processed, the implementation of these analyses is in fact dependent upon the automatization of the recording of the discursive surface. In my view, there is no alternative, and any preliminary or arbitrary reduction of surface [...] by means of techniques of the 'code résume' type is to be avoided because it presupposes a knowledge of the very result we are trying to obtain [...]" (PÊCHEUX, HAK & HELSLOOT, 1995, p.121). [32]

With SAUSSUREs distinction of signifier and signified he argues that discourse has to be studied by observing language within its contexts of production and its use with as little pre-assumptions as possible. Approaches which just count predefined symbol frequencies assigned to categories suffer from the underlying (false) assumption of a bi-unique relation between signifier and signified—thus are considered as "pre-Saussurean" (PÊCHEUX et al., 1995, p.65). Meaning instead is "an effect of metaphoric relations (of selection and substitution) which are specific for (the conditions of production of) an utterance or a text" (HELSLOOT & HAK, 2007, §25). In the 1970s and following decades, AAD was developed further as a theoretical framework of discourse study as well as an empirical tool to analyze texts. This class of text analysis tools is often labeled lexicometrics. [33]

Lexicometric approaches in discourse studies aim to identify major semantic structures inductively in digital text collections. Linguists apply lexicometric measures in the field of corpus linguistics to quantify linguistic data for further statistical analysis. Other social scientists who are interested in analyzing texts for their research adapted these methods to their needs and methodologies. DZUDZEK, GLASZE, MATTISSEK and SCHIRMEL (2009) mention four fundamental methods of lexicometrics: 1. frequency analysis for every term of the vocabulary of the collection to identify important terms, 2. concordance analysis to examine local contexts of terms of interest (results usually are returned as keyword-in-context, so-called KWIC-lists, which display n words to the left and to the right of each occurrence of an examined key term), 3. identification/measuring of characteristics of sub corpora which are selected by meaningful criteria (e.g. different authors, time frames etc.), and finally 4. co-occurrence analysis to examine significant contexts of terms on a global (collection) level. Significance thereby is measured with a statistical test showing which terms occur together more frequently within the corpus than simply by random chance. Multivariate methods may complement these techniques e.g. to identify clusters of co-occurring terms or measure their "keyness," the importance of specific terms for a given document (the more sophisticated these methods get, the more they may be assigned to the fourth category of this typology called "text mining"). [34]

In contrast to CCA, where development of categories, category markers, code plans etc. takes place before the automated analysis, the interpretive part of lexicometric text analysis is conducted after the computational part. First, quantitative relations between lexical units are computed in a purely data-driven manner from a carefully selected document corpus. Although computation itself is data-driven and thus, not prone to research bias, the selection of corpus documents of course is susceptible to it, as well as parameter settings and threshold values of the algorithms. However, these adjusting screws are essential to consciously control the process and fit it to the researchers needs. Then, the computed results are examined further and interpreted in the light of the research question (DZUDZEK et al., 2009, p.234). Compared to CCA, the exchange of these steps in the research process allows that the researcher even has a chance to develop an understanding of how meaning is constructed in the empirical data. This makes these tools compatible with a range of poststructuralist methodological approaches of text analysis like (Foucauldian) discourse analysis, historical semantics, grounded theory methodology, or frame analysis. Especially in France (and other French speaking countries) discourse studies combining interpretive, hermeneutic approaches with lexicometric techniques are quite common (GUILHAUMOU, 2008). [35]

In the Anglo-Saxon and German-speaking qualitative research community the methodical current of critical discourse analysis (CDA) developed a branch which incorporates lexicometric methods of corpus linguistics successfully into its analysis repertoire:

> "The corpus linguistic approach allows the researcher to work with enormous amounts of data and yet get a close-up on linguistic detail: a 'best-of-both-worlds' scenario hardly achievable through the use of purely qualitative CDA, pragmatics, ethnography or systemic functional analysis" (MAUTNER, 2009, p.125). [36]

In a lexicometric CDA study of the discourse about refugees and asylum seekers in the UK the authors conclude on their mixed method:

> "Importantly, the project demonstrated the fuzzy boundaries between 'quantitative' and 'qualitative' approaches. More specifically, it showed that 'qualitative' findings can

> be quantified, and that 'quantitative' findings need to be interpreted in the light of existing theories, and lead to their adaptation, or the formulation of new ones" (BAKER et al., 2008, p.296). [37]

More with linguistic than with social science interest the works of the German semtracks research group applied methods of corpus linguistics incorporated into a methodology of discourse analysis. Noah BUBENHOFER (2009) sketched a framework of purely data-driven corpus linguistic discourse analysis which seeks to identify typical repetitive patterns of language use in texts. These patterns of significant co-occurrences provide the basis for intersubjectively shared knowledge or discursive narratives within a community of speakers. For political scientists of special interest is the project PolMine by the University of Duisburg-Essen which makes protocols of German federal and state parliaments digitally available and also provides corpus linguistic/lexicometric analysis functions. In a first exploratory study, Andreas BLÄTTE (2012) investigated empirically overlapping and delimitation of policy fields with these data and compared his findings with theoretical assumptions on policy fields in political science literature. For a study of the (post-) colonial discourse in France Georg GLASZE (2007) suggested a procedure to operationalize the discourse theory of Ernesto LACLAU and Chantal MOUFFE by combining interpretive and lexicometric methods. [38]

Although these examples show that lexicometric approaches gain ground in the analysis of qualitative data, they have been largely ignored over a long time and still are not very common outside the Francophone world.[4] Besides the fact, that no methodological standard yet exists, these methods require a certain amount of technical skills, which excludes quite a bit of social scientists not willing to dive into this topic. Yet, lexicometric approaches are quite flexible to be integrated into different research designs and are compatible with epistemological foundations of well-established manual qualitative data analysis approaches. Methodologically, lexicometrics and corpus linguistics differ from the most manual qualitative methods in how they handle their text corpus. Qualitative methods often investigate open corpora. Whenever the researcher has found new interesting material or has the assumption that his/her data already analyzed does not cover the topic completely, he/she is able to extend the collection. In contrast, corpus linguistics analyzes closed corpora—means a fixed set of documents is necessary to make the results of text statistical analysis comparable. When applying these methods, researchers may work around this problem by selecting different sub corpora, thus, slightly dissolving this problem. [39]

Overall, the application of lexicometrics is of medium complexity. Some matured software packages exist, allowing its use for the technically interested social scientists without any help of computer linguistic experts.[5] In contrast to CCA,

[4] For example, the bi-annual conference "Journées internationales d’analyse statistique des données textuelles" (JADT) is relatively well-known in the Francophone world, but only recently opens up to participants who prefer English as primary language for scientific exchange. Another hint is that estimated ¾ of lexicometric software products I know were developed by Francophone research teams.

[5] Popular programs are for example Alceste, WordSmith or TextQuest.

lexicometric approaches preserve linguistic contexts of the observed lexical units to a certain degree and thus allow investigation of the constitution of their meaning as well as their evolvement. But the notion of context may be further extended for more sophisticated (semi-) automatic text analyses. [40]

3.4 Text mining: Pattern- and model-based latent context calculation

The process of extracting knowledge represented and expressed within text is achieved by human readers intuitively. It can be seen as a process of structuring, by identifying relevant textual fragments, collecting and assigning them to newly created or predefined concepts in a specific field of knowledge. Accordingly, text mining can be defined as a set of "computer based methods for a semantic analysis of text that help to automatically, or semi-automatically, structure text, particular very large amounts of text" (HEYER, 2009, p.2). [41]

Until the 1980s computer scientists and linguists tried to reproduce the rules of human language in a structuralist manner—and largely failed. The structure of human language turned out too complex and too dynamic to be represented by first-order logic and hand-written rules. Thus, during the 1980/90s statistical approaches to natural language processing (NLP) became popular and much more successful (SAMUELSSON, 2004, p.358). While it is beyond the scope of this article to explain details of text mining, I can give only some basic ideas of how computational extraction of semantic knowledge is achieved. In NLP corpora of actual human originated text, spoken or written, are the basis for identifying structures by applying statistical methods. This requires a fundamentally different view on text in contrast to what qualitative oriented researchers are used to. For most approaches, text has to be transformed into numbers—eventually, it has to be handled as vectors and matrices. For example, you can count the occurrence of every word (token) in a document and thus represent it by a mathematical vector—an ordered list of summed up occurrences of each unique word form (type). Hence, a set of documents may be represented as a set of vectors, or as mathematical matrix. Text mining algorithms now combine elaborated statistical methods on those matrices with knowledge about statistical characteristics of language and text-external knowledge manually coded by researchers (e.g. categories or example sets). Machine learning (ML) algorithms applied to those data may, for example, infer rule sets or statistical probabilities of typical characteristics from hand coded input texts, thereby "learning" to retrieve or annotate information in unknown material. If an algorithm uses for its analysis just the textual data itself, without interference of external data or human control, it is called "unsupervised." You may think of a cluster algorithm grouping your document set in k different clusters each containing similar documents but as distinctive as possible to the documents of another cluster. In contrast, an algorithm is called "supervised" if it integrates external information or its intermediate results are controlled and evaluated by analysts during processing. Here, in contrast to unsupervised clustering, you may have a given set of categories and some documents labeled with them. From this "training set," the machine-learning algorithm may learn features to classify new unlabeled documents. In combination with pattern based approaches, powerful visualizations and user-friendly browsers those algorithms are capable to extend traditional qualitative research designs and open them up to large document collections. [42]

In contrast to CCA or lexicometrics, researchers are not obliged to restrict their analysis to single lexical units when using text mining. The representation of documents as vectors and document sets as matrices allows the preservation of linguistic contexts to a large extent. Context hereby is not only co-text, defined as a rather small snippet of some terms surrounding a lexical unit. Instead, context may be a sentence, a complete document or even the entire corpus. Furthermore various kinds of external data might be included into the analysis—like time indices of documents allowing for the data-driven identification of evolvement-patterns of linguistic data, or text snippets manually annotated with information of interest like category labels, sentiment or valence scales. Depending on the analysis of interest, the data to be included as well as the type of results to be produced determines the selection of a suitable algorithm or statistical model. In contrast to corpus linguistic methods, many text mining approaches do not rely on closed corpora. Instead, they may be applied to dynamic sets of input documents or to continuous flows of input streams. This enables researchers not to restrict themselves on fixed document sets. Instead they may incorporate new qualitative data, if at one point of the research process it seems suitable. [43]

In general, one may distinguish in tasks of clustering, classification and information extraction of texts which might be applied to social science research interests in different ways. [44]

Classification of documents into a given set of categories is a standard application of media and content analysis. Using a supervised support vector machine (SVM) classification approach, Michael SCHARKOW (2012) has shown that for a relatively simple code set of news-article types (training sets annotated with "politics," "economy," "sports," etc.) the automatic classification achieves accuracy up to 90 percent of correct document annotations—a pretty good value for a machine learning approach, although better machine learners than SVMs already exist. But even if classification accuracy is below 90% (which it is often), results may be useful for social scientists. HOPKINS and KING point to the fact, that social scientists are not primarily interested in correct classification of single documents. Instead they want to infer generalization on the whole document set like proportions of the identified categories—which introduces new problems:

> "Unfortunately, even a method with a high percent of individual documents correctly classified can be hugely biased when estimating category proportions. By directly optimizing for this social science goal, we develop a method that gives approximately unbiased estimates of category proportions even when the optimal classifier performs poorly" (2010, p.229). [45]

With their approach they measured the sentiments (five classes ranging from extremely negative to extremely positive) on more than 10,000 blog posts on the candidates of the US-American presidential election in 2008. Therefore only 442 posts were read and hand coded by researchers. They then were used as a training set for the ML algorithm which classified the remaining posts. In another project, philologists classified newspaper articles from a complete time indexed corpus of the German magazine *DIE ZEIT* between 1949 and 2011 by applying a relatively sophisticated dictionary approach. Using selected parts of an onomasiological dictionary they identified and annotated the mentioning of tropic frames (e.g. health, criminality, family, virtue and order) in more than 400,000 articles. The increases and decreases, as well as the co-occurrences of these frames over time give some interesting insights (SCHARLOTH, EUGSTER & BUBENHOFER, 2013): Their method reveals long-term developments in societal meta-discourses in Germany independently from the close observation of specific societal events which could not have been shown by solely qualitative analysis. To support a qualitative study about a small Finnish coffee firm, JANASIK, HONKELA and BRUUN (2009) employed an unsupervised clustering approach with self organizing maps (SOM). With the help of SOMs they visually arranged their interview data by textual similarity on a two-dimensional map to disclose the topological structure of the data and infer data-driven "real types" (in contrast to theory-led "ideal types") of their interviewees. Interestingly, they argue for parallels of their approach with grounded theory methodology (pp.436f.). [46]

Also for *information extraction* some interesting case studies can be found in the literature. ADAMS and ROSCIGNO (2005) applied the commercial text mining tool TextAnalyst on a corpus of website documents from US neo-Nazis and Ku-Klux-Klan chapters to investigate identity patterns of both groups. The software, usually used in applied information science, creates semantic networks on the basis of automatic extraction of relevant terms and their co-occurrences. These networks represent knowledge structures on a transtextual level, giving insight into how the ideologies of both groups are constructed and in what way they differentiate or share ideas. Another two-class-divided document set is investigated with two rule learning and one decision tree learning algorithm in a study of POLLAK, COESEMANS, DAELEMANS and LAVRAC (2011). They strive to learn about the local and the international media discourse on the topic of Kenyan elections in 2008. The results of their automatic text analysis represent text features which are most distinctive for both classes. Their interpretation allows interesting insights into the differences of Kenyan news framing and its reception in the Anglo-Saxon world. [47]

One last, but most promising approach for information extraction to mention here are topic models. Topic models are an approach to identify global co-occurrence structures within text corpora. These structures are assigned to a given number of (previously unknown) categories, representing semantic connections which may be interpreted as topics (BLEI, NG & JORDAN, 2003). Topic models may be applied for a variety of further analysis like term extraction, topic evolution over time as a data-driven operationalization of discourse theory or for retrieval of similar documents. The way topic models may change our access to large document collections is well described by its developer David BLEI—somehow his description resembles an empiricist reformulation of a theoretical discourse comprehension:

> "Imagine searching and exploring documents based on the themes that run through them. We might 'zoom in' and 'zoom out' to find specific or broader themes; we might look at how those themes changed through time or how they are connected to each other. Rather than finding documents through keyword search alone, we might first find the theme that we are interested in, and then examine the documents related to that theme" (2012, p.77). [48]

One study showing the potential of topic models for social sciences has been conducted by the political scientist Justin GRIMMER (2010) who calculated topic proportions of more than 25,000 press releases from members of the US Congress and correlated the findings with text external data like partisanship and rural vs. urban election districts. Unfortunately, text mining approaches in general and topic models in particular can get relatively complex to handle. So far, their application is an undertaking of dual-disciplinary nerds or a joint cooperation of computer linguists and social scientists in larger projects. [49]

4. Large-Scale QDA as Mixed Method Text Analysis

The examples above show, that computer-assisted analysis of qualitative data may become a very complex venture. From simple counts of occurrences of character strings in single documents to complex statistical models with latent variables over huge document collections a long road has been traveled. The complexity of the methods just mirrors the complexity of natural language itself. Still, these methods are based on very simplified models of how natural language takes effect cognitively (on a micro level) and socially (on a macro level). Nonetheless, today's methods come quite a bit closer to the aim of extracting meaning from text—the basis for understanding as meta objective of qualitative research. In contrast to computer-assisted manual methods with CAQDAS, a quantitative perspective on the data necessarily has to be taken into account. Which knowledge by the use of language is expressed within a concrete speech act can only be understood by comparing it to a large set of other linguistic data. Manual QDA relates on expert and world knowledge of the researcher for that (implicitly quantified through the assumption of its relevance), whereas computer-assisted (semi-) automatic methods need a lot of qualitative data, incorporating quantities explicitly. Thus, analyzing big data in QDA only makes sense as mixed method text analysis. [50]

The typology suggested above also shows that CATA is quite flexible not only in terms of methodological compatibility but in its procedures as well. Thereby, it is not primarily decisive whether research designs use them inductively or deductively, for corpus-driven hypothesis testing or data-driven pattern identification, for data exploration or explanation. Numerous analysis techniques of the textual data may be combined, depending on the research interest. Far more decisive is to develop an understanding how the research process is guided by the CATA approach and how the analysis may be controlled by the researcher. In contrast to purely automatic coding of CCA computational approaches like lexicometrics and applications of text mining allow for inductive data-driven and semi-automatic analysis procedures. Hereby, the combination of supervised learning procedures and automatic codification in a semi-automatic approach is very promising. Christophe LEJEUNE from the University of Liége, for example, has built the software Cassandre to annotate texts with qualitative categories defined by textual markers (2011). The software deduces the lexical features determining a category while the researcher qualitatively annotates them. Thus, a semi-automatic annotation process of large data sets becomes feasible without losing the control or direct connection to the empirical data—as he puts it: combining the best from "automatic" and "reflective coding" (§10). [51]

For qualitatively oriented social science research this is essential: the *semi*-automatic analysis or the supervised text mining process is the sticking point where computer-assisted text analysis grows from naïve word counting, which does not help much to attain useful cognition, to a tool enabling the researcher to answer his/her questions in a new powerful, controlled way guided by theoretical or empirical foundations. Well, it seems obvious that computers will not be able to really understand texts in ways reconstructivist social scientists strive for. Algorithms may deploy only little contextual knowledge from outside the text they shall analyze, compared to the experience and common sense knowledge a human analysts can rely on. Thus, the extraction of "latent meaning" in the sense qualitative hermeneutic methods aim at, is truly not within the scope of automatic text analysis. Reconstructive methods like objective hermeneutics which produce a lot of material on the basis of rather short text excerpts combined with the world knowledge of the researcher may not profit directly from automatic text analysis. But text-reducing methods, like various approaches of discourse analyses, which operate on a transtextual level do have a good chance to benefit from computer-assisted methods if they are not shy of quantification; by the way, quite a commonplace in France—birthplace of postmodern discourse analysis—, where the qualitative-constructivist vs. the realist-positivist divide never took that much effect (consequently, sometimes discourse analysis is labeled as quasi-qualitative method; ANGERMÜLLER, 2005). Hence, the conceptual differences of the distinctive types of CATA have to be made clearer in the discussion on research methods. If, for example, method experts elaborate and highlight explicitly their different underlying epistemologies and their compatibilities with methods like GTM, CDA or qualitative content analysis, acceptance for new mixed method text analysis in the QDA community may grow. [52]

The complexity of this undertaking advises not to do this in single person projects or restricted to one discipline. Recent developments have shown that these methods are developed and tested best by interdisciplinary research teams bringing together social scientists, linguists and computer scientists. In a current funding line of the German Federal Ministry of Education and Research (BMBF), 24 interdisciplinary projects in the field of "digital humanities" are funded for three years. At least six of them have a dedicated social science background, thus fulfilling the requirement of the funding line which explicitly had called qualitatively researching social scientists for participation (BMBF, 2011). So far these projects make clear: there is no "out-of-the-box" solution on the way to answer their research questions—neither from a technical perspective, nor from a methodological one. Each has to develop its own way of proceeding, as well as to reinvent or adapt existing analysis technologies for their specific purpose. There is no, and probably will never be a "one button" solution to CATA, because of the simple fact, that generic approaches are not appropriate to satisfy specific and complex research needs. [53]

But if qualitative oriented social science research takes the plunge to join into this interdisciplinary cooperation it may be of fruitful benefit for all participants. Computer scientists and linguists may sharpen their methods and tools on "real world" problems gaining knowledge on applicability of their approaches. Social science may further blur the obstructive and rather artificial distinction of the qualitative vs. quantitative research paradigm towards a fruitful integration of both. For the future of computer assisted text analysis I expect, that 1. the more the applied algorithms are able to dig into "latent" meaning rather than counting surface observations they help to bridge the gap between qualitative and quantitative QDA, and 2. as long as they are able to keep the link between the qualitative input data and their quantified results, they enable the researcher to build confidence in this approach. Given these conditions "distant" and "close reading" may interact fruitfully and quantitative text analysis may keep a "qualitative quality." [54]

Acknowledgments

For the opportunity to gain experience in this interdisciplinary field I would like to thank all my colleagues of the ePol project—an eHumanities research project dedicated to investigate the hypothesis on post-democracy in Germany in a long time-frame. Furthermore, the workshop "Political Science and the Methods of the eHumanities" in November 2012 in Hamburg was very inspiring concerning the appliance of mass textual analysis in different social science disciplines—thanks to all its participants. Last but not least, thanks to Alexander REISENBICHLER for his helpful comments on this article.

References

Adams, Josh & Roscigno, Vincent J. (2005). White supremacists, oppositional culture and the world wide web. *Social Forces*, *84*(2), 759-778.

Angermüller, Johannes (2005). "Qualitative" methods of social research in France: Reconstructing the actor, deconstructing the subject. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *6*(3), Art. 19, http://nbn-resolving.de/urn:nbn:de:0114-fqs0503194 [Accessed: May 14, 2013].

Baker, Paul; Gabrielatos, Costas; Khosravi; Nik, Majid; Krzyzanowski, Michael; McEnery, Tony & Wodak, Ruth (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & Society*, *19*(3), 273-306.

Belsky, Gary (2012). Why text mining may be the next big thing. *TIME*, http://business.time.com/2012/03/20/why-text-mining-may-be-the-next-big-thing [Accessed: June 8, 2012].

Bergmann, Gustav (1952). Two types of linguistic philosophy. *The Review of Metaphysics*, *5*(3), 417-438.

Blätte, Andreas (2012). Unscharfe Grenzen von Policy-Feldern im parlamentarischen Diskurs. Messungen und Erkundungen durch korpusunterstützte Politikforschung. *Zeitschrift für Politikwissenschaft*, *22(*1), 35-68.

Blei, David M. (2012). Probabilistic topic models. Surveying a suite of algorithms that offer a solution to managing large document archives. *Communications of the ACM*, *55*(4), 77-84.

Blei, David M.; Ng, Andrew Y. & Jordan, Michael I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993-1022, http://www.cs.princeton.edu/~blei/papers/BleiNgJordan2003.pdf [Accessed: May 14, 2013].

BMBF (2011). *Bekanntmachung des Bundesministeriums für Bildung und Forschung von Richtlinien zur Förderung von Forschungs- und Entwicklungsvorhaben aus dem Bereich der eHumanities*, http://www.bmbf.de/foerderungen/16466.php [Accessed: January 16, 2012].

Bonzio, Roberto (2011). Father Busa, pioneer of computing in humanities with Index Thomisticus, dies at 98. *Forbes*, http://www.forbes.com/sites/robertobonzio/2011/08/11/father-busa-pioneer-of-computing-in-humanities-dies-at-98 [Accessed: May 4, 2013].

Brier, Alan & Hopp, Bruno (2011). Computer assisted text analysis in the social sciences. *Quality & Quantity*, *45*(1), 103-128.

Bubenhofer, Noah (2009). *Sprachgebrauchsmuster. Korpuslinguistik als Methode der Diskurs- und Kulturanalyse*. Berlin: de Gruyter.

Busa, Roberto A. (2004). Foreword: Perspectives on the digital humanities. In Susan Schreibman, Raymond George Siemens & John Unsworth (Eds.), *A companion to digital humanities* (pp.xvi-xxi). Malden, MA: Blackwell.

Crane, Gregory (2006). What do you do with a million books? *D-Lib Magazine*, *12*(3), http://www.dlib.org/dlib/march06/crane/03crane.html [Accessed: January 9, 2012].

Dzudzek, Iris; Glasze, Georg; Mattissek, Annika & Schirmel, Henning (2009). Verfahren der lexikometrischen Analyse von Textkoprora. In Georg Glasze & Annika Mattissek (Eds.), *Handbuch Diskurs und Raum. Theorien und Methoden für die Humangeographie sowie die sozial- und kulturwissenschaftliche Raumforschung* (pp.233-260). Bielefeld: transcript.

Evers, Jeanine C.; Silver, Christina; Mruck, Katja & Peeters, Bart (2011). Introduction to the KWALON experiment: Discussions on qualitative data analysis software by developers and users. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *12*(1), Art. 40, http://nbn-resolving.de/urn:nbn:de:0114-fqs1101405 [Accessed: May 14, 2013].

Flick, Uwe (2007). Zur Qualität qualitativer Forschung – Diskurse und Ansätze. In Udo Kuckartz (Ed.), *Qualitative Datenanalyse computergestützt. Methodische Hintergründe und Beispiele aus der Forschungspraxis* (2nd ed., pp.188-209). Wiesbaden: VS Verlag für Sozialwissenschaften.

Friese, Susanne (2011). Using ATLAS.ti for analyzing the financial crisis data. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *12*(1), Art. 39, http://nbn-resolving.de/urn:nbn:de:0114-fqs1101397 [Accessed: May 14, 2013].

Glasze, Georg (2007). Vorschläge zur Operationalisierung der Diskurstheorie von Laclau und Mouffe in einer Triangulation von lexikometrischen und interpretativen Methoden. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *8*(2), Art. 14, http://nbn-resolving.de/urn:nbn:de:0114-fqs0702143 [Accessed: February 2, 2012].

Grimmer, Justin (2010). A Bayesian hierarchical topic model for political texts. Measuring expressed agendas in senate press releases. *Political Analysis*, *18*(1), 1-35.

Guilhaumou, Jacques (2008). Geschichte und Sprachwissenschaft. Wege und Stationen (in) der "analyse du discours". In Reiner Keller, Andreas Hirseland, Werner Schneider & Willy Viehöver (Eds.), *Handbuch sozialwissenschaftliche Diskursanalyse 2. Forschungspraxis* (3rd ed., pp.21-67). Wiesbaden: VS Verlag für Sozialwissenschaften.

Helsloot, Niels & Hak, Tony (2007). Pêcheux's contribution to discourse analysis. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research, 8*(2), Art. 1, http://nbn-resolving.de/urn:nbn:de:0114-fqs070218 [Accessed: June 14, 2012].

Heyer, Gerhard (2009). Introduction to TMS 2009. In Gerhard Heyer (Ed.), *Text mining services. Building and applying text mining based service infrastructures in research and industry; proceedings of the Conference on Text Mining Services 2009 at Leipzig University* (pp.1-14). Leipzig: LIV (Leipziger Beiträge zur Informatik, 14).

Hopkins, Daniel J. & King, Gary (2010). A method of automated nonparametric content analysis for social science. *American Journal for Political Science*, *54*(1), 229-247.

Janasik, Nina; Honkela, Timo & Bruun, Henrik (2009). Text mining in qualitative research. Application of an unsupervised learning method. *Organizational Research Methods*, *12*(3), 436-460.

Kelle, Udo (1997). Theory building in qualitative research and computer programs for the management of textual data. *Sociological Research Online*, *2*(2), http://www.socresonline.org.uk/2/2/1.html [Accessed: January 9, 2012].

Kelle, Udo (2008). Computergestützte Analyse qualitativer Daten. In Uwe Flick (Ed.), *Qualitative Forschung. Ein Handbuch* (6th ed., pp.485-502). Reinbek: Rowohlt.

Kelle, Udo (2011). Computerunterstützung in der qualitativen Forschung. In Ralf Bohnsack, Winfried Marotzki & Michael Meuser (Eds.), *Hauptbegriffe Qualitativer Sozialforschung* (3rd ed., pp.29-31). Opladen: Budrich.

Kracauer, Siegfried (1952). The challenge of qualitative content analysis. *Public Opinion Quarterly*, *16*(4), 631–642, http://www.jstor.org/stable/2746123 [Accessed: December 2, 2012].

Krippendorff, Klaus (2013). *Content analysis. An introduction to its methodology* (3rd ed.). Los Angeles, CA: Sage.

Kuckartz, Udo (2007). QDA-Software im Methodendiskurs: Geschichte, Potenziale, Effekte. In Udo Kuckartz (Ed.), *Qualitative Datenanalyse computergestützt. Methodische Hintergründe und Beispiele aus der Forschungspraxis* (2nd ed., pp.15-31). Wiesbaden: VS Verlag für Sozialwissenschaften.

Kuckartz, Udo (2010). *Einführung in die computergestützte Analyse qualitativer Daten* (3rd ed.). Wiesbaden: VS Verlag für Sozialwissenschaften.

Kuş Saillard, Elif (2011). Systematic versus interpretive analysis with two CAQDAS packages: NVivo and MAXQDA. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *12*(1), Art. 34, http://nbn-resolving.de/urn:nbn:de:0114-fqs1101345 [Accessed: June 22, 2012].

Lejeune, Christophe (2011). From normal business to financial crisis ... and back again. An illustration of the benefits of Cassandre for qualitative analysis. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *12*(1), Art. 24, http://nbn-resolving.de/urn:nbn:de:0114-fqs1101247 [Accessed: February 25, 2011].

Lowe, Will (2003). The statistics of text: New methods for content analysis. *Midwest Political Science Association Conference. Chicago, April 3-6, 2003*. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.129.5225&rep=rep1&type=pdf [Accessed: October 4, 2012].

Mautner, Gerlinde (2009). Checks and balances: How corpus linguistics can contribute to CDA. In Ruth Wodak & Michael Meyer (Eds.), *Methods of critical discourse analysis* (pp.122-143). London: Sage.

Mayring, Philipp (2010). *Qualitative Inhaltsanalyse. Grundlagen und Techniken* (11th ed.). Weinheim: Beltz.

Moretti, Franco (2000). Conjectures on world literature. *New Left Review*, *1*, 54-68.

Moretti, Franco (2007). *Graphs, maps, trees. Abstract models for literary history*. London: Verso.

Mühlmeyer-Mentzel, Agnes (2011). Das Datenkonzept von ATLAS.ti und sein Gewinn für "Grounded-Theory"-Forschungsarbeiten. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *12*(1), Art. 32, http://nbn-resolving.de/urn:nbn:de:0114-fqs1101325 [Accessed: May 29, 2012].

Mühlmeyer-Mentzel, Agnes & Schürmann, Ingeborg (2011). Softwareintegrierte Lehre der Grounded-Theory-Methodologie. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *12*(3), Art. 17, http://nbn-resolving.de/urn:nbn:de:0114-fqs1103171 [Accessed: October 25, 2011].

Pêcheux, Michel (1969). *Analyse automatique du discours*. Paris: Dunod.

Pêcheux, Michel; Hak, Tony & Helsloot, Niels (1995). *Automatic discourse analysis*. Amsterdam: Rodopi.

Pollak, Senja; Coesemans, Roel; Daelemans, Walter & Lavrac, Nada (2011). Detecting contrast patterns in newspaper articles by combining discourse analysis and text mining. *Pragmatics*, *21*(4), 647-683.

Samuelsson, Christer (2004). Statistical methods. In Ruslan Mitkov (Ed.), *The Oxford handbook of computational linguistics* (pp.358-375). Oxford: Oxford University Press.

Scharkow, Michael (2012). *Automatische Inhaltsanalyse und maschinelles Lernen*. Berlin: epubli.

Scharloth, Joachim; Eugster, David & Bubenhofer, Noah (2013/in print). Das Wuchern der Rhizome. Linguistische Diskursanalyse und Data-driven Turn. In Dietrich Busse & Wolfgang

Teubert (Eds.), *Linguistische Diskursanalyse. Neue Perspektiven*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Schönfelder, Walter (2011). CAQDAS and qualitative syllogism logic—NVivo 8 and MAXQDA 10 compared. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *12*(1), Art. 21, http://nbn-resolving.de/urn:nbn:de:0114-fqs1101218 [Accessed: October 2, 2012].

Stone, Phillip J. (1997). Thematic text analysis: New agendas for analyzing text content. In Carl W. Roberts (Ed.), *Text analysis for the social sciences. Methods for drawing statistical inferences from texts and transcripts* (pp.35-54). Mahwah: Erlbaum.

Tamayo Korte, Miguel; Waldschmidt, Anne; Dalman-Eken, Sibel & Klein, Anne (2007). 1000 Fragen zur Bioethik – Qualitative Analyse eines Onlineforums unter Einsatz der quantitativen Software MAXDiction. In Udo Kuckartz (Ed.), *Qualitative Datenanalyse computergestützt. Methodische Hintergründe und Beispiele aus der Forschungspraxis* (2nd ed., pp.163-174). Wiesbaden: VS Verlag für Sozialwissenschaften.

Teubert, Wolfgang (2006). Korpuslinguistik, Hermeneutik und die soziale Konstruktion der Wirklichkeit. *Linguistik Online*, *28*(3), http://www.linguistik-online.de/28_06/teubert.html [Accessed: January 30, 2012].

Züll, Cornelia & Mohler, Peter (2001). Computerunterstützte Inhaltsanalyse: Codierung und Analyse von Antworten auf offene Fragen. *GESIS-How-to-8*, http://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/howto/how-to8cz.pdf [Accessed: April 10, 2012].

Author

Gregor WIEDEMANN studied political science and computer science in Leipzig. Currently he is working on his doctoral thesis about the application of computer-assisted approaches for qualitative data analysis in social sciences. As team member of the Natural Language Processing Group at the University of Leipzig he is involved in an interdisciplinary research project which investigates the evolvement of political justifications in the German public media between 1949 and 2011 (ePol project).

Contact:

Gregor Wiedemann, M.A.

NLP Group | Department of Computer Science
University of Leipzig
Augustusplatz 10
04109 Leipzig, Germany

E-mail: gregor.wiedemann@uni-leipzig.de
URL: http://asv.informatik.uni-leipzig.de/en/staff/Gregor_Wiedemann

Citation

Wiedemann, Gregor (2013). Opening up to Big Data: Computer-Assisted Analysis of Textual Data in Social Sciences [54 paragraphs]. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *14*(2), Art. 13, http://nbn-resolving.de/urn:nbn:de:0114-fqs1302231.