

## 2. Computer-Assisted Text Analysis in the Social Sciences

Despite there is a long tradition of Computer Assisted Text Analysis (CATA) in social sciences, it followed a rather parallel development to QDA. Only a few years ago, realization of TM potentials for QDA started to emerge slowly. In this chapter, I reflect on the debate of the use of software in qualitative social science research together with approaches of text analysis from the NLP perspective. For this, I shortly elaborate on the quality versus quantity divide in social science methods of text analysis (2.1). Subsequently, perspectives and technologies of text analysis from NLP perspective are introduced briefly (2.2). Finally, I suggest a typology of computer-assisted text analysis approaches utilized in social science based on the notion of context underlying the analysis methods (2.3). This typology helps to understand why developments of qualitative and quantitative CATA have been characterized by mutual neglect for a long time, but recently opened perspectives for integration of both research paradigms—a progress mainly achieved through advancements in Machine Learning (ML) for text. Along with the typology descriptions example studies utilizing different kinds of CATA approaches are given to introduce on related work to this study.

### 2.1. Text as Data between Quality and Quantity

When analyzing text, social scientists strive for inference on social reality. In contrast to linguists who mainly focus on description of language regularities itself, empirical language use for sociologists or political scientists is more like a window through which they try to re-

construct the ways speaking actors perceive themselves and the world around them. Systematic reconstruction of the interplay between language and actors' perception of the world contributes to much deeper understanding of social phenomena than purely quantitative methods of empirical social research, e.g. survey studies, could deliver. Consequently, methodical debates on empirical social research distinguish between reconstructivist and hypothesis testing approaches (Bohnsack, 2010, p. 10). While research approaches of hypothesis testing aim for intersubjectively reliable knowledge production by relying on a quantitative, statistical perspective, reconstructivist approaches share a complicated relationship with quantification. As already mentioned in the introduction, it is a puzzling question why social science, although having put strong emphasis on analyzing textual data for decades, remained skeptical for so long about computer-assisted approaches to analyze large quantities of text. The answer in my opinion is two-fold, comprising a methodological and a technical aspect. The methodological aspect is reflected in the following, while I highlight on the technical obstacles in Section 2.3.

In the German as well as in the Anglo-Saxon social research community a deep divide between quantitative and qualitative oriented methods of empirical research has evolved during the last century and is still 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. Following a positivist epistemological conceptualization of the world, media scientists up to the mid 20th century perceived qualitative data only as a sequence of symbols, which could be observed and processed as unambiguous analysis units by non-skilled human coders or computers to produce scientific knowledge. Analyses were run on a large numbers of cases, but tended to oversimplify complex societal procedures by application of fixed (deductive) categories. As a counter model, during the 1970s, the post-positivist paradigm led to the emergence of several qualitative text analysis methodologies seeking to generate an in-depth comprehension of a rather small number

**Table 2.1.:** Examples for two kinds of software products supporting text analysis for linguistic and social research.

Data management	Data processing
Atlas.ti, MAXQDA, QDA-Miner, NVivo, QCAmap, CATMA, LibreQDA	MAXDictio, WordStat (QDAMiner), WordSmith, Alceste, T-LAB, Lexico3, IRaMuteQ, Leipzig Corpus Miner

of cases. Knowledge production from text was done by intense close reading and interpretation of trained human analysts in more or less systematic ways.

Computer software has been utilized for both paradigms of text analysis, but of course, provided very distinct functions for the analysis process. Analogous to the qualitative-quantitative divide, two tasks for Computer Assisted Text Analysis can be distinguished:

- data management, and
- data processing.

Table 2.1 illustrates examples of software packages common in social science for qualitative and quantitative text analysis.

*Data processing* of large document sets for the purpose of quantitative content analysis framed the early perception of software usage for text analysis from the 1960s onward. For a long time, using computers for QDA appeared somehow as retrogression to protagonists of truly qualitative approaches, especially because of their awareness of the history of flawed quantitative content analysis. Software for *data management* to support qualitative analysts by annotating parts of text with category codes has been accepted only gradually since the late 1980s. On the one hand, a misunderstanding was widespread that such programs, also referred to as Computer Assisted Qualitative Data Analysis (CAQDA), should 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 advantageous to increase the number of cases in qualitative research designs by using computer software. To generate insight into their subject matter, researchers should not concentrate on as many cases as possible, but on as most 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 CAQDA 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 which circumstances quantification may be applied to coding results.

To overcome shortcomings of both, the qualitative and the quantitative research paradigm, novel ‘mixed method’ designs are gradually introduced in QDA. Although the methodological perspectives of quantitative content analysis and qualitative methods are almost diametrically opposed, application of CATA may be fruitful not only as a tool for exploration and heuristics. Functions to evaluate quantitative aspects of empirical textual data (such as the extension MAXDictio for the software MAXQDA), have been integrated in all recent versions of the leading QDA software packages. Nevertheless, studies on the usage of CAQDA indicate that qualitative researchers usually confine themselves to the basic features (Kuckartz, 2007, p. 28). Users 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 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, critics of qualitative research often see these rather ‘soft’ criteria as a shortcoming of QDA compared to what they conceive as ‘hard science’ based on knowledge represented by numeric values and significance measures.

Proponents of ‘mixed methods’ do not consider both paradigms as being contradictory. Instead, they stress advantages of integration of both perspectives. 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; translation GW). His perspective suggests that qualitative and quantitative approaches of text analysis should not be perceived 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 which 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. 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.

In contrast to the early days of computer-assisted text analysis which spawned the qualitative-quantitative divide, in the last decades computer-linguistics and NLP have made significant progress incorporating linguistic knowledge and context information into its analysis routines, thereby overcoming the limitations of simple “term based analysis functions” (ibid., p. 218). Two recent developments of computer-assisted text analysis may severely change the circumstances which in the past have had been serious obstacles to a fruitful integration of qualitative and quantitative QDA. Firstly, the availability and processability of full-text archives enables researchers to generate insight from quantified qualitative analysis results through comparison of different sub populations. A complex research design as suggested in this study is able 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 ( $n \approx 1,000$ ) may be drawn, or even the full corpus ( $n \gg 100,000$ ) 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 here is the algorithmic extraction of *meaning*, which is approached by the inclusion of different levels of *context* into a complex analysis workflow integrating systematically several TM applications of distinct types. How meaning is extracted in NLP will be introduced in the next section. Then, I present in detail the argument why modern TM applications contribute to bridge the seemingly invincible qualitative-quantitative divide.

## 2.2. Text as Data for Natural Language Processing

For NLP, text as data can be encoded in different ways with respect to the intended algorithmic analysis. These representations model semantics distinctively to allow for the extraction of meaning (2.2.1). Moreover, textual data has to be preprocessed taking linguistic knowledge into account (2.2.2), before it can be utilized as input for TM applications extracting valuable knowledge structures for QDA (2.2.3).

### 2.2.1. Modeling Semantics

If computational methods should be applied for QDA, models of semantics of text are necessary to bridge the gap between research interests and algorithmic identification of structures in textual data. Turney and Pantel (2010, p. 141) refer to semantics as “in a general sense [...] the meaning of a word, a phrase, a sentence, or any text in human language, and the study of such meaning”. Although there was some impressing progress in the field of artificial intelligence and ML

in recent decades, computers still lack of intelligence comparable to humans regarding learning, comprehension and autonomous problem solving abilities. In contrast, computers are superior to human abilities when it comes to identify structures in large data sets systematically. Consequently, to utilize computational powers for NLP we need to link computational processing capabilities with analysis requirements of human users. In NLP, three types of semantic representations may be distinguished:

1. patterns of character strings,
2. logical rule sets of entity relations, and
3. distributional semantics.

Text in computational environments generally is represented by *character strings* as primary data format, i.e., sequences of characters from a fixed set which represent meaningful symbols, e.g., letters of an alphabet. The simplest model to process meaning is to look for fixed, predefined patterns in these character sequences. For instance, we may define the character sequence *United States* occurring in a text document as representation of the entity ‘country United States of America’. By extending this single sequence to a set of character strings, e.g. “United States”, “Germany”, “Ghana”, “Israel”, . . . , we may define a representation of references to the general entity ‘country’. Such lists of character sequences representing meaningful concepts, also called ‘dictionaries’, have a long tradition in communication science (Stone, 1996). They can be employed as representations of meaningful concepts to be measured in large text collections. By using regular expressions<sup>1</sup> and elaborated dictionaries it is possible to model very complex concepts.<sup>2</sup> In practice, however, success of this

<sup>1</sup>Regular expressions are a formal language to fulfill ‘search and replace’ operations. With a special syntax complex search patterns can be formulated to identify matching parts in a target text.

<sup>2</sup>The pattern `\d+ (protester|people|person) [\w\s]* (injured|hurt|wounded)`, for example, would match text snippets containing a number (`\d+`) followed by mentioning of a group together with verbs indicating injury in any permutation where only word characters or spaces are located between them (`([\w\s]*)`).

approach still depends on the skill and experience of the researcher who creates such linguistic patterns. In many cases linguistic expressions of interest for a certain research question follow rather fixed patterns, i.e. repeatedly observable character strings. Hence, this rather simple approach of string or regular expression matching can already be of high value for QDA targeted to manifest content.

A much more ambitious approach to process semantics is the employment of *logic frameworks*, e.g., predicate logic or first-order logic, to model relations between units represented by linguistic patterns. Instead of just searching for patterns as representatives for meaning in large quantities of text, these approaches strive for inference of ‘new’ knowledge not explicitly contained in the data basis. New knowledge is to be derived deductively from an ontology, i.e., a knowledge base comprising of variables as representatives of extracted linguistic units and well-formed formulas. Variables may be formally combined by functions, logical connectives and quantifiers that allow for reasoning in the ontology defined. For example, the set of two rules 1)  $car(b) \wedge red(b)$ , 2)  $\forall x(car(x) \rightarrow vehicle(x))$  would allow to query for the red vehicle  $b$ , although the knowledge base only contains explicit information about the red car  $b$  (rule 1), because the second rule states that all cars are vehicles. Setting up a formal set of rules and connections of units in a complete and coherent way, however, is a time consuming and complex endeavor. Quality and level of granularity of such knowledge bases are insufficient for the most practical applications. Nevertheless, there are many technologies and standards such as Web Ontology Language (OWL) and Resource Description Framework (RDF) to represent such semantics with the objective to further develop the internet to a ‘semantic web’. Although approaches employing logic frameworks definitely model semantics closer to human intelligence, their applicability for QDA on large data sets is rather limited so far. Not only that obtaining knowledge bases from natural language text is a very complex task. Beyond manifest expressions content analytic studies are also interested in latent meaning. Modeling latent semantics by formal logic frameworks is a very tricky task, so far not solved for NLP applications in a satisfying manner.



Most promising for QDA are distributional approaches to process semantics because they are able to cover both, manifest and latent aspects of meaning. *Distributional semantics* is based on the assumption that statistical patterns of human word usage reveal what people mean, and “words that occur in similar contexts tend to have similar meanings” (Turney and Pantel, 2010). Foundations for the idea that meaning is a product of contextual word usage have been established already in the early 20th century by emerging structural linguistics (Saussure, 2001; Harris, 1954; Firth, 1957). To employ statistical methods and data mining to language, textual data needs to be transformed into numerical representations. Text no longer is comprehended as a sequence of character strings, instead character strings are chopped into lexical units and transformed into a numerical vector. The Vector Space Model (VSM), introduced for IR (Salton et al., 1975) as for many other NLP applications, encodes counts of occurrences of single terms in documents (or other context units, e.g., sentences) in vectors of the length of the entire vocabulary  $V$  of a modeled collection. If there are  $M = |V|$  different word types in a collection of  $N$  documents, then the counts of the  $M$  word types in each of the documents leads to  $N$  vectors which can be combined into a  $N \times M$  matrix, a so-called Document-Term-Matrix (DTM). Such a matrix can be weighted, filtered and manipulated in multiple ways to prepare it as an input object to many NLP applications such as extraction of meaningful terms per document, inference of topics or classification into categories. We can also see that this approach follows the ‘bag of words’ assumption which claims that frequencies of terms in a document mainly indicate its meaning; order of terms in contrast is less important and can be disregarded. This is certainly not true for most human real world communication, but works surprisingly well for many NLP applications.<sup>3</sup>

---

<sup>3</sup>The complete loss of information on word order can be mitigated by observing  $n$ -grams, i.e. concatenated ongoing sequences of  $n$  terms instead of single terms while creating a DTM.

### 2.2.2. Linguistic Preprocessing

Analyzing text computationally in the sense of distributional semantics requires the transformation of documents, i.e. sequences of character strings, into numerical data suitable for quantifying evaluation, statistical inference or modeling. Usually, for such a transformation documents need to be separated into single lexical units, which then are counted. Depending on the application, the analysis unit for counts may be altered from documents to paragraphs or single sentences to narrow down certain contexts, or document sets for aggregating information on higher discursive levels. After definition of the analysis unit and its corresponding data separation, e.g. detecting sentence boundaries in documents, single lexical units, also known as *tokens*, need to be identified. This process, called ‘tokenization’, separates all distinct word forms present in the entire text corpus. Such distinct word forms are called *types*. Again, counts of types for every analysis unit can be encoded and stored in a vector—collections in a DTM respectively.

The way in which text is tokenized mainly influences posterior analysis steps as it defines the atomic representatives of semantics. Tokens might be single terms, punctuation marks, multi-word units, or concatenations of  $n$  tokens, so called  $n$ -grams encoding different aspects of semantics numerically. Computer linguistics comprises of a variety of procedures to preprocess textual data before encoding it in a DTM. After initial encoding, the DTM may be further preprocessed mathematically, e.g. to weight terms by their contribution to document meaning. Linguistic and mathematical preprocessing of the DTM prepare subsequent TM analysis. The following list briefly introduces the most common preprocessing steps:

- Sentence segmentation: For certain TM applications, single sentences need to be identified in documents. The simplest approach would be to separate by locating punctuation marks or full stops. However, this produces false separations in certain cases, e.g. abbreviations or date formats. More sophisticated approaches utilize probabilistic models to determine whether punctuation marks in-

dicating ends of sentences by observing their context (Reynar and Ratnaparkhi, 1997).

- **Tokenization:** Separation of text into single tokens can be achieved in many languages simply by separating at white space characters. However, this base line approach misses separation of punctuation marks from single terms or does not cover recognition of Multi Word Units (MWUs). Again, more sophisticated approaches utilize probabilistic models trained on manually tokenized data to decide on boundaries of lexical units more accurately.
- **Cleaning:** For specific use cases, not all identified types of lexical units contribute to the desired level of meaning. For example, stop words such as articles or pronouns often do not cover relevant aspects of meaning in a ‘distant reading’ perspective. The same can be valid for punctuation marks or numbers in the text. If useful, such types of lexical units can be omitted to reduce the amount of data and concentrate on the most meaningful language aspects for subsequent analysis.
- **Unification:** Lexical units occur in different ways of spelling and syntactical forms. Variants of the same noun may occur in singular, plural or different cases, verbs may be inflected. Unification procedures reduce such forms to a single basic form, to treat occurrences of variances in the data as identical event for all further applications. Common forms of unification are reduction of characters to lowercase, stemming and lemmatization. For stemming word stems of terms are guessed by cutting suffixes from tokens according to a language specific rule set. For lemmatization, large language specific lists which contain assignments of inflected forms to corresponding dictionary forms are utilized to look up and replace any occurrence of a token by its lemma.
- **Part of Speech (POS):** In POS-tagging any token in a sequence of tokens, e.g. in a sentence, is labeled with a part of speech label, e.g. NN for nouns, ADJ for adjectives, VA for auxiliary verb (Heyer

et al., 2006, p. 126). POS labels may be utilized as filter during preprocessing, e.g. to just concentrate on nouns for certain analysis. They also can be helpful for disambiguation of homonyms (e.g. *can\_VA* versus *can\_NN*), hence, contributing to capture desired semantics more accurately.

- Pruning: Characteristics of term distributions in natural language can be formally described by Zipf's law (Heyer et al., 2006, p. 87). From Zipf's law it can be inferred that most frequent types do not contribute much to specific constitution of meaning in a text, and that roughly half of the types only occur once in the entire corpus. Hence, pruning the most and least frequent terms for DTM generation while preprocessing helps to keep data objects manageable in size and concentrate on the most meaningful lexical units. Pruning can be done in absolute manner (omitting terms occurring more or less than  $n$  times in the corpus) or relative manner (omitting terms occurring in more or less than  $p$  percent of documents of the corpus).

These procedures of preprocessing distinctively shape the set of types to be counted to prepare a DTM by identifying, transforming and filtering lexical units with respect to linguistic knowledge. There is no ideal or correct configuration of such a preprocessing chain. Instead, each application demands its own parameter settings to yield optimal results. For example, in QDA scenarios stemming might contribute to performance gains in a classification task through extensive feature unification while it produces artificial homonymy and unpleasant term stubs in co-occurrence analysis. Often it is necessary to experiment with different parameters for preprocessing before deciding which results fit best to study requirements.

### 2.2.3. Text Mining Applications

Once a document collection is encoded in a numerical DTM format, it can be utilized as input for various TM applications. Regarding TM

applications, I distinguish in lexicometric and Machine Learning approaches. Lexicometric approaches calculate statistics on closed data sets, rank observed events and highlight on those where observations deviate from expectations. ML approaches ‘learn’ data regularities by inferring discriminative or generative probabilistic models. Such models can be applied to previously unseen data to identify structures or patterns. Within this study, I refer to both, lexicometric and ML applications, as Text Mining applications.

### Lexicometrics

Lexicometric analysis on digital text has been utilized since the early beginning of computational text processing and is widely used in corpus linguistics. Over the decades the method toolbox has been extended from simple frequency counts to more elaborated statistical methods:

- *Frequency analysis*: In this application observations of events, e.g. specific terms or concepts occurring in documents, are counted and counts are compared across dimensions, e.g. time. Observing term frequencies in a longitudinal view over several decades may reveal peaks and dips in term usage, and corresponding concepts. Events for observation can be defined in distinguished ways, e.g. as raw term frequencies or as document frequencies where multiple occurrences of one term in the same document are counted only once. Beyond just single terms, more meaningful concepts can be counted by defining sets of terms as events which either must occur together in a specific context unit, or are treated as a list of synonyms. Utilization of such lists is also called dictionary analysis (Stone et al., 1966).
- *Key term extraction*: This application identifies important terms in documents or entire collections by applying statistical measures (Archer, 2008). The established method of difference analysis compares term frequencies in a target text (or an entire collection) to frequencies in a reference corpus, e.g. a collection of general texts of

the same language without a bias to any topic. Deviations between expectations based on the comparison text and observations in the target text are evaluated by a statistical test resulting in lists of terms ranked by ‘keyness’. Terms of such lists can be displayed in Key Word in Context (KWIC) views which allow for quick qualitative assessment of usage contexts of terms in a collection (Luhn, 1960).

- *Co-occurrence analysis*: For co-occurrence analysis<sup>4</sup>, joint occurrence of events in a well defined context unit is observed and evaluated by a statistical test (Bordag, 2008; Büchler, 2008). For any word type it reveals a ranked list of other words which co-occur with it, e.g. in a sentence or as its left / right neighbor, more often than expected under the assumption of independence. In accordance with structuralist linguistic theory, this reveals semantic fields of syntagmatically related terms. Comparing and ranking such semantic fields by similarity further may reveal paradigmatically related terms, i.e. words occurring in similar contexts (Heyer et al., 2006, p. 19ff).
- *Dimension reduction*: The idea of co-occurrence of two terms can be extended to observation of co-occurrence of multiple terms to infer on latent structures. For this, various methods of dimension reduction from data mining are also applicable to DTMs extracted from text collections. In Principal Component Analysis (PCA), Multi Dimensional Scaling (MDS) or Correspondence Analysis continuous or categorical data incorporated in a DTM can be reduced to its main components or projected into a two-dimensional space. The reduced two dimensions of the vocabulary, for example, may be utilized to visualize semantic proximity of terms. A higher number of reduced dimensions may be utilized to infer on similarity of documents in a latent semantic space. As Latent Semantic Analysis (LSA), dimension reduction has also been utilized in Information Retrieval (Deerwester et al., 1990).

---

<sup>4</sup>In linguistic contexts it is also referred to as collocation analysis.

## Machine learning

While lexicometric approaches are widely used in corpus linguistics, the exploration of ML applications for QDA in social science is just at its beginning. Tom Mitchell formally defines ML as follows: “A computer program is said to learn from experience  $E$  with respect to some class of tasks  $T$  and performance measure  $P$ , if its performance at tasks in  $T$ , as measured by  $P$ , improves with experience  $E$ ” (1997, p. 2). While lexicometric measures lack of the ‘learning’ property through ongoing ‘experience’ of data observation, ML incorporates such experience in model instances. Model parameters can be updated with new units of observed data which make the concept interesting especially for large data sets and streaming data. For analyzing textual data, several kinds of ML applications have been developed.

Analogue to data mining, we can distinguish *unsupervised* from *supervised* methods for data analysis. Unsupervised, data-driven approaches identify previously unknown patterns and structures emerging from the data itself. They provide a clustering of data points satisfying certain similarity criteria (e.g. similarity of documents based on word usage, or similarity of terms based on their contexts). Supervised classification methods in contrast utilize document external knowledge, e.g. information on class membership of a document, to model the association between that external observation and features of the document. This allows to assign category labels to new, unknown documents (or document fragments), analogously to manual coding in a content analysis procedure. These methods resemble research paradigms in data analysis for social sciences. While the unsupervised methods help to explore structures in large amounts of unknown data, thus supporting *inductive* research approaches of text analysis, supervised methods may take into account external, theory-led knowledge to realize *deductive* research workflows.

Useful Text Mining applications for QDA following the paradigm of *unsupervised learning* are:

- *Document clustering*: For cluster analysis, context units such as sentences or documents have to be grouped according to similarity of

their content, e.g. based on common term usage (Heyer et al., 2006, p. 195ff). Clusters should have the property of optimal similarity of documents within the cluster and maximum difference of documents between clusters. Variants exist for strict partitioning versus hierarchical or overlapping (soft) clustering. For some algorithms the number of clusters for partitioning has to be given as external parameter, some try to identify an optimal number of clusters on their own. With the help of clustering analysts can separate large collections into manageable sub-collections, explore collections by semantic coherence and concentrate on the most meaningful ones.

- *Topic Models*: refer to a set of “algorithms for discovering the main themes that pervade a large and otherwise unstructured collection of documents [...] topic models can organize the collection according to the discovered themes” (Blei, 2012, p. 77). Since the initial Latent Dirichlet Allocation (LDA) model developed by Blei et al. (2003) a variety of more complex topic models has been introduced (Blei and Lafferty, 2006; Mcauliffe and Blei, 2008; Grimmer, 2010; Mimno et al., 2011). All these models assume a generative process of text production governed by a probability distribution of topics within a document and a probability distribution of terms for each topic. Via a complex inference mechanism on the observed words per document in a collection, they infer on semantic coherent terms representing topics, and proportions of topics contained in each document as latent variables. In analogy to LSA, “LDA can also be seen as a type of principal component analysis for discrete data” (Blei, 2012, p. 80). Among other things, topic models provide two valuable matrices as a result of the inference process:
  - matrix  $\beta$  of the dimensions  $|V| \times K$  containing a posterior probability distribution over the entire vocabulary  $V$  for each of the  $K$  modeled topics,
  - matrix  $\theta$  of the dimensions  $N \times K$  containing a posterior probability distribution over all  $K$  topics for each of the  $N$  documents in a collection.



With these results analysts can reveal thematic structures, filter large document collections or observe changes in topic proportions over time. Consequently, topic models can be seen as a kind of soft or fuzzy clustering giving the likelihood of a document belonging to a certain thematic cluster.

- *Dimensional Scaling*: Especially for political science analysis, dimensional scaling strives for assigning a measurement to documents of a collection representing a relative position on a one-dimensional scale (Benoit and Laver, 2012). For thematically coherent collections, e.g. parliamentary speeches on a single draft bill or party manifestos, this measurement shall represent political position between a left and a right pole of the spectrum. Based on word usage, centrist or outer political positions of single speeches, parliamentarians or parties may be determined.

Useful Text Mining applications for QDA following the paradigm of *supervised learning* are:

- *Classification*: While clustering assigns any unit of analysis to a group of other units based on the emergent structure from within the data itself, supervised classification relies on external information to group the units. This external knowledge usually is a set of categories (classes) and assignments of these categories to a set of training data entities, e.g., documents. Based on this knowledge supervised ML algorithms can learn to predict which category a new unobserved document belongs to (Sebastiani, 2002). Again, instead of documents also paragraphs, sentences or terms may be useful context units for classification. For QDA, e.g. sentence classification can be a worthwhile extension to manual content analysis in which human coders assign category labels to texts.
- *Named Entity Recognition / information extraction*: This application strives for the identification of person names, organizations or locations in a document. Usually, it is realized by probabilistic sequence classification determining the most probable category for

any token in a sentence (Finkel et al., 2005). For QDA, Named Entity Recognition (NER) is useful to identify actors or places associated with any other information identified in a text, e.g. certain vocabulary use, an activity or a quote. The method of sequence classification surely is not restricted to named entities. It may be applied to any other information occurring in a contextual sequence in a structural way, e.g. currency amounts or dates.

- *Sentiment Analysis*: A specific application for supervised classification is sentiment analysis, the identification of subjective information or attitudes in texts (Pang and Lee, 2008). It may be realized as a ML classification task assigning either a positive, neutral, or negative class label to a document. Another wide-spread approach is the use of so-called sentiment lexicons or sentiment dictionaries which are basically word lists with additionally assigned sentiment weights on a positive–negative scale (Remus et al., 2010).

While such applications represent deeply studied problems in NLP and computer linguistics, only few studies exist so far which apply such techniques for social science. Moreover, little knowledge exists on their systematic and optimal integration for complex analysis workflows.

## **2.3. Types of Computational Qualitative Data Analysis**

So far, the TM applications briefly introduced above have been utilized for social science purposes with varying degrees of success and in a rather isolated manner. The method debate in social science tries to identify different types of their usage by constructing method typologies. In the literature on Computer Assisted Text Analysis (CATA), several typologies of software use to support QDA 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. A very first differentiation of CATA dates back to the *Annenberg Conference on Content Analysis* in the late 1960s. There Content Analysis (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 hand (Stone, 1997). More fine grained, a famous text book on Content Analysis (CA) by Krippendorff suggests the differentiation into three types: 1. retrieval functions for character strings on raw text, 2. Computational Content Analysis (CCA) with dictionaries and 3. Computer Assisted Qualitative Data Analysis (CAQDA) for data management supporting purely manual analysis. Although published recently in its third edition (2013), it largely ignores latest developments of Machine Learning (ML). The typology from Lowe (2003) additionally incorporates computer-linguistic knowledge by covering aspects of linguistics and distributional semantics. Algorithmic capabilities are differentiated into 1. dictionary based CCA, 2. parsing approaches, and 3. contextual similarity measures. Scharkow (2012, p. 61) proposes the first typology including ML distinctively. He distinguishes three dimensions of computational text analysis: 1. unsupervised vs. supervised approaches. Within the supervised approaches he distinguishes 2. statistical vs. linguistic, and 3. deductive vs. inductive approaches. Unquestionably, this typology covers important characteristics of CATA approaches used for QDA. Yet, the assignments of single techniques to the introduced categories of his typology is not convincing in all cases. For example, he categorizes supervised text classification supporting manual CA as inductive approach (p. 89) although it is described as a process of subsuming contents into previously defined content categories. On the other hand, full-text search is categorized as deductive approach (p. 81), although it remains unclear to which extent document retrieval contributes to a deductive research design as isolated technique. Last but not least, the rather arbitrary distinction between statistical and linguistic approaches does not cover the fact that most TM applications combine aspects of both, for example in linguistic preprocessing and probabilistic modeling of content.

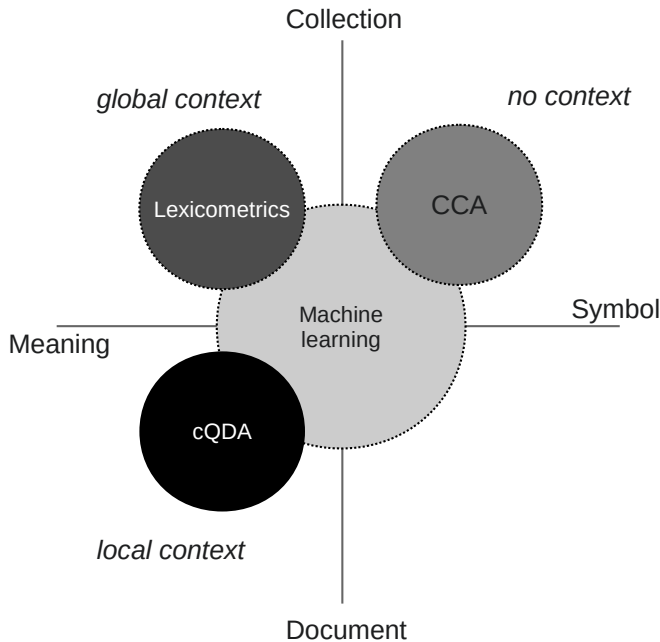
The difficulty in constructing a convincing typology for CATA is that the technical perspective and the applied social science perspective are intermingling. While the distinctions *supervised* versus *unsupervised* as well as *statistical* versus *linguistic* relate to technical aspects of NLP algorithms, the distinction *inductive* versus *deductive* captures methodological aspects. Although there might be some overlapping of category dimensions from both disciplines, they do not give guidance for clear separation.<sup>5</sup> To capture important characteristics of recent CATA approaches from an application perspective of social science research, I suggest another typology along two dimensions: complexity of meaning and textual quantity. As displayed in Figure 2.1, I distinguish between four types of CATA:

1. frequency observations of manifest expressions (fixed character strings) for CCA in large collections,
2. data management tools supporting manual coding of local contexts within single documents (CAQDA),
3. lexicometric approaches capturing aspects of (latent) meaning on a collection level, and, finally,
4. machine learning approaches incorporating characteristics of all three aforementioned types.

The horizontal dimension of this typology highlights the complexity of meaning extraction capabilities. It visualizes the progress that has been made from observation of document surfaces by simple word counts in CCA to more complex lexicometric approaches seeking to identify meaningful structures in document collections. Manually

---

<sup>5</sup>There are conceptual parallels between the pairs unsupervised/inductive and supervised/deductive with respect to usage of prior knowledge for structure identification. Nevertheless, NER, for instance, is technically realized as a supervised ML approach based on previously annotated training data. The results, however, lists of named entities associated to certain context units, can be employed methodologically in an exploratory step of QDA as well as for deductive hypothesis testing.



**Figure 2.1.:** Two-dimensional typology of analysis software for text.

conducted CAQDA, of course, strives for inference of meaningful structures identified in close reading processes on the document level. On the quantitative dimension CAQDA operates on small corpora manually manageable, while CCA and lexicometrics extract their structures from large collections. Machine learning approaches encompass an interesting intermediate position along these two dimensions as they operate on single documents and large collections at the same time by modeling single document contents with respect to collection-wide observations. This characteristic can be described further by the character of ‘context’ incorporated into the analysis.

At the beginning of the quantitative–qualitative divide, Kracauer (1952) criticized the methodological neglect of substantial meaning in quantitative CA. 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 more profound research interests. In this respect, upcoming methods of qualitative CA 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.

Since the linguistic turn took effect in social science (Bergmann, 1952), it became 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. At the same time, 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. Luckily, identifying patterns in digital data is one major strength of computers.

However, not all types of approaches are able to capture context or patterns of language use alike. Boyd and Crawford (2012) even warn that big data is losing its meaning, if taken out of context. Hence, concentrating on this important aspect, all four distinguished types of CATA capture aspects of context in their own way with severe consequences for their utilization in QDA:

1. Early approaches of Computational Content Analysis (CCA) just observed character strings in digital text for frequency analysis,

while largely ignoring context at all. More complex definitions for event observation, e.g. occurrence of term  $x$  near to term  $y$  in a distance of  $d$  or less terms, may include simple context aspects.

2. CAQDA software for *manual coding* of carefully selected small document sets allows for comprehensive consideration of linguistic and situational contexts. Understanding of expressed meaning is achievable through cognitive abilities of the human coder who also includes text external knowledge for interpretation. Analysis primarily concentrates on deep understanding of single cases through investigation of their *local context*.
3. Lexicometric applications such as key term extraction, co-occurrence analysis or LSA allow for inductive *exploration* of statistically prominent patterns of language data. Instead of local contexts in single documents, they extract *global context* observable only through examination of an entire document collection.
4. Characteristics of context extracted via Machine Learning (ML), both supervised and unsupervised, reside in an interesting middle positions between the other three types. ML works on the basis of local context by observing textual events in single documents or smaller sequences (e.g. sentences). Through aggregation and joint observation of multiple text instances, knowledge conceivable only on the collection-level is learned and incorporated into model instances representing global context. At the same time, learned global knowledge again is applied on individual text instances, e.g. by assigning globally learned categories to documents.

Through consideration of context defined as surrounding language patterns of observed events, lexicometric as well as ML approaches are able to capture more complex semantics than CCA. Combined observation of linguistic units overcomes isolated counting of overt meanings in manifest expressions. Instead, it digs down into ‘latent meaning’ either represented as statistical co-occurrence significance measure relating an observed linguistic unit to multiple other ones, or

as non-observable variables from statistical dimension reduction on observable variables. The special characteristics of ML approaches in the two-dimensional typology positioned between symbol observation and meaning extraction, as well as between document and collection level, makes it a perfect connective link to the other three CATA approaches. On the one hand, the visualization contributes to understanding why utilization of CCA and lexicometrics was spurned longtime in the QDA community, since they all operate on different levels of context and semantics. On the other hand, it helps to understand that with the advancement in ML for text, QDA is definitely confronted with a new technology bridging the gap between such formerly rather parallel developments of text analysis. ML models oscillate between local and global contexts on document and collection level to learn characteristics from individual analysis units, while also applying globally learned knowledge to them. Technically these algorithms comply with human cognitive procedures of generating textual understanding better than any prior approach.

In the following section, I will explain characteristics of these types in detail and give examples of social science studies applying such kinds of methods.

### **2.3.1. Computational Content Analysis**

Quantitative approaches of content analysis first originated 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, such as 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 proportions or quantitative developments 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 by hand. 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” (Krippendorff, 2013, p. 24). Replicability is to be achieved by determining highest possible inter- and intracoder-reliability—two metrics which calculate the matches of code assignments between several coders or the same coder in repeated coding processes.

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 (ibid., p. 19). The latter strand paved the way for the overly successful CCA software THE GENERAL INQUIRER during the 1960s (Stone et al., 1966). While neglecting implicit meaning through concentration on linguistic surfaces, CCA simply observed character string occurrences and their combinations in digital textual data. Researchers therefore create lists of terms, so called *dictionaries*, describing categories of interest. A search algorithm then processes large quantities of documents looking for those category-defining terms and in case of detection, increases a category counter. The process can be fine-tuned by expanding or narrowing the dictionary, applying pattern rules (e.g. observation of one, several or all category-defining terms; minimum  $1 \dots n$  times per document). Category counts in the end allow for assertions on the quantitative development of the overall subject-matter. Thus, developing valid dictionaries became the main task of the research process in a CCA designs.

In social science research, the method is applicable when large corpora of qualitative data need to be investigated for rather manifest content expressions. Züll and Mohler (2001) for example have used the method to summarize open questions of a survey study on the perception of aspects of life in the former GDR. Tamayo Korte

et al. (2007) evaluated tens of thousands of forum postings of a public campaign on bioethics in Germany. The project is interesting insofar as it embeds CCA in a framework of discourse analysis. The development of the categories of interest was conducted in an abductive manner. At first, recurring discourse and knowledge structures were inferred from observed lexical units inductively. These structures, operationalized as dictionaries in MAXDictio, then were tested as hypothesis against the empirical data. The project shows that CCA is not constrained to a pure nomothetic research paradigm.

Scharloth et al. (2013) have classified newspaper articles from a complete time indexed corpus of the German magazine *Die Zeit* between 1949 and 2011 by applying a dictionary approach with rather abstract categories. Using selected parts of an onomasiological dictionary, they identified and annotated the mentioning of tropic frames (e.g. health, criminality, family, virtue, 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: Their method reveals long-term developments in societal meta-discourses in Germany. At the same time, results of the rather data-driven study are hard to interpret qualitatively due to the fact that causes of the identified long-term trends remain obscure.<sup>6</sup>

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

---

<sup>6</sup>In fact, dictionary application itself cannot be considered as a data-driven approach. But selection of interesting tropic frames to describe discourse developments in the FRG was realized in a data-driven manner by ranking time series of all frames with respect to best compliance with ideal long-term trends, e.g. steady in-/decreases during the investigated time period.

Count<sup>7</sup>, 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 QDA. 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 have 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 optimum 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” (ibid., 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.).

### 2.3.2. Computer-Assisted Qualitative Data Analysis

As a counter-model to CCA and its methodological flaws, methods of QDA have emerged followed by corresponding software to support it. For this, software packages like MAXQDA, NVivo or ATLAS.ti have been developed since the 1980s. They provide functions for 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 CAQDA software 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).

---

<sup>7</sup><http://www.liwc.net>

Most of the software packages are relatively flexible concerning the research methodologies they are employed with. Early versions usually had concrete QDA methodologies in mind which should be mapped onto a program-guided process. Data representations and analysis functions in ATLAS.ti for example were mainly replicating concepts known from Grounded Theory Methodology (GTM) (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 persist in the qualitative research community (Kuş Saillard, 2011). Nonetheless, the use of CAQDA software in social science is nowadays widely accepted. Anxious debates from the 1980s and early 1990s, whether or not computers affect qualitative research negatively *per se*, have been settled. A study by Fielding and Lee (1998) 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).

The KWALON experiment conducted by the journal FQS in 2010 largely confirmed this assumption. The experiment sought to investigate the influence of different CAQDA programs on research results in a laboratory research design (same data, same questions, but different software packages and research teams). Regarding the results, Frieze (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 to predefined concepts the software advertises.

Taking context of analysis objects into account is not determined by CAQDA programs, but by the applied method. Due to its focus on support of various manual analysis steps, it is flexible in methodological regard. Linguistic context of units of interest are part of the analysis simply because of the qualitative nature of the research process itself.

Situational contexts, such as 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. However, this kind of CATA limits the researcher to a narrow corpus. Although CAQDA software guidance 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 reliability in the codification process is the objective of the other types of CATA, strongly incorporating a quantitative perspective on the qualitative data. The current method debate on CATA highlights this trade-off between qualitative deep understanding of small corpora and the rather shallow analysis capabilities of automatic big data analysis (Boyd and Crawford, 2012; Fraas and Pentzold, 2015). Taking the best of both worlds, more and more researchers advocate for combined analysis approaches of ‘close’ and ‘distant’ reading (Lemke and Stulpe, 2015; Lewis et al., 2013; Wettstein, 2014).

### **2.3.3. Lexicometrics for Corpus 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 CATA. In the late sixties, the historian Michel Pêcheux (1969) 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 its late translation in 1995 (Helsloot and 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 [...] any preliminary or arbitrary reduction of surface [...] by means of techniques of the ‘code résumé’ type is to be avoided because it presupposes a knowledge of the very result we are trying to obtain [...]” (Pêcheux et al., 1995, p. 121).

With Saussure’s 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 and Hak, 2007, § 25). In the 1970s and following decades, *Analyse Automatique du Discours* (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.

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, Matzisek, and Schirmel (2009) identify four fundamental methods of lexicometrics: 1. frequency analysis for every term of the vocabulary in the collection to identify important terms, 2. concordance analysis to examine local contexts of terms of interest,<sup>8</sup> 3. identification/measuring of characteristics of sub-corpora which are selected

---

<sup>8</sup>Results usually are returned as Key Word in Context (KWIC) lists (Luhn, 1960), which display  $n$  words to the left and to the right of each occurrence of an examined key term.

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. Dzudzek (2013) extends this catalog by applying the dimension reduction approaches Principal Component Analysis (PCA) and Correspondence Analysis on the vocabulary of an investigated corpus. By aggregating documents of one year from a diachronic corpus into meta-documents, she visualizes semantic nearness of terms as well as their correspondence with years in two-dimensional plots displaying the two principal components of the investigated semantic space.

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. Compared to CCA, the exchange of these steps in the research process allows the researcher a chance to understand how meaning is constructed in the empirical data. This makes these tools compatible with a range of poststructuralist methodological approaches of text analysis such as (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). In the Anglo-Saxon and German-speaking qualitative research community, the methodical current of Critical Discourse Analysis (CDA) has 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).

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

“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).

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. With rather linguistic research interest 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. In his view, extracted 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 Pol-Mine<sup>9</sup> which makes protocols of German federal and state parliaments digitally available and provides lexicometric analysis functions over an R interface. In a first exploratory study, Blätte (2012) investigated empirically overlaps and delimitations of policy fields with this data and compared his findings with theoretical assumptions on policy fields in political science literature. Lemke and Stulpe (2015) study the change of meaning of the political concept ‘social market economy’ in the German public discourse over the last six decades by exploring frequencies and co-occurrences of the term in thousands of newspaper articles.

Although these examples show that lexicometric approaches gain ground in QDA, they have lived a marginalized existence in the social science method toolbox for a long time. Their recent awakening largely is an effect of manageable complexity by nowadays software packages<sup>10</sup> together with the availability of long-term digital corpora allowing for tracing change of words and concepts in new ways.

---

<sup>9</sup><http://polmine.sowi.uni-due.de>

<sup>10</sup>Popular programs are for example Alceste, WordSmith or TextQuest as well as the packages *tm* (Feinerer et al., 2008) and *PolmineR* for R.



Besides the fact that no methodological standard yet exists, these methods require a certain amount of technical understanding, 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 QDA approaches. In addition to traditional manual QDA approaches, lexicometrics are able to enlighten constitution of meaning on a global context level augmenting insights from hermeneutic-interpretive analysis of single paradigmatic cases.

### 2.3.4. Machine Learning

The cognitive process of extracting information represented and expressed within texts is achieved by trained human readers very intuitively. It can be seen as a structuring process through identifying of relevant textual fragments and assigning them to predefined or newly created concepts, by and by forming a cognitive map of knowledge. Analogue to human processing, TM can be defined as a set of methods that (semi-)automatically structure very large amounts of text. ML approaches for TM brought *syntactic* and *semantic* analysis of natural language text decisive steps forward (McNamara, 2011).

Important computer-linguistic applications to identify syntactic structures are POS-tagging, sentence chunking or parsing to identify meaningful constituents (e.g. subject, predicate, object) or information extraction (e.g. NER to identify person names or locations). Sequence classification allows for analysis beyond the ‘bag of words’-assumption by taking order of terms into account through conjoint sequence observation. These computer-linguistic procedures by themselves are not really useful for QDA as single analysis. Instead, they may contribute to subsequent analysis as useful preprocessing steps to filter desired contexts by syntactic criteria.<sup>11</sup>

---

<sup>11</sup>Part-of-speech tagging for example can be utilized to filter document contents for certain word types before any subsequent TM application. Term extraction or topic models then can just concentrate on nouns or verbs, for example.

Semantic structures directly useful for QDA can be inferred by procedures of clustering and classification, e.g. to identify thematic coherences or label units of analysis with specific content analytic codes. Units of analysis can be of different granularity, e.g. single terms, phrases, sentences, paragraphs, documents or sub-collections. As introduced in Section 2.2.3, ML approaches can be distinguished in unsupervised clustering and supervised classification. ML approaches try to infer on knowledge structures interpretable as representations of global context by joint observation of the entire set of analysis units. At the same time, the learned model is applied to each individual unit of analysis, either by assigning it to a cluster or a classification category. For structure inference, not only linguistic contexts of modeled analysis units can be taken into account. Additionally, various kinds of external data might be included into models—for instance, time stamps of documents allowing for the data-driven identification of evolvment-patterns of linguistic data, or manually annotated category labels per analysis unit such as sentiment or valence scales. This interplay between local document contexts, global collection contexts together with possibilities of integrating external knowledge provides genuinely novel opportunities for textual analysis. For a few years now, pioneering studies utilizing ML have entered social science research.

### **QDA and Clustering**

Thematic structures within document collections and characteristic similarities between documents can be inferred in a purely data-driven manner by clustering algorithms. Clustering for a dedicated qualitative research interest has been employed by Janasik et al. (2009). They studied interviews conducted in a small Finnish coffee firm with self organizing maps (SOM). With the help of SOMs they visually arranged their interview data by textual similarity on a two-

---

Syntactic parsing may be utilized to identify desired subject-object relations to differentiate between certain contents dependent on word order (“In America, you watch Big Brother.” versus “In Soviet Russia, Big Brother watches you!”).

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. Methodologically, the authors argue for parallels of their approach with GTM (Janasik et al., 2009, pp. 436f).

Topic models as a variant of soft clustering have been recognized for their potential in the Digital Humanities (Meeks and Weingart, 2012), but also have received criticism from the DH community for lacking coherence and stability (Schmidt, 2012; Koltcov et al., 2014). Experience so far suggests not to apply clustering algorithms naively onto text collections, but rather to acquire decent knowledge of the algorithm along with its parameter adjustments and to critically evaluate its results. Early applications of topic models simply described topics and evaluated on thematic coherence of their highest probable terms. For example, Hall et al. (2008) investigate a large collection of historical newspapers from the USA to study topic trends over time. Another model for political science studies, incorporating authors as observed variable in addition to word usage in documents, has been introduced by Grimmer (2010). He analyzes more than 25,000 press releases from members of the US Congress. By also modeling authorship of parliamentarians, topics could be correlated with external information such as partisanship and rural versus urban election districts. Incorporating such external information allowed for a hypothesis testing research design. A more inductive study with topic models is done by Evans (2014) who analyzed US newspapers on issues denoted as “unscientific” in public discourse. A broad sample of articles selected by key terms such as “not scientific”, “non-science” etc. was clustered by a topic model revealing interpretable topics such as “evolution”, “climate change”, or “amateur sports” as issues where allegations of unscientific knowledge seem to play a major role. Slowly topic model results are not only evaluated on their own, but integrated with other TM methods for more complex analysis. With a dedicated research interest in “net policy” as an emerging policy field Hösl and Reiberg (2015) utilize topic models in combination with a dictionary approach to identify core topics with respect to their degree of politicization.

A special kind of ML clustering for political science use is dimensional scaling (Benoit and Laver, 2012) which relates texts or corresponding authors to each other on a one-dimensional scale, e.g. to determine their political left/right attitude. But, as prerequisites on text collections for valid scaling models are rather hard (collections need to be very coherent thematically) and information reduction through one-dimensional scaling is severe, benefits of methods such as *Wordscores* (Laver et al., 2003; Lowe, 2008) or *Wordfish* (Slapin and Proksch, 2008) are not clear—at least from QDA perspective targeted towards deepening of understanding instead of mere quantification.

### **QDA and Classification**

Much more useful for QDA are approaches of classification of documents, or parts of documents respectively. *Classification of documents* into a given set of categories is a standard application of media and content analysis. Methodically the combination of manual CA with supervised ML into a semi-automatic process is, for example, reflected in Wettstein (2014). Using Support Vector Machine (SVM) (2012) and Naive Bayes (2013) approaches for classification, Scharkow has shown that for simple category sets of news-article types (e.g. “politics,” “economy,” “sports,”) automatic classification achieves accuracy up to 90 % of correct document annotations. Unfortunately, conditions for successful application of classification in typical QDA environments are somewhat harder than in Scharkow’s exemplary study (see Section 3.3). Hillard et al. (2008) applied a variety of classifiers on Congressional bills for classification of 20 thematic policy issues. They also report on accuracy up to 90 % using ensemble classification with three learning algorithms (SVM, Maximum Entropy and BoosTexter). Moreover, they showed that SVM classification alone is able to predict category proportions in their data set relatively well. For semi-automatic classification of a much more complex category, ‘neo-liberal justifications of politics’ in newspaper data of several decades, Lemke et al. (2015) applied an approach of active learning within the aforementioned *ePol*-project. In iterated steps of manual annotation

followed by automatic classification, we extended an initial training set of around 120 paragraphs to more than 600 paragraphs representing our desired category. This training set provides a valid basis to measure the category in various sub-populations of complete newspaper archives. With the trained model we are able to identify trends of usage of “neoliberal justifications” in different policy fields. Exemplary studies utilizing syntactic information from parsing for classification have been conducted on large text collections as well. To extract semantic relations between political actors in Dutch newspapers, van Atteveldt et al. (2008) used a parsing model which grouped identified actors with respect to their syntactic role along with certain activities (e.g. “Blair *trusts* Bush”). Kleinnijenhuis and van Atteveldt (2014) employed parsing information on news coverage of the middle east conflict to distinguish speech acts expressing Israel as an aggressor against Palestine or vice versa.

Recently, classification of online communication such as Twitter posts became a popular field of interest especially in computational social science. For example, Johnson et al. (2011) analyzed around 550,000 twitter posts on Barack Obama and cross-correlated their findings with national survey data on popularity of the president. Their findings suggest that short term events affecting Twitter sentiments do not necessarily relate to president’s popularity in a sense of significant correlation. Tumasjan et al. (2010) classified sentiment profiles of politicians and parties of the German parliamentary elections in 2010 by analyzing sentiments in more than 100,000 Twitter posts. Surprisingly, they also claimed that mere frequency of mentioning of major parties pretty accurately predicted election results. Since then, a bunch of studies using Twitter as primary data source have been published. From QDA perspective, these early studies based on social media data are questionable, as most of them rely on overly simple categories or try to reproduce measurements formerly collected in quantitative (survey) studies. As long as they do not strive for a more complex investigation of textual meaning they do not contribute to a deeper understanding of communication contents in a qualitative sense.

But not only the result of a classification process, i.e. labels for individual documents, can be used for qualitative analysis. The global knowledge inferred from a collection incorporated in an ML model can also deliver interesting information for investigation. Pollak et al. (2011) study a document set with rule based classifiers (J48, decision tree). Their document set consists of two classes: local and international media articles on the Kenyan elections in 2008. For their analysis, they investigate the rules learned by the classifier to distinguish between the two text sets. The most discriminating features allow for intriguing insights into the differences of Kenyan news framing and its reception in the Anglo-Saxon world.

For social science purpose, Hopkins and King point to the fact that CA studies often are not primarily interested in correct classification of single documents (Hopkins and King, 2010). Instead they want to infer generalization on the whole document set like proportions of the identified categories. This introduces additional problems: “Unfortunately, even a method with a high percent of individual documents correctly classified can be hugely biased when estimating category proportions” (ibid. p. 229). To address this problem, they introduce an approach which does not aggregate results of individual document classification, but estimates proportions directly from feature distributions in training and test collections via regression calculus. With this method they measured the sentiments (five classes ranging from extremely negative to extremely positive) on more than 10,000 blog posts reporting on candidates of the 2008 US-American presidential election. Their proportion prediction is more accurate than aggregating individual classification results.<sup>12</sup> My suggested procedure for application of classification in an active learning paradigm presented in Section 3.3 also deals with the question of reliable measurement of category proportions, but further extends it to the reliable measurement of category trends.

---

<sup>12</sup>Actually, their method need severe conditions to be fulfilled to produce accurate results (ibid. 242). Complying with these prerequisites leads to the consequence that their method is not much more useful than random sampling for proportion estimation (see Section 3.3 for more information on this problem).

Text Mining for Qualitative Data Analysis in the Social  
Sciences

A Study on Democratic Discourse in Germany

Wiedemann, G.

2016, XVII, 294 p. 24 illus., Softcover

ISBN: 978-3-658-15308-3