Kritische Studien zur Demokratie

Gregor Wiedemann

Text Mining for Qualitative Data Analysis in the Social Sciences

A Study on Democratic Discourse in Germany

Springer VS

5. V-TM – A Methodological Framework for Social Science

Chapter 3 has introduced a selection of Text Mining (TM) procedures and integrated them into a complex workflow to analyze large quantities of textual data for social science purposes. In Chapter 4 this workflow has been applied to a corpus of two newspapers to answer a political science question on the development of democratic discourses in Germany. In this final chapter, I extend the workflow to a general methodological framework describing requirements, high-level design and evaluation strategies to support Qualitative Data Analysis (QDA) with the help of TM.

The method framework is drafted in analogy to concepts from the field of Software Engineering (SE) or, more specific, to Requirements Engineering (RE). In (business-oriented) computer science RE is a well-established field of research. It is defined as "a coordinated set of activities for exploring, evaluating, documenting, consolidating, revising and adapting the objectives, capabilities, qualities, constraints and assumptions that the system-to-be should meet based on problems by the system-as-is and opportunities provided by new technologies" (van Lamsweerde, 2007, p. 6). Heyer et al. (2014) suggested that employing concepts from SE and RE for Digital Humanities (DH) can help to develop interdisciplinary research designs more systematically.

Picking up this idea, I will conceptualize the methodological framework along the *V-Model* known from SE. The V-Model of the software development cycle distinguishes two phases, the verification phase and the validation phase, to coordinate development and testing activities during the entire process (Bucanac, 1991). By assigning a testing task to each development task around the actual implementation of

G. Wiedemann, *Text Mining for Qualitative Data Analysis in the Social Sciences*,
Kritische Studien zur Demokratie, DOI 10.1007/978-3-658-15309-0_5

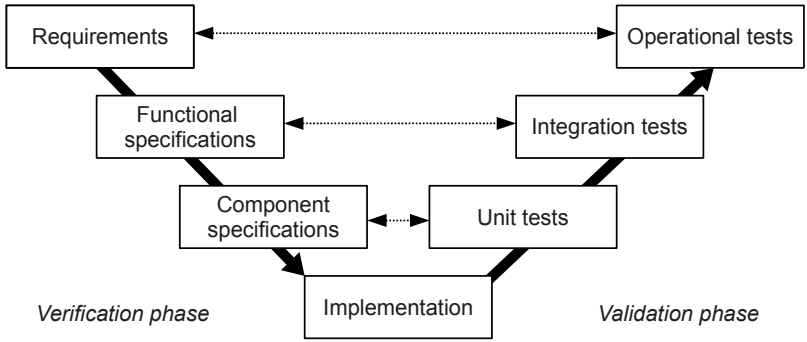

Figure 5.1.: The V-Model of software development process (Liversidge, 2005)

code, the model can be visualized in its typical V-shape (Fig. 5.1). For some years now, there has been much criticism on the V-Model and its predecessor, the *waterfall model*, from a software development perspective because of its rigidity and inflexibility (see for example Liversidge, 2005). Nonetheless, instead of utilizing it for software product management, I rely on strengths of the model as a rough guideline to conceptualize a methodological framework.

The V-Model became attractive due to clear and logic separation of specification phases as well as their close coupling with tests for quality assessment during the development process. We can adapt these ideas and modify the model for our purpose. Figure 5.2 shows the adapted version of the V-Model to guide the methodological framework for TM integration into QDA. For adaption, the verification phase is exchanged with a *research design phase* and the validation phase with an *evaluation phase*. Further, we adapt some of the specific sub-phases and their corresponding tests with processes fitting our domain. Analogue to the V-Model, *requirements* of the analysts, or in business logic terms requirements from the perspective of clients and other stakeholders, have to be identified. But instead of *opera-*

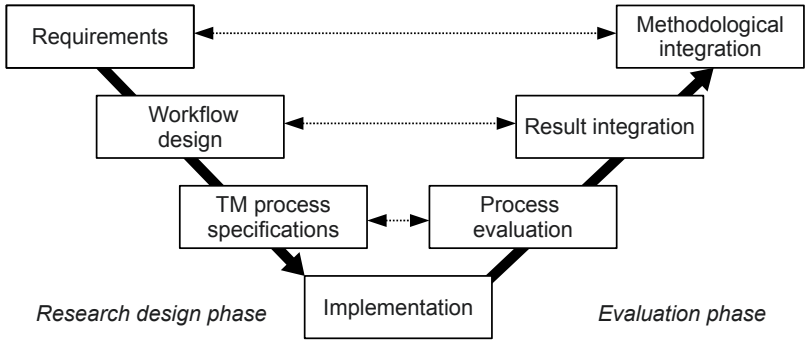

Figure 5.2.: The V-TM framework based on the V-Model to guide the methodological integration of TM into QDA. The upper two levels are discussed in this chapter while the lower two levels were already subject to discussion in Chapter 3.

tional testing in a software development process, social scientists need for integration of their study design with *methodological background* assumptions. The phase of *functional specifications*, sometimes also called high-level design, is exchanged for *workflow design* to define a sequence of analysis methods to answer a social science question. During evaluation phase, it corresponds to *result integration*, i.e. the task of compiling outcomes of different TM processes as input for subsequent processes, or in a final step summarizing outcomes for a comprehensive answer to the research question. The *component specifications* are adapted to *specifications of the selected TM processes*, each corresponding to its own specific *process evaluation* criterion.

This adaption of the V-Model will be named V-TM framework in the following descriptions. Its lower two levels *TM process specifications* and their corresponding *evaluation* followed by suggestions for concrete *implementations* have already been described intensively in Chapter 3. To complete the V-TM framework, descriptions provided in this chapter deliberately concentrate on the upper two levels of the V-

formation to elaborate on processual and methodological questions of integrating TM into QDA. In the upcoming sections, I outline on these sub-phases of the V-TM framework in more detail:

- Requirements (Section 5.1)
- Workflow design (Section 5.2)
- Result integration (Section 5.3)
- Methodological integration (Section 5.4).

5.1. Requirements

In order to create a valid and reliable research design integrating a range of computer-assisted analysis procedures for large amounts of qualitative data, we need to identify requirements first. Requirements analysis is an essential sub-task of RE to determine the needs of users, or stakeholders in general. As we deal within an interdisciplinary research environment, it is useful to reflect on two roles of stakeholders: the role of technology providers from a computer science perspective and the role of data analysts from a social science perspective—both with their own interests in research and views on the subject.

Social scientists can rely on a variety of QDA methodologies and procedures to analyze textual data such as interviews, newspapers data, or any other document manually. Surely, depending on their research background, they would like to draft TM integration into their processes as closely to their established procedures as possible. Developers of computer linguistic and NLP procedures, in contrast, have their own standards, established applications and resources to analyze textual data. Probably, both groups do not even share the same theoretical notions of semantics. Different views on the joint task, divergence between analyst requirements and technical capabilities to meet them, should be made clear in an early stage of aspired cooperation processes. Established procedures from both perspectives need to be adapted in order to fit into an integrated research design.

Requirements describe the *what* of a system, but not the *how* of its realization. In this respect, RE distinguishes into functional and non-functional requirements of a project. Functional requirements describe what a system should *do*, while non-functional requirements put emphasis on how a system should *be*. For the V-TM framework, the discipline specific notions, views and demands mentioned above mainly translate into non-functional requirements. For TM supported QDA application in scientific research domains the following **non-functional requirements** can be identified:

- *Compliance with established research quality criteria:* Participating disciplines from social science, linguistics and computer science have their own discourse on quality criteria of their research. Newly introduced methods should strive for compliance with these established criteria to increase acceptance. In quantitative social science these are basically validity, reliability and objectivity (Flick, 2007)—criteria largely compatible with numerical evaluation strategies based on established gold standards dominating in NLP. In QDA on the other hand, there are rather soft criteria such as theoretical contextualization of findings from empirical data, generalizability of findings and inter-subjective traceability of systematically applied methods (Steinke, 1999; Mayring, 2000).

- *Integration with established methodologies:* Although social science procedures of manual QDA are similar in many respects, they mostly differ in their methodological background (see Chapter 2). Integrating TM into QDA should not mean to give up on these methodological reflections. In contrast, through investigation of large amounts of texts, epistemological questions on the character of investigated speech acts as representatives of super-individual knowledge structures or social reality become even more vital.

- *Control over analysis procedures:* Algorithmic extraction of knowledge structures from text should not be a black box to the analyst. Many commercial TM software packages applied in business contexts do not disclose their algorithms and hide key parameters

to control the process in favor of usability. Analysts in research domains instead need full control of NLP preprocessing pipelines, single analysis algorithms and their key parameters, as well as chaining them into complex analysis sequences fitting to their research interest.

- *Reliability:* TM algorithms should be deterministic to get reproducible and comparable results complying to the reliability criterion of research. In fact, for many ML applications, reliability is not that easy to achieve, because they rely on random parameter initialization or sampling processes (e.g. LDA topic inference or k-means clustering). Dependent on structures within the data, these moments of chance may result in varying outcomes of the algorithms. Close investigation of the results produced by multiple runs of such algorithms may reveal, if outcomes comprise of stable structures, or instead are rather a product of chance.

- *Reproducibility:* TM workflows can get very complex quickly. Outcomes are influenced by data cleaning, linguistic pre-processing of the text, mathematical pre-processing of DTMs, key parameters of analysis algorithms and chaining of intermediate results with possible filter or selection steps in between. Further, supervised ML processes, e.g. POS-Tagging for linguistic pre-processing, may rely on models built with external training data. To achieve perfect reproducibility of results for other researchers, one would need the raw data, an exact documentation of the entire algorithmic chain and its key parameters, as well as utilized model instances.

- *Usability:* There is a danger that complexity of TM application for QDA is at the expense of the reflection on the social science aspects of the research question. Producing new, hitherto unknown types of results and nifty visualizations of extracted data does not necessarily come with deeper insight. To allow for concentration on discipline specific important aspects of the analysis work, application of TM methods for QDA should give respect to a certain degree of usability.

- *Scalability:* TM usually deals with large amounts of text. Data management procedures, workflows, analysis algorithms and result visualizations should ensure the ability to handle expected amounts of data.

This list may be incomplete. Probably, more non-functional requirements can be identified in the light of concrete analysis scenarios. Nevertheless, this list represents core requirements important to most scenarios of integrating TM into a social science research process.

What may be **functional requirements** of the research design? This highly depends on the concrete research question and its operationalization strategy. They describe which core capabilities of NLP the method design needs to provide. There is, on the one hand, a rather generic requirement of *data management* for large text collections. On the other hand, there are specific functional requirements along with specific *data analysis goals*. The next two sections briefly introduce functional requirements of both aspects.

5.1.1. Data Management

Functional requirements can be identified with respect to data management. To handle large amounts of textual data for TM processes, a system needs efficient storage and retrieval capabilities. For this, relational data bases[1], document oriented data bases[2], or a combination of both might be the right choice. Storage includes the text data itself along with all its metadata further describing the content (e.g. publisher, author, publishing date, or geo-coordinates). However, not only initial texts need to be stored. A solution is also needed for large amounts of data created during several analysis steps, which might serve as intermediate result for subsequent processes or as basis for the final analysis and visualization.

[1] Open source solutions are for example MariaDB (http://mariadb.org) and MySQL (http://www.mysql.com).

[2] Open source solutions are for examples CouchDB (http://couchdb.org) and MongoDB (http://www.mongodb.org).

For context-specific selection of documents as a starting point for any study, fast and convenient access might be provided by specialized full text indexes[3] on a full corpus collection (e.g. entire volumes of a newspaper), which allows for key term search, or more complex queries involving metadata, phrase search and term distance criteria.

To actually analyze textual data by NLP methods, text represented as character strings usually needs to be transformed into vector representations, i.e. Document-Term-Matrices (DTMs). This conversion is achieved by procedures of linguistic preprocessing resulting in a DTM object which might be transformed further by some mathematical preprocessing before computer-linguistic and statistical NLP analyses are applied. To structure and coordinate these processes, the application of an NLP framework is advised.[4] These frameworks provide functions for data reading, preprocessing and chaining of analysis tools. With their help, researchers are enabled to connect to selected data sets and quickly set up workflows for experimentation and analysis.

5.1.2. Goals of Analysis

Functional requirements also can be identified with respect to data analysis. At the beginning, there is a research question and the presumption that answers to that question can be found in a large set of documents. Requirements for analysis may be formulated along with analysis goals which arise directly from the operationalization of the research question. Operationalization of a qualitative research question with the help of TM needs to define units of analysis as a basis for structure identification. These units may be terms or phrases, concepts, actors or locations, activities, statements representing specifically defined categories of content, semantic fields, topics,

[3] Open source solutions are for example Apache Lucene/Solr (http://lucene.apache.org) and Elastic(https://www.elastic.co).

[4] Open source solutions are for example Apache UIMA (https://uima.apache.org.), the *tm*-packge for R (Feinerer et al., 2008) or the Natural Language Toolkit (NLTK; http://nltk.org) for the Python programming language.

or combinations of all these types. Goals of analysis with respect to such analysis units might be

- identification of documents relevant for the research question
- retrieving paradigmatic example texts for manual analysis
- extraction of meaningful vocabulary
- observation of semantic fields by term co-occurrences
- observation of changes in semantic fields over time
- observation of changes in semantic fields between topics, actors etc.
- observation of sentiments in documents or towards semantic units
- identification of actors or other named entities in certain contexts
- identification of topics in document collections
- identification of content analytic categories and semantic concepts
- measuring category proportions
- measuring category frequencies for time series analysis
- measuring correlation and co-occurrence of categories
- correlating quantified structures with external data.

Such analysis goals may be translated into well-defined, traceable and evaluable sub-tasks to prepare the later workflow design.

To give a more precise exemplary description, I recur to the goals of the exemplary study on democratic demarcation (see Chapter 4). The overall research question *'How was democratic demarcation performed in Germany over the past six decades?'* was operationalized two-fold: 1) an inductive step clarified on qualitative changes of topics and language use on democratic demarcation over time; 2) a deductive step measured categories of democratic demarcation derived from political science theory on political spectrums. This operationalization translated into the following analysis goals and corresponding tasks:

1. *Identification of relevant documents:* From a corpus of around 600,000 documents from two newspapers, those relevant to the topic of democratic demarcation needed to be identified. As this topic is not easily definable by a handful of key terms, an IR approach based on language use in reference documents needed to be developed. Further steps were compiling the reference corpus of paradigmatic documents, scoring relevancy of target documents by the IR approach, deciding on the size of relevant document set for subsequent analysis and evaluation of the retrieved set. Requirements of this step are described in Section 3.1.1.

2. *Exploration of contents:* To assess on contents of around 29,000 retrieved documents qualitatively, exploratory methods needed to be applied. Tasks were: identification of meaningful time periods, identification of thematic clusters, visualization of significant meaningful patterns in these thematic clusters, and extraction of paradigmatic sentences representing good examples for identified patterns. Qualitative descriptions can be retrieved from synoptic review of all partial results. Requirements of this step are described in more detail in Section 3.2.1.

3. *Manual annotation of content categories:* Statements on democratic demarcation needed to be classified by a semi-automatic process, which combines steps of manual training set creation and automatic machine classification. For the first step, five categories were derived from theoretical assumptions and defined in a code book. Afterwards, documents for initial coding needed to be selected, and then annotated manually. For manual annotation a user-friendly tool for document retrieval and coding is advised, since there are several hundred documents to read. To ensure coding quality, intercoder-reliability or intracoder-reliability has to be evaluated.

4. *Active learning of content categories:* The initial training set from manual coding is to be extended within an active learning process using supervised machine classification. The classifier needs to be chosen, feature types extracted from the training set, meaningful

features need to be selected. Then, the classifier can start to identify new meaningful text samples from the selected document set to extend the training set. In iterated steps of batch active learning suggestions for new training examples generated by the classifier need to be evaluated manually until evaluation criteria of precision, recall and size of the training set fit to required quality criteria of the process defined beforehand. Requirements of this step are described in more detail in Section 3.3.1.

5. *Final analysis:* The trained classifier from the previous step is applied to the original data set to analyze developments of the measured categories over time and in selected subsets of the data. Results from the final classification process need to be filtered by meta-data, to prepare them for visual and statistic evaluation. Visualizations for time series and co-occurrences of categories need to be provided and statistics for more comprehensive analysis should be computed. Quantified results need to be interpreted in the light of the research question and hypothesis formulated at the beginning of the study. Triangulation of the findings with external literature and qualitative investigation of the classified text units for each category help to assure the quality of the overall process chain and backup the interpretations. An example for such a comprehensive analysis is given in Section 4.3.

This list of analysis goals and corresponding tasks may serve as the basis to define specific analysis workflows in the next step.

5.2. Workflow Design

After having identified analysis goals and further refined them into tasks, we now can start to model the research design on a high level as a chain of workflows.

5.2.1. Overview

The chain of workflows realized in the example study on democratic demarcation can be seen as a specific realization of the V-TM framework to integrate several TM methods for answering a complex research question. Yet, not all steps need to be conducted to generate desired outcomes for other types of studies. Instead, single steps might be combined differently, or just can be omitted if they are considered as unimportant for the analysis. Analogue, new steps might be introduced to augment the workflow chain by other analysis approaches considered as helpful to answer the posed research question.

To visualize workflow patterns in a generic way, Figure 5.3 displays five abstract categories of previously identified analysis goals, which can be chained in varied sequential orders to realize different research designs. In accordance with the relevancy of context in CATA application (see Chapter 2), as a first step, research designs needs to conduct a certain context selection strategy to define the base population of texts for subsequent analyses. Apparently, this is a generic step to be realized by any research strategy. Subsequent steps then may be conducted to realize either inductive or deductive research strategies, or a combination of both. For this, semantic patterns representing desired units of analysis can be conceived as some kind of abstract, rather vaguely defined 'category'. Inductive exploratory analysis steps can be conducted to identify such patterns (step 2). For some research designs, qualitative description of patterns retrieved in a data-driven manner might be the final goal. Alternatively, the workflow chain can be augmented with steps of category specification to fixate identified patterns (step 3). This can be compared to the derivation of real types from empirical data, or to selective coding in GTM. Well-specified categories can be employed in a further deductive analysis step. For this, category representative analysis units can be retrieved in the entire corpus (step 4). This can be done in conjunction with or without previous data exploration and inductive category specification. In a last step of any research design, all generated categorical data from each of the conducted steps has to be evaluated in a synoptic analysis

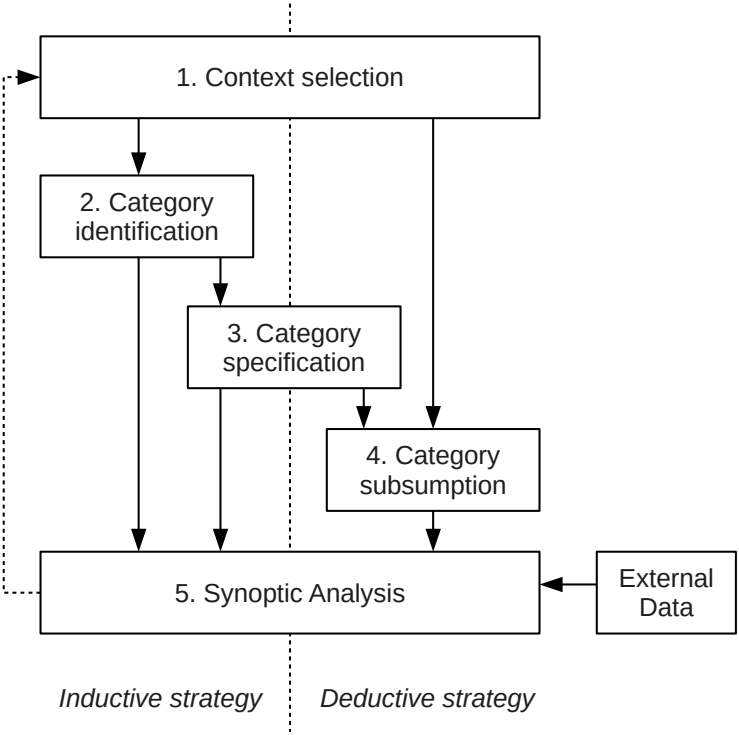

Figure 5.3.: Generic workflow design of the V-TM framework for QDA integrating Text Mining: Single steps might be realized by different technologies and under different methodological assumptions to follow either inductive or deductive research strategies, or a combination of both.

(step 5). Here also certain external data could be taken into account, e.g. for correlating patterns extracted from a document collection with text external observations.[5]

Analogue to the example study, Figure 5.4 presents the schematic visualization of a specific V-TM workflow chain. It can be perceived as an instantiation of the generic workflow design (Fig. 5.3) integrating all analysis steps, both inductive and deductive. Additionally, it shows results as resources produced by one step and utilized as input for subsequent processes to achieve following analysis goals. In accordance with the V-Model's emphasis on testing, each sub-goal needs to be evaluated separately to ensure the overall quality of results within the entire workflow chain. The principle 'garbage in, garbage out' governs any of the applied sub-processes. Serious flaws produced at the beginning of the process cannot be compensated by later processes. To assure quality of the entire analysis process, measures and procedures for quality assessment have to be distinctively defined for each goal. Defining these evaluation procedures and sufficient quality criteria is a decisive part of any workflow design. It guarantees that validity of a procedure depends on its implementation or parameters only, and not on variances in quality of its input. Consequently, the example workflow design presented in Fig. 5.4 mentions types of evaluation procedures and makes suggestions for concrete approaches.

As we are on the second level of the V-TM framework dealing with high-level research design, the following workflow descriptions operate on classes of algorithms rather than mentioning specific algorithms. Input data and outcome of results are given along with substantiated imperative descriptions of analysis tasks as steps to achieve the analysis goal. The decision for specific algorithms and their key parameters would be part of the subsequent level of TM process specifications. Nonetheless, in some cases, I refer to specific algorithmic approaches

[5]Petring (2015), for example, cross-correlates patterns in news coverage related to topics of social justice and inequality with the development of the Gini-coefficient in Germany. In a time series analysis, he can show a significant relationship between economic inequality expressed by the Gini-coefficient and increases of media reports on the topic, peaking after five months.

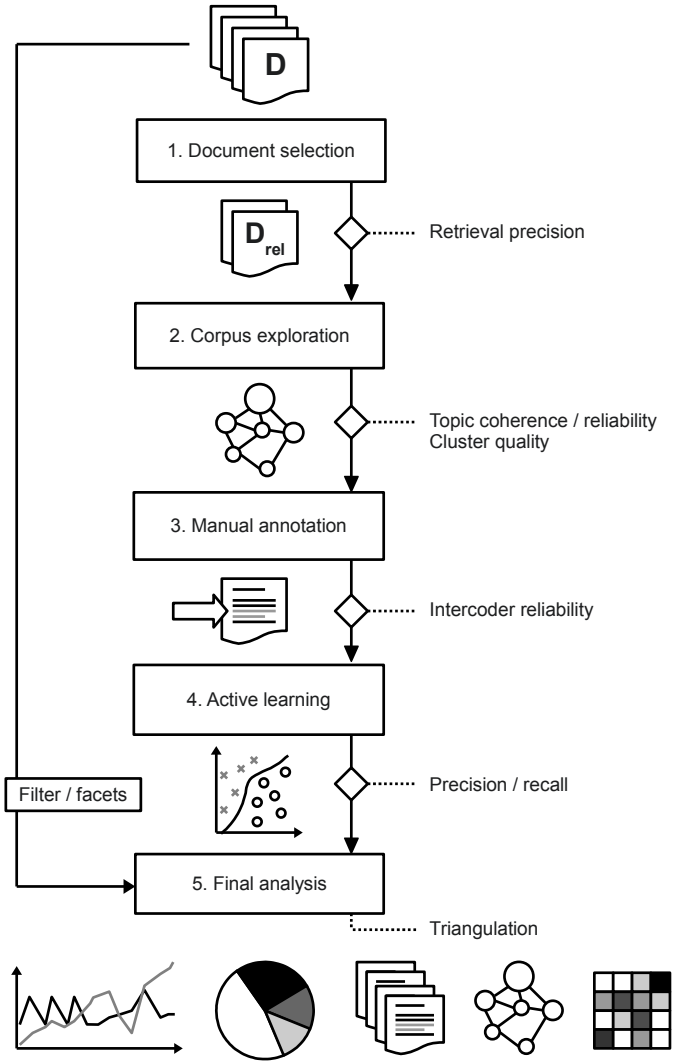

Figure 5.4.: Specific workflow design of the V-TM framework consisting of five integrated tasks. For each task, potential results are given and corresponding evaluation steps are determined to ensure validity of the results of the entire process chain.

used in the exemplary study and also mention alternatives to this design decision. Opting for such alternatives might be a good idea, if requirements change, operationalizations allow for simpler solutions or demand more complex ones.

5.2.2. Workflows

Document Selection

Workflow 1 substantiates on steps for the goal of identifying relevant documents in a population of documents $\mathcal{D}$. In the example study, scoring of relevancy is done by utilizing a contextualized dictionary of key terms extracted from a paradigmatic reference corpus $\mathcal{V}$. This reference corpus consists of five German governmental reports of the BfV on perceived threats for democracy. 'Contextualized dictionary' means the method not only relies on containment of key terms in the target collection, but rewards if key terms occur in similar contexts such as observed in the reference corpus. This proves useful in cases where target collections cannot be identified by a small set of key terms, and relevancy instead is determined by a variety of expressions and contextual language use. But surely there are research problems, where operating with a simple full text index and document selection by key term occurrence would be sufficient. Assume you want to investigate the debate on minimum wages in Germany, then simply selecting all documents containing the term 'Mindestlohn' will deliver an appropriate document set for further studies. For some research designs, retrieval may require additional filtering of certain contexts after selecting the n most relevant documents from a collection. Close reading samples may reveal that the retrieved set contains thematically relevant, yet unwanted elements with respect to the research question. In the example study, I needed to filter out documents concerned with democratic demarcation, but related to foreign countries. If one operates with key terms of ambiguous meaning, it also may be advisable to filter sub-collections with the help of a topic model. Topic models are able to identify coherent latent semantics in the

data. Hence, they may be utilized to filter (un-)desired contexts from collections.

Workflow 1: Document selection

Data: $\mathcal{D}$ – corpus of documents; n – number of documents to select

Result: $\mathcal{D}'$ – Sorted list of (potentially) relevant documents

1 Score each document $d \in \mathcal{D}$ with relevancy function.
2 Sort $\mathcal{D}$ by decreasing relevancy.
3 Return list $\mathcal{D}' = (d_1, \ldots, d_n)$ from $\mathcal{D}$ as relevant documents.
4 Option: Filter $\mathcal{D}'$ for undesired contexts, e.g. by dictionary lists or topic models.
5 Evaluate retrieval precision.

Evaluation: *Retrieval precision* of document selection should be evaluated to ensure that the selected document set contains a sufficiently large number of relevant documents. For automatic performance evaluation of a retrieval process, a gold standard of relevant documents is helpful. It can be utilized to estimate the share of relevant documents among the top ranks of the retrieved set. If a reference corpus for extraction of a contextualized dictionary is utilized, parts of this corpus could be used as a pseudo-gold set (see Section 3.1.6). Another method is suggested by Nuray and Can (2006), who propose to generate a set of pseudo-gold set by an automatic approach. They aggregate highly ranked result items found by different retrieval systems to define a set of 'pseudo-relevant' documents. The simplest and most reliable method would be to manually determine relevancy of the documents in different ranges of ranks in the retrieval list, also known as 'precision at k' (Baeza-Yates and Ribeiro-Neto, 2011, p. 140).

Corpus Exploration

Workflow 2 substantiates on steps for corpus exploration of the retrieved documents from the previous workflow. It basically relies on data-driven clustering methods for identification of time periods

and thematic structures inherent to the collection. Once thematic clusters per time period have been identified, they can be visualized in a 'distant reading' paradigm to give analysts easy access to contents of sub-collections. Section 3.2 proposes Semantically Enriched Co-occurrence Graphs for this goal. Most probable terms for each topic from an LDA topic model are observed in documents containing a topic to a substantial share in a given time period. Co-occurrence of terms in sentences of such documents is then drawn into a graph network and enriched by sentiment and term significance information. The process could be varied in multiple ways, either by relying on other topic models than LDA, other methods for clustering time periods, or modifications of the graph visualization. To provide QDA researchers with concrete text examples to anchor their interpretations, semantic patterns revealed visually on the global context level can be linked back again to the local context level (see Section 2.3). For this, sentences (or any other analysis unit) containing extracted global patterns can be retrieved as paradigmatic references for any temporal and thematic cluster.

Evaluation: As clustering is an unsupervised, data-driven process, there is not really a strict anchor for quality. For analysts it should be possible to bring identified time periods in line with their prior knowledge on important events in history. Cluster quality indices can be utilized to guide the analysist evaluation of clustering outcomes. For clustering of time periods, I utilized the Calinski-Harabasz index (Caliński and Harabasz, 1974) to determine which number of time periods yields best clustering of years. For thematic clustering, inferred topics should contain terms from coherent semantic fields. If needed, semantic coherence of topics could be evaluated experimentally in user studies (Chang et al., 2009).[6] But usually, a close qualitative investigation combined with numeric evaluation measures, such as held out likelihood (Wallach et al., 2009), would be sufficient for optimal model selection. A recommendable evaluation measure for

[6] An established approach is to deliberately insert random terms into lists of high probable topic terms. Topic coherence is high, if human evaluators are able to identify the artificially inserted word.

Workflow 2: Exploratory clustering

Data: A corpus $\mathcal{D}'$ of relevant documents with time stamps
Result: Thematic clusters of documents $d \in \mathcal{D}'$; Time periods with similar thematic structures; Semantically Enriched Co-occurrence Graphs

1 Compute thematic clusters of documents $d \in \mathcal{D}'$ (e.g. by LDA).
2 Aggregate thematic cluster assignments of documents per time span (e.g. mean topic probability per year).
3 Cluster on thematic aggregates per time span to get periods with similar thematic structures.
4 Split $\mathcal{D}'$ into subsets per thematic and temporal cluster.
5 **for** *each time period* **do**
6 Rank thematic clusters (e.g. topic probability or $rank\_1$).
7 Identify most important clusters by rank and/or manual selection.
8 **for** *m most important thematic clusters* **do**
9 Draw SECGs revealing global context patterns in documents of the current temporal and thematic cluster (for details on SECGs see Workflow 6 in the Appendix).
10 Extract samples of paradigmatic text snippets containing revealed global context patterns in their local context.
11 Evaluate cluster quality and topic model reproducibility.

topic coherence is suggested by Mimno et al. (2011) (see Eq. 3.15). Social scientists also should be careful with the reliability of topic models which may produce varying results due to random sampling for topic inference (Koltcov et al., 2014). To determine the reliability of a model, a measure for reproducibility of topics between repeated inferences can be computed by matching topic pairs (Niekler, 2016). Nevertheless, instead of relying on data-driven parameter selection and numeric evaluation only, intuition of analysts appears to be very important for adjustments of exploratory processes to generate the most valuable outcomes with respect to the analysis goal.

Manual Annotation

Workflow 3 substantiates on steps for initial document annotation with semantic categories for content analysis to prepare machine classification. Selecting documents for annotation appropriately at this point is essential for preparation of the following active learning step. Chances that machine classification predicts category trends in a valid and reliable way are much higher, if the initial training samples contain as much variety of category-related expressions as possible. If one wants to analyze documents from a time span of more than sixty years, it is advisable to rely on a training set sampled from the entire time span. Moreover, it is likely that categories occur within certain thematic contexts more frequently than in others. If the topics themselves are not highly frequent in the base population of documents, random sampling from the entire collection will deliver a poor starting point for generating a good training set. Accordingly, Workflow 3 proposes to utilize topics automatically extracted by corpus exploration in the previous step. Documents are selected by their topic probability and faceted by time period. In the example study, I identified five time periods. For each time period, I selected the 10 topics containing most interesting content with respect to the research question. Selecting the five most probable documents per topic and time period produced a list of 250 documents to read and annotate. This document set for initial annotation contains sufficient variety of category relevant language use, in both thematic and temporal manner. Context units for annotation can be of different kind, e.g. sentences, paragraphs or complete documents.

Evaluation: Manual codings of text are evaluated by *intercoder-* or *intracoder-reliability* measures. They measure agreement of codings between human coders or a single coder at different times. Established measurements in QDA are the Holsti index which just calculates the share of codings two coders agree on. Two more elaborated measures, Cohen's κ and Krippendorff's α, correct basic agreement counts for agreement by chance (Krippendorff, 2013). As a rule of thumb, it can be expected that machine classification in the next step will not

Workflow 3: Manual annotation

Data: Ranked thematic clusters of relevant documents in distinct time periods; Category system $\mathcal{C}$ for text annotation

Result: Text samples representing content categories which capture a wide range of language use determining the categories

```
1 for each time period do
2     Select the n best ranked thematic clusters.
3     for each selected cluster do
4         Select the m most representative documents (e.g. by topic
          probability).
5         for each selected document do
6             Read through document and annotate units of analysis
              representing content categories.
7 Evaluate intercoder-reliability (Cohen's κ, Krippendorff's α, ...).
```

perform better than humans do. But if categories are defined in a way that allows for reliable coding by humans, machine learning algorithms will probably be able to learn category characteristics for correct classification, too. Conversely, if humans fail to reliably identify categories, algorithms do not stand a good chance either.

Active Learning

Workflow 4 substantiates on steps for active learning of content categories in texts by supervised learning. The goal is to extend the initial training set from manual coding in the previous step with more positive and negative examples. As category systems for content analysis often are not fully complete and disjoint to describe the empirical data, we train a separate binary classifier for each category to decide whether a context unit belongs to it or not. Training examples are generated in an iterated loop of classifier training, classification of the relevant document set, selection of category candidates and manual evaluation

of these candidates. This process should be repeated until we have at least $minExamples$ positive training examples identified. It should also run at least $minIter$ times to guarantee that dubious outcomes of classification in early iteration phases are corrected. During early iterations on small training sets, one can observe that the supervised learning algorithm assumes presence of single features as absolute indicator for a specific category. Imagine the term 'right' as feature to determine statements on demarcation against far-right politics. Provided with an initial training set originating from documents of political contexts only, the classifier will learn the feature occurrence of the term 'right' as a good feature. In early active learning steps, we now can expect suggestions of sentences containing the term 'right' in the context of spatial direction or as synonym for 'correct'. Only through manual evaluation of such examples as negative candidates in ongoing iterations, the classifier will learn to distinguish between such contexts by taking dependency of the term 'right' with occurrence of other terms into account. The final training set generated by this workflow will contain valuable positive *and* negative examples to validly identify category trends. Experimentally, I identified $minIter = 6$ and $minExamples = 400$ as a good compromise between prediction performance and annotation cost (see Section 3.3).

Evaluation: Supervised classification usually is evaluated in terms of *precision*, *recall* and their harmonic mean, the F_1-measure (Baeza-Yates and Ribeiro-Neto, 2011, p. 327). To improve comparability to intercoder reliability, Cohen's κ between (human annotated) training data and machine predicted codes would also be a valid evaluation measure. As the number of positive examples often highly deviates from the number of negative examples in annotated training sets of CA categories, the application of *accuracy*, i.e. the simple share of correctly classified positive *and* negative analysis units based on all analysis units, is not advisable as evaluation measure.[7] As training

[7] Imagine a toy example of a test set containing 100 sentences, 10 belonging into the positive and 90 into the negative class. A classifier not learning any feature at all, but always predicting the negative class as outcome, would still achieve an accuracy of 90%.

Workflow 4: Active learning

Data: Corpus $\mathcal{D}'$ of relevant documents; Manually annotated samples $\mathcal{S}$ with $\mathcal{S}_+ \subset \mathcal{S}$ positive examples representative for a content category $c_p \in \mathcal{C}$

Result: *minExamples* or more positive examples $\mathcal{S}_+$ for c_p extracted from $\mathcal{D}'$; a reasonable number of 'good' negative examples $\mathcal{S}_-$ for c_p.

```
1  i ← 0
2  while |S+| < minExamples OR i < minIter do
3      i ← i + 1
4      Train machine learning classification model on S (e.g. using
       SVM).
5      Classify with trained model on U ← D' \ S.
6      Select randomly n classified results u ∈ U with
       P(+|u) >= 0.3.
7      for each of the n examples do
8          Accept or reject classifiers prediction of the class label.
9          Add correctly labeled example to S.
10 Evaluate classifier performance (F1, Cohen's κ, ...).
```

sets are rather small, it is advisable to use a process of k-fold cross validation (Witten et al., 2011, p. 152), which splits the training data into k folds, $k-1$ one for training and one for testing. Precision, recall and F_1 are then calculated as the mean of these values out of k evaluation runs, where the test set split is changed in each iteration.

Synoptic Analysis

Workflow 5 substantiates on final analysis steps incorporating results from unsupervised exploration of the retrieved relevant document set $\mathcal{D}'$ and classification of categories on the entire data set $\mathcal{D}$. For generation of final results from supervised learning, a classifier is trained with the training set $\mathcal{S}$ generated in the previous workflow

(again, separately for each category). Then, the classifier model is applied to the entire corpus $\mathcal{D}$, our initial starting point in Workflow 1. Predicted label results are assigned to each document $d \in \mathcal{D}$. Labels of the entire collection or subsets of it can be aggregated to frequency counts, e.g. per year or month, to produce time series of category developments. Subsets can be filtered beforehand by any meta-data available to a document, e.g. its publisher. Instead of category frequencies, it is advisable to use document frequencies, i.e. counting documents containing one or more positively classified context units. Document frequencies eliminate the effect of unequal category densities within documents and dampen the influence of unequal document lengths (e.g. between different publications). Further, it is advisable to normalize absolute frequency counts to relative frequencies for time series, since the original document collection may be distributed unequally over time, yielding misleading peaks or trends in the trend line. For visualization of trend lines, using smoothing of curves is advisable, because granularity of data points may produce too complex plots. To reveal more information from the category measurements, pairs of categories can be observed together on an aggregated distant level and on document level. On the distant level, Pearson's correlation between trend lines can identify concurrent, but not necessarily linked, discourse flow patterns. Beyond that, linkage of categories becomes observable by evaluating on their co-occurrence in documents. Co-occurrence can be counted as frequency, but similar to term frequencies is better judged on by a statistic, e.g. the Dice coefficient. Observing conditional probability of categories, i.e. the chance of observing B if having observed A before, can reveal even more insight on (un-)equal usage of categories in documents (see Section 4.3.3).

For synoptic analysis, findings from supervised classification of categories should be reflected in the light of the findings from exploratory analysis. Final results together with intermediate results from each workflow provide the basis for a comprehensive and dense description of the contents in qualitative as well as quantitative manner. Analogue to manual methods such as Critical Discourse Analysis (Jäger, 2004, p. 194), a synoptic analysis of textual material representative

Workflow 5: Synoptic analysis

Data: Corpus $\mathcal{D}$ of all documents; Samples of texts $\mathcal{S}$ representative for a content category $c_p \in \mathcal{C}$ retrieved by active learning

Result: Measures of trends and co-occurrence of categories in $\mathcal{D}$

1 Train ML classification models for all $c_p \in \mathcal{C}$ on S.
2 Classify each $d \in \mathcal{D}$ with the trained models.
3 Option: filter classification results on $\mathcal{D}$ by desired meta data, e.g. 1) time periods identified during exploratory clustering, 2) publication, 3) thematic cluster, or 4) mentioning of a certain actor.
4 Aggregate frequencies of positively classified context units as document frequencies by time span (e.g. years).
5 Option: Normalize absolute frequencies to relative ones. Visualize category trends as frequencies over time.
6 Count co-occurrence of c_p with other categories in documents.
7 Calculate statistic (e.g. Dice) or conditional probability of joint category co-occurrence.
8 Visualize co-occurrence statistic (e.g. as heatmap or graph network).
9 Substantiate on findings from supervised learning with those from unsupervised exploration in the previous workflow.
10 Check on findings in the light of preexisting literature or by triangulation with different QDA methods.

for relevant categories aims at providing deeper understanding of contents and underlying social formations investigated. Quantification of categories and discourse patterns allows for long term observations, comparison between categories and their dependency or relation to certain external factors. Again, the interplay of qualitative and quantitative dimensions of the retrieved data is what makes this approach appealing. A simple close reading of a sample of the extracted positively classified analysis units is very useful to further elaborate on the extracted contents. More systematically, a concept for method *triangulation* could be set up, to compare findings generated by TM supported analysis with findings made by purely qualitative research methods on (samples of) the same data (Flick, 2004).

Of course, classification of CA categories does not have to be the end of the entire workflow chain. Positively classified analysis units easily can be utilized as input to any other TM procedure. For example, automatically classified sentences representing a certain category might be clustered to get deeper insights in types of category representatives. Or documents containing a certain category might be subject to another topic model analysis to reveal more fine-grained thematic structures within a category. In some research designs it might be interesting to identify specific actors related to categories. In this case, applying a process of Named Entity Recognition (NER) to extracted context units can be a very enlightening approach to determine persons or organizations playing a vital role. Once the full range of TM application is at hand to the analyst, there is plenty of room to design extensions and new variants of the workflow chain.

5.3. Result Integration and Documentation

For quality assurance and compliance with rules of scientific work, the validity of the overall research design not only depends on isolated parts of the workflow chain, but also on their interplay. Moreover, reliability and reproducibility requires detailed documentation of analysis steps.

5.3.1. Integration

In the previous section, a TM supported research workflow design has been introduced, together with suggestions for specific algorithmic approaches and evaluation strategies. Thereby, evaluation mainly focused on the quality of encapsulated single analysis goals. In the proposed V-TM framework, such evaluations correspond to the level of unit tests in the V-Model of SE. On the next level up, workflow design corresponds with result integration during the evaluation phase (see Fig. 5.2). This level does not put emphasis on the results of analysis goals in isolated manner. Instead, it judges on the validity of combining intermediate results. Outputs of single analysis workflows often need to be filtered or transformed in a specific way to serve as input for the next workflow. Regarding this, decisions have to be made with respect to the concrete circumstances and requirements of the workflow design.

Document retrieval, for example, produces a ranked list of documents from the entire collection with respect to a relevancy scoring. If this ranking should serve as a basis to select a sub-corpus of relevant documents for the upcoming analysis, there is the need for determining a number n of documents to select from the ranking. This decision should be made carefully by evaluating and investigating the retrieval results. Dependent on the retrieval strategy, it might be absolutely valid to select the entire list containing a single key word (think again of the 'minimum wage' example). But if retrieval was performed by a large list of (contextualized) key terms producing a large list of documents, such as for democratic demarcation in the example study, clearly restricting the selection to the top ranks would be advisable.

After corpus exploration via temporal and thematic clustering of the retrieved document set, there are several ways to rank identified topics per cluster and documents within a cluster, e.g. for close reading or manual coding of categories. These rankings are not inherent to the clustering processes as such and may be even guided by researchers intuition rather than determined by data-driven numeric criteria. In this respect, specified thresholds and steps taken for selection should

be justified in a transparent way. For QDA purposes it is decisive to always maintain the connection between patterns identified on the global context level quantitatively and their underlying qualitative local contexts. To anchor interpretations based on exploratory analysis steps, it is advised to extract paradigmatic text examples containing such globally retrieved patterns. Close reading of such paradigmatic examples helps to backup interpretations qualitatively and to much better understand the contents underlying the 'distant reading' procedures.

Final evaluations based on classification of an entire population by a category system may be realized in different ways. If classification identifies sentences as representative statements of a given category, frequencies of positive sentences in all documents could be counted. By adjusting the threshold for positive label probability of the classifier, it is possible to control classified sets for precision or recall. If a study should rather concentrate on high precision of individual sentence classifications, higher thresholds for label probability might be a valid strategy to restrict outcomes.

For time series analysis, instead of sentence frequencies transformation to document frequencies might be preferred, because documents are the more natural context unit in content analysis. To restrict the final set to those documents highly representative for a certain category, it might be a valid approach to only count documents containing at least two or more positively classified sentences. At the same time, we should keep in mind that due to unequal mean lengths of articles in different publications, like in the daily newspaper *FAZ* compared to the weekly paper *Die Zeit*, higher frequency thresholds may lead to distorted measurements.

Last but not least, absolute counts preferably should be converted into relative counts, to make proportional increases and decreases of category occurrence visible independent of the data distribution in the base population. Here, different normalization strategies are applicable, such as normalization by the entire base population, by the retrieved set of relevant documents, or by the set of documents containing at least one of the categories classified. All strategies may

provide viable measures, but need to be interpreted differently. Making wrong decisions during this step may lead to false interpretations.

As briefly sketched in this section, there are many pitfalls in combining results of TM processes. Usually, there is no single best practice—only the advice to think carefully about valid solutions and provide reasonable justifications.

5.3.2. Documentation

Integration of isolated TM applications into complex workflows not only needs sound justification. To comply with demands for reliability and reproducibility, researchers need to document data inputs, chains of linguistic and mathematical preprocessing, and TM algorithms used together with settings of key parameters as detailed as possible. For complete reproducibility, it would also be necessary to provide external data utilized during the processes such as stop word lists, models utilized for tokenization and POS-tagging, etc. Strict requirements in this manner pose hard challenges to the applicability of TM methods. Complexity of the overall workflow design makes it almost impossible to completely document all decisive settings, decisions and parameters. Furthermore, there might be severe license issues concerning the disclosure of raw data, like in the case of newspaper data,[8] or issues for passing on data and models from linguistic (pre-)processing.

Hence, a complete fulfillment of the reproducibility requirement is hardly achievable, if it demands for exact reproduction of results. One possible solution could be the utilization of Open Research Computing (ORC) environments which allow for 'active documents' containing verbal descriptions of research designs and results together with scripts and raw data for evaluation.[9] Subject to the condition

[8] For this project I utilized the newspaper data of the project 'ePol – post-democracy and neoliberalism'. Unfortunately, license agreements with the publishers does not allow for data use outside the ePol project.

[9] For example, the platform 'The Mind Research Repository' (openscience.uni-leipzig.de) provides such an integrated environment of data/analysis packages along with research papers for cognitive linguistics.

that raw data can be provided together with these documents, this would allow for perfect reproducibility of published research results.

Unfortunately, until such ways of scientific publication further matured, we need to stick to verbal descriptions of workflows and parameters as systematic and detailed as possible. Hence, reproducibility as quality criterion for empirical research has to be conceptualized somewhat softer. As exact reproduction of measurements and visualizations is too strict, requirement of reproducibility should rather refer to the possibility for secondary studies to generate analysis data of the same kind as produced by the primary study. This would allow to compare whether trends and interpretations based on processed results correspond to each other. To achieve this, method standards for documentation are needed in social science disciplines. Which information at least needs to be specified depends on the concrete workflow design. For example, for linguistic preprocessing (see Section 2.2.2), this means to document the use of tokenization strategies, lower case reduction, unification strategies (e.g. stemming, lemmatization), handling of MWUs, stop word removal, n-gram-concatenation and pruning of low/high frequent terms. For co-occurrence analysis, it is the context window, minimum frequencies and the co-occurrence statistic measure. For LDA topic models, it would be the number of topics K together with the prior parameters α and η (if they are set manually an not estimated automatically by the model process).

For documentation of the realization of a specific research design, I suggest *fact sheets* as part of the V-TM framework. Such specifically formatted tables allow for a comprehensive display of

- the research question,
- the data set used for analysis,
- expected result types,
- analysis goals split up into workflows of specific tasks,
- chains of preprocssing used for task implementation,

- analysis algorithms with their key parameters, and finally,
- analysis results together with corresponding evaluation measures.

Figure 5.5 gives a (mock-up) example for the display of a V-TM fact sheet. Further method debates in social science need to be hold to determine a common set of standards and criteria for documentation and reproducibility as quality criteria.

5.4. Methodological Integration

As Chapter 2 has shown, there is definitely a growing number of studies which exploit several TM techniques for exploration and analysis of larger document collections. Some of them are designed along specific QDA methodologies, but most rather explore potentials of certain technologies, while lacking a methodological embedding. Further, up to now most studies just employ a comparatively small set of analysis techniques—if not just one single TM procedure only. To overcome the state of experimentation with single analysis procedures, the V-TM framework not only asks for integration of various procedures on the high level workflow design, but for methodological integration of the entire study design. Methodological integration not only contributes to interoperability between manual QDA methods and computer-assisted analysis, but also gives guidance for researchers what to expect from their workflow results. Alongside with identification of requirements in the research design phase (see Fig. 5.2), methodological integration of the evaluation phase asks:

1. how input data and (quantified) output of semantic structures identified by TM relate to the overall research question,
2. which parts of knowledge discovery need to be conducted in rather inductive or deductive manner, and
3. whether or how ontological and epistemological premises of the study reflect the concrete method design.

Research question	How accepted is the idea of a statutory minimum wage in the news coverage of the *Frankfurter Allgemeine Zeitung* (FAZ)?					
Data set	Set of 1.1 million FAZ articles from 1991-2015 (D)					
Expected result types	a) Semantic clusters around the minimum wage debate in Germany, 2) time series of acceptance / rejection of a statutory minimum wage (MW), 3) statistical dependency between minimum wage acceptance and employees in the low-wage sector					
Goals	**Tasks**	**Data**	**Preprocessing**	**Algorithms / Param.**	**Results + Evaluation**	**Notes**
1. Document selection	Key term retrieval	Collection of 1,1 million documents between 1991 and 2015 (D)	none	Regular expressions search	12,032 documents containing the string "Mindestlohn" were retrieved. Qualitative evaluation of 100 sample documents shows that 87% of them relate to domestic politics.	Filter for foreign affairs not needed, as 13% of documents relating to foreign countries can be tolerated.
2. Corpus exploration	1. Thematic clustering 2. Remove documents associated with bad clusters	12,032 documents containing the string "Mindestlohn" (D')	Lemmatization, stop word removal, relative pruning (Min. 1%, Max. 99%), DTM: 3012 types in 12,032 documents	LDA with K = 15, α = 0.2, η = 0.02	15 topics, 10 of them of relevance for the question, e.g. related to construction industry, unemployment or economic growth D'' = D' minus 3,201 documents mainly belonging to 5 bad topics Reproducibility of model topics: 73,3% (t = 0.2, n = 100)	LDA was fine, but let us try a non-parametric model next time
3. Manual annotation	1. identify documents to annotate 2. annotate documents	8,831 documents (D'') and Category system with 2 categories "(A) MW supporting", "(B) MW opposing"	Selection of 10 random articles from each year of the investigated time frame for annotation	none	Identification of 272 sentences for A, 301 sentences for B Intercoder reliability Cohen's κ = 0.71	Category system should be augmented by category "MW ambivalent" next time
4. Active learning	Extend training set to at least 400 documents	D'' and 272 (A) / 301 (B) positive sentences, 1926 negative sentences in 250 initially annotated documents	Stemming, no stop word removal, absolute pruning (MinFrq = 1), unigrams/bigrams DTM: 34289 types in 8,831 documents	Features with Chi-Square(c_{pl}) >= 6 SVM, C = 1 7 iterations of active learning	423 (A) / 513 (B) positive sentences, 2701 negative sentences in the final training set for the category system P = 0.70, R = 0.50, F_1 = 0.58	Reached enough examples already after 5 iterations of active learning for (B), category (A) took 7 iterations
5. Final analysis	1. Classification of categories A and B for time series 2. Cross-Correlation of A with external data	D'' and Final training data set Employees in the low-wage sector from 1991-2015	Stemming, absolute pruning (MinFrq = 1), unigrams/bigrams DTM: 34289 types in 8,831 documents	1. Features with Chi-Square(c_{pl}) >= 6, SVM, C = 10 2. Cross-Correlation	A correlates with low-wage employment rate highest in a time lag of 13 month, i.e. 13 month after an increase of employment in the low-wage sector an increase in ML approval is observable (r = 0.46).	Correlation is statistically significant ($p < 0,01$)

Figure 5.5.: Mock-up example of a V-TM fact sheet for documenting a V-TM analysis process.

The example study on democratic demarcation has shown that computer-assisted analysis of qualitative data may become a very complex endeavor. From simple counts of occurrences of character strings in single documents to complex statistical models with latent variables over huge document collections and supervised classification of hundreds of thousands of sentences, a long road has been traveled. The integration of TM revealed that nowadays algorithms are capable to extract quite a bit of meaning from large scale text collections. In contrast to manual methods of QDA, a quantitative perspective is necessarily inherent to these algorithms, either because they reveal structures in unsupervised approaches or classify large quantities of textual units in supervised approaches. Which meaning is expressed within a concrete speech act can only be understood by relating it to context, i.e. comparing it to a large set of other (linguistic) data. Human analysts in manual QDA rely on their expert and world knowledge for that, whereas computer-assisted (semi-)automatic methods need a lot of qualitative data. Thus, analyzing big data in QDA with the help of TM only makes sense as mixed method analysis combining qualitative and quantitative aspects.

In this respect: What kind of resources of large text collections are valuable resources for social science data analysis, and what kinds of research questions can be answered with them? Surely, there are collections of newswire text, as utilized in this example study, covering knowledge from the public media discourse. Other valuable resources are, for instance, web and social media data, parliamentary protocols, press releases by companies, NGOs or governmental institutions. All of these resources encompass different knowledge types and, more important, can be assigned to different actor types on different societal discourse levels. Consequently, they allow for answering of different research questions. What they all have in common is that investigation of this textual material assumes inter-textual links of the contained knowledge structures. Such links can be revealed as structures by human interpretation as well as with the help of TM algorithms.

Identification of structures also is part of any manual QDA methodology to some extent. Yet, the character of structure identification

in qualitative social research can follow different logics with respect to underlying epistemological assumptions. Goldkuhl (2012) distinguishes three epistemologies in qualitative research: 1) interpretive, 2) positivist, and 3) critical. He states, the method debate in social science mainly is concerned with the dichotomy between interpretivism and positivism:

> "The core idea of *interpretivism* is to work with [...] subjective meanings already there in the social world; that is to acknowledge their existence, to reconstruct them, to understand them, to avoid distorting them, to use them as building-blocks in theorizing. [...] This can be seen as a contrast to *positivistic* studies, which seem to work with a fixed set of variables." (ibid., p. 138)

In methodological ways of proceeding, this dichotomy translates into two distinct logics: *subsumptive* versus *reconstructive* logic. Subsumptive logic strives for assignment of observations, e.g. speech acts in texts, to categories; in other terms, "subsuming observations under already existing ones" (Lueger and Vettori, 2014, p. 32). In contrast, reconstructive logic aims for deep understanding of isolated cases by "consider as many interpretive alternatives as possible. [...] Methodologically, this means to systematically search for the various latencies a manifest expression may carry out" (ibid.). Nevertheless, even reconstructive logic assumes structure for its dense description of single cases in its small case studies, to reveal typical patterns from the language data investigated. But, in contrast to QDA in subsumptive logic, researchers do not necessarily strive for generalization of identified patterns to other cases.[10] In this characterization, both logics of QDA relate differently to *inductive* and *deductive* research designs. While reconstructive approaches always need to be inductive, subsumptive approaches may be both, either inductive by subsuming under open, undefined categories, or deductive by subsuming under

[10] For me, the main difference between the two logics seems to be the point of time for creation of types or categories during the research process. While subsumptive approaches carry out category assignments in the primary analysis phase directly on the investigated material, reconstructive approaches rather develop types on secondary data based on interpretation of the primary text.

closed, previously defined categories. This brief categorization of QDA methodologies makes clear that mainly subsumptive research logics profit from TM applications on large collections, while strictly reconstructive approaches[11] cannot expect to gain very much interesting insights.

Since TM refers to a heterogeneous set of analysis algorithms, individually adapted variants of the V-TM framework may contribute to a range of methods from subsumptive QDA. It seems obvious that computers will not be able to actually 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 analyst can rely on. But they can learn to retrieve patterns for any specific category constituted by regular and repetitive language use. Methodologies for QDA which epistemologically assume trans-textual knowledge structures observable in language patterns do have a decent chance to benefit from computer-assisted methods, if they are not shy of quantification. The integration of TM appears to be useful with qualitative methodologies such as Grounded Theory Methodology (Glaser and Strauss, 2005), Qualitative Content Analysis (Mayring, 2010), Frame Analysis (Donati, 2011), and, most promising, variants of (Foucauldian) Discourse Analysis (Foucault, 2010; Jäger, 2004; Mautner, 2012; Blätte, 2011; Bubenhofer, 2008; Keller, 2007).

Especially Discourse Analysis fits with TM because of its theoretical assumption on super-individual knowledge structures determining individual cognition and social power relations to a large extent. Michel Foucault, the french philosopher who described characteristics of his conceptualization of discourse as primary productive power for social reality, sharply distinguished between *utterance* and *statement* (Foucault, 2005). Only by repetition of utterances following certain regularities within a group of speakers, statements emerge which are able to transport social knowledge, hence, power to interfere

[11] For example, Objective Hermeneutics (Oevermann, 2002) or the Documentary Method (Bohnsack, 2010, p. 31ff)

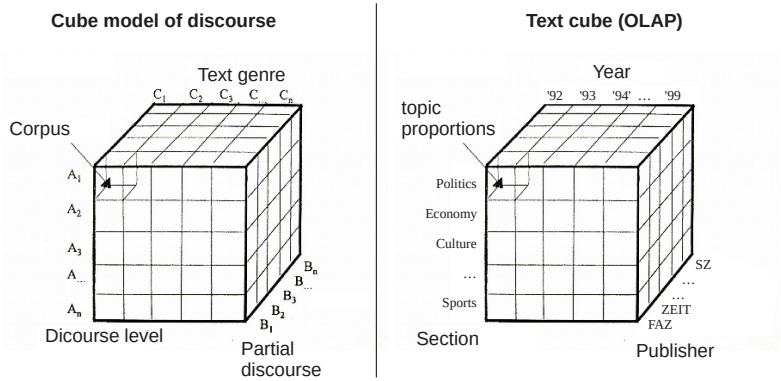

Figure 5.6.: Discourse cube model (Jung, 2011), and OLAP cube for text in the style of (Zhang et al., 2011).

in social reality. Consequently, a quantitative dimension is at least implicitly contained in Discourse Theory, wherefore it sometimes also is labeled a 'quasi-qualitative method' (Angermüller, 2005). Jung (2011) provides an interesting visual model of discourse as a cube (Fig. 5.6). It identifies corpora as slices from an entire discourse D separable along n different dimensions, for instance (topic specific) partial discourses, text genres or discourse levels (e.g. political, academic or media discourse). Every selectable (sub-)corpus comprises of certain distributions of utterances which may follow identifiable patterns, hence, statement formations following certain language regularities interesting for the analysis. By looking for patterns of language constituting important discursive statements, it would be desirable to analyze corpora along the cube dimensions. Normally, discourse analysts manually analyze small corpora of text to identify such patterns (Jäger, 2004, p. 158ff). But, being restricted to small corpora prevents to utilize the clear advantages of this systematic model.

Interestingly, the discourse cube model pretty much resembles data models of Online Analytical Processing (OLAP) cubes in data warehousing. In fact, NLP literature contains of some approaches to model text collections in OLAP cubes, which allow for computation of key

figures based on selected sub-collections by specified meta-data along cube dimensions, e.g. time, publisher or section/genre (Keith et al., 2006; Zhang et al., 2009, 2011). The task is to retrieve aggregated data (e.g. term frequencies, mean sentence length or topic proportions) from these selected sub-corpora rapidly, either computed in advance or if possible in real time. Slicing and dicing operations on such cubes allow for fast selection of sub-corpora, and more importantly, for comparison of aggregated key figures. If it is manageable to integrate meaningful data for QDA as atomic facts into such a model, it would allow for great increases of analysis capabilities through comparison of aggregated results from sub-collections easily split along multiple dimensions. The TM applications introduced and integrated in the exemplary study (see Chapter 3) provide interesting data, to be analyzed along such cube dimensions. Chapter 4 conducts analysis of measured content analytic categories together with discourse patterns along the dimensions time, publication and thematic structure. Using OLAP databases and fast indexes for text might be an interesting supplement for implementations of TM analysis infrastructures to conveniently discover semantic patterns. Certainly, equivalences between the two models highlight interesting parallels between models of (inter-)textual semantics in computer science and social science methodologies.