Cathleen Kantner and Maximilian Overbeck

Exploring Soft Concepts with Hard Corpus-Analytic Methods

Abstract: Corpus-analytic studies are currently experiencing a veritable boom in the social sciences. However, three crucial barriers impede the methodological quality and long-term reputation of these promising new technologies: Firstly, creating and pre-processing very large text corpora is still a laborious and costly enterprise. Secondly, the semantically valid operationalization of complex theoretical concepts remains a problem. Thirdly, scholars need flexible data output and visualization options to connect the data generated by computational methods with the discipline's existing research. We will conclude that it is possible to solve these problems. However, hermeneutically sensitive uses of computational methods will take much more time, work, and creativity than often assumed. The reflected appropriation of big data methods in the social sciences has only just begun.

Zusammenfassung: In diesem Beitrag argumentieren wir, dass drei methodische Barrieren die Verbreitung korpusanalytischer Methoden in den Sozialwissenschaften erschweren. Erstens bereitet es immer noch einen sehr hohen Aufwand, große Textkorpora zu erstellen und aufzubereiten. Zweitens ist das Problem der semantisch validen Operationalisierung komplexer geistes-, sozial- und kulturwissenschaftlicher Begriffe noch völlig unzureichend gelöst. WissenschaftlerInnen sind daran interessiert, über die Analyse manifester Textinhalte komplexe gesellschaftliche Sinnzusammenhänge zu rekonstruieren. Drittens erlauben viele der für linguistische Fragestellungen designten Tools kaum eine sozialwissenschaftlich anschlussfähige Ergebnisdarstellung. Die Effizienz der neuen Methoden bleibt hinter den hohen Erwartungen zurück, wenn es nicht gelingt, diese mit jedem neuen Forschungsprojekt verbundenen zeitraubenden Arbeitsschritte zu standardisieren. Allerdings kann es hierfür keine one size fits all Lösungen geben,

Note: An earlier version of this contribution in German language can be found in: Cathleen Kantner and Maximilian Overbeck (2018). "Die Analyse 'weicher' Konzepte mit 'harten' korpuslinguistischen Methoden". In: *Computational Social Science: Die Analyse von Big Data*. Ed. by Andreas Blätte, Joachim Behnke, Kai-Uwe Schnapp, and Claudius Wagemann. Baden-Baden: Nomos Verlag, pp. 163–189.

Cathleen Kantner, Department of International Relations and European Integration, University of Stuttgart

Maximilian Overbeck, Department of Communication and Journalism, The Hebrew University of Jerusalem

Open Access. © 2020 Cathleen Kantner und Maximilian Overbeck, published by De Gruyter This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International license.
<https://doi.org/10.1515/9783110693973-008>

weil aus der Perspektive unterschiedlicher wissenschaftlicher Forschungsfragen unterschiedliche methodische Entscheidungen zu treffen bleiben.

1 Introduction

The increasing accessibility of textual data, as well as numerous innovations in computational text analysis tools, are currently creating new opportunities for big data research in the social sciences.¹ Yet the great potential of corpus-analytic² methods and tools remains largely unused until now, although a growing number of researchers are interested in large or even big data³ content analyses.

In this paper, we discuss the opportunities and challenges of conducting computerized corpus-analytic research guided by complex theoretical concepts. We will call them “soft concepts” because they are structurally contested in the social scientific research community. We will put special emphasis on how to use the new, seemingly easy to use corpus-analytic methods without violating the exigent quality standards of empirical research in the social sciences. In ‘classical’, ‘qualitative’, or manual textual analysis, there are well-established best practices that comply with the methodological standards of reliability, validity, and representativeness.⁴ These standards have been established in long foundational debates and the methodological development of the discipline. For the analysis of

1 In this paper, we discuss the methodological challenges from a social sciences perspective. However, we assume that our argument is equally valid for the humanities and other research disciplines that analyze abstract theoretical concepts within larger text corpora.

2 We will use the terms “corpus-linguistic” and “corpus-analytic” instead of “computational-linguistic” since computers support the data processing, yet these methods relate to linguistic properties of human language, expressed in a *corpus* of text. We define a corpus as “a collection of spoken and written expressions. The corpus material is digitized, which means it is saved on computers and can be read by machines. The corpus elements consist of the material itself as well as metadata, describing the text material, and of linguistic annotations which were applied to the textual material” (Lemnitzer and Zinsmeister 2006, p. 8, *author translation*).

3 There does not exist a strict numerical threshold in text analysis for when to speak of small or large corpora. We will refer to a corpus as small if the documents can be manually annotated with reasonable efforts and if the researcher has an overview over the full corpus by close reading. In contrast, we refer to a corpus as large if it cannot longer be manually explored and annotated, and it therefore requires some kind of computer assisted support in order to gain an understanding of the corpus. Large corpora deal with masses of text or even big data.

4 *Reliability* deals with the question of whether data generation remains stable over time and can be precisely reproduced by different researchers. (Semantic) *validity* enquires whether the collected data actually measures the theoretically derived concepts (Krippendorff 2018, pp. 277, 361). *Representativeness*, or “sampling validity” (Krippendorff 2018, pp. 368–373) verifies whether

large amounts of textual data, however, similar ‘hard methods’ have yet to be developed.

In our research project, *eIdentity*, we created and analyzed a multilingual corpus of 460 917 newspaper articles dealing with wars and humanitarian military interventions since the end of the Cold War. We specifically focused on multiple collective identities as well as discussions on religious identities within these public debates. The analyzed corpus includes a conservative and a liberal newspaper each from six countries (Germany, Austria, France, United Kingdom, Ireland, and USA) extending over the entire period from January 1990 to December 2012. The corpus includes controversial debates about well-known conflicts like those in the former Yugoslavia, Iraq, or Libya, but also less intensively discussed conflicts, such as armed conflicts in Kashmir, Sierra Leone, or Kurdistan. For our analyses, we developed generic tools and best practices for future empirical research: the *Exploration Workbench*, designed to create and preprocess large text corpora, and the *Complex Concept Builder*, which supports researchers in their analysis and operationalization of theoretical concepts.⁵

We will argue that three crucial methodological obstacles impede the dissemination of corpus-analytic methods in the social sciences. Firstly, it still takes a lot of effort to create, manage, and preprocess large text corpora before the actual analysis can even begin. The performance of the new methods will lag behind high-flying expectations until we succeed in standardizing those time-consuming and recurring processes. Having said that, there can be no ‘*one size fits all*’ solution for corpora creation and preprocessing because scholars must adjust these processes to their respective research questions.

Secondly, adequate attention has not yet been paid to the semantically valid operationalization of theoretical concepts. Social scientists are not *per se* interested in the lexical or grammatical patterns of a text or a corpus of texts. Instead, they want to learn about society from manifest text content (Krippendorff 2018, pp. 29–40). Moreover, they are conducting their research usually from a very specific theoretical point of view within a scientific community characterized by theoretical pluralism. Therefore, social scientists search for expressions by ‘real world’

a selected set of textual data represents the whole spectrum of variability inside a given (textual) population (Biber 1993, p. 243).

⁵ For further information, see: <https://www.sowi.uni-stuttgart.de/abteilungen/ib/forschung/elidentity/> (last accessed on April 16th 2020). Follow-up projects such as PolMine and CRETA already reuse parts of our tools. Moreover, in the medium term, some of the tools will be accessible in the CLARIN-D infrastructure.

speakers in everyday language which can with good reasons count as indicators for the occurrence of *abstract theoretical* concepts (e. g. ‘national identity’).⁶

Thirdly, many of the tools designed for linguistic research are barely suitable for specific social scientific needs. Flexible data output and visualization options are required to relate corpus-linguistic data to existing research in the discipline. This is a central prerequisite for making meaningful contributions to ongoing debates and gaining acceptance for the new methods.

In this paper, we will proceed as follows: We will start by presenting three generations of text-analytic research. Every new generation improves upon the preceding one through more sophisticated research techniques. Then we focus on the youngest generation of textual analysis, as it most profoundly deals with the challenging task of combining innovative computational tools with interpretative techniques in a reflected way. This task is crucial because we could seriously harm the methodological quality and the long-term reputation of these methods in our disciplines if we do not manage to find viable solutions for *the three fundamental problems identified above*.

To date, there are a considerable number of scholars who are motivated to employ computational, corpus-analytic methods for their own research questions but who often surrender because of three seemingly insurmountable barriers. Social scientists need tools that address the three barriers and find solutions to them. We conclude that it is possible to find solutions for these challenges but, even then, using corpus-analytic methods in the social sciences will take much more time, work, and creativity than is widely expected.

2 Three Generations of Computer-Assisted Text Analysis

In comparison to other fields, social scientists are generally very open-minded about the use of new computational methods for their own research questions. This is certainly because text analysis already has a long-standing tradition in the social sciences. The increasing availability of digitized text archives and innovative computational tools has important repercussions for social science research. Simultaneously, computational methods have spread and developed in recent

⁶ One would for instance rarely express ‘national identity’ directly like “because of my Spanish identity, I...” but rather in expressions such as “Spain should...” or “our country faces the challenge of...”

Tab. 1: Three generations of computer-assisted text analysis

Generation	Corpus size	Methodology	Strengths	Problems
1st generation <i>A Good Piece of Craftsmanship</i>	Small	Manual annotation of small corpora with qualitative methods (e.g. content analysis, discourse analysis)	High validity & reliability	Low representativeness of the analyzed data (small corpora, opaque sampling) Low (inter-coder-) reliability, if interpretations are not made inter-subjectively transparent
2nd generation <i>The Lawn Mower Approach</i>	Large / big data	'Counting words' within large corpora instructed by quantitative mostly corpus linguistic methods (e.g. dictionaries, key terms, collocates, etc.)	High reliability	Low semantic validity of research instruments
3rd generation <i>The Swiss Army Knife Approach</i>	Large / big data	Combination of qualitative- hermeneutic and quantitative text analyses on large corpora via: – Corpus linguistics / text mining & discourse analysis (inductive) – Semantic field analysis (deductive) – Semi-automatic classification and text mining via machine learning algorithms	Full samples or large representative corpora Analyzing complex concepts on large or big data corpora validly, reliably and transparently	Reliability sometimes uncertain, in case of untransparent usage of complex algorithms (<i>black box</i>)

years. Overall, this has led to a steady improvement in the research instruments. In the following section, we distinguish between three generations of computer-assisted text analysis and we present their methodological strengths and weaknesses. We will assess how each generation deals with the trade-offs between the quality criteria of validity, reliability, and representativeness (see Table 1 for an overview).

2.1 The First Generation: A Good Piece of Craftsmanship

Max Weber, already in 1910, suggested “measuring with scissors and compass, how the content of newspapers has shifted [...] from what is now published as news compared to what was previously published as news” (Weber 1910, *author translation*; cf. Rössler 2010, p. 13). It was impossible, though, to carry out systematic analyses on a larger scale. Things changed in the early 1990s when the first software packages such as *ATLAS.ti* or *MAXQDA* appeared on the market. User-friendly interfaces for qualitative text analysis and manual annotation enabled scholars to use their limited resources (i. e. money, time, and annotators) more efficiently.

The first generation had direct contact with text. Manual work and attention to detail and quality characterized this approach – almost as in traditional craftsmanship. Scholars manually annotated small text corpora using qualitative content or discourse analysis methods. If annotation guidelines (codebooks) were well elaborated and pretested, coders well trained, and intercoder reliability checks performed continuously, corpus-analytic data could attain high quality in terms of validity and reliability. Many political science scholars used to apply first-generation methodological approaches within their research. Their usage can be illustrated within the particularly vivid debate on transnational public spheres (see e. g. Pfetsch and Heft (2015) for an overview).⁷

The first generation, however, faced the severe issue of sampling bias in order to get a somehow manageable amount of text. Ideally, researchers should draw a *text-sample* from the relevant corpus which represents a “down-scaled but

⁷ Further examples of first-generational research practice can be found in the publications of the Europub project (<http://europub.wzb.eu>) and the SFB 597 project “The Transnationalization of Public Spheres in the EU: Citizens’ (re)actions” (<http://www.sfb597.uni-bremen.de/pages/forProjektBeschreibung.php?SPRACHE=en&ID=9>). Moreover, the study by Rucht et al. (1992) used first-generational approaches within their exceptionally large-scale “Prodat” project (<https://www.wzb.eu/en/research/completed-research-programs/civil-society-and-political-mobilization/projects/prodat-dokumentation-und-analyse-von-protestereignissen-in-der-bundesrepublik>, last accessed on April 16th 2020).

structurally equivalent image of the basic population” (Rössler 2010, p. 58, *author translation*).⁸ Studies, however, often selected samples which were not representative of the whole spectrum of variability inside a given population. Many researchers did not even specify their selection criteria for what they called the ‘relevant’ texts. Therefore, if the sample is not representative of the investigated social phenomenon, the conclusions of the study cannot make valid claims about society.

As soon as more texts were analyzed within a study and more than one coder was annotating the texts, another problem occurred: intercoder reliability. In most projects, however, intercoder reliability was checked – if at all – superficially and only during the preparation of the coding phase. Continuous intercoder reliability tests were and are still an exemption. This limits, of course, the validity of any results if one cannot tell whether differences found are indeed in the data or rather due to different coders’ readings.

2.2 The Second Generation: The Lawn Mower Approach

Having learned from the first generation that sampling corpora for qualitative analysis carries a risk of sampling bias, second generation scholars took up the challenge of analyzing larger corpora via text statistical methods (e.g. simple word counts, key terms analyses or collocates). For their quantitative analyses, scholars often used dictionaries.⁹ Dictionaries are lists of words or combinations of words that represent an abstract concept in a text corpus and can be subsequently counted by the computer (Neuendorf and Skalski 2009, p. 207). Such term lists could be either self-generated or based on externally available sources.

The second generation, while having employed perfectly replicable methods to categorize linguistic features along social-scientific concepts, still encountered limits. Keyword lists and dictionaries have rarely been cleaned to account for word ambiguities and misleading semantic connotations.¹⁰ It was tempting to use second-generation computational tools and run them – like a lawn mower

⁸ Analysts used different sampling techniques (e.g. relevance based, full random, or snowball sampling). Sample sizes in the first generation usually vary from several hundred to a few thousands of newspaper articles.

⁹ Moreover, dimensional scaling methods such as ‘Wordscores’ (Laver et al. 2003; Lowe 2008) or ‘Wordfish’ (Slapin and Proksch 2008) have become popular methods within quantitative text analysis.

¹⁰ An exception is the study from Klüver (2009) who compared the validity of the quantitative ‘Wordfish’ and ‘Wordscores’ techniques with manual annotation on a small control sample.

– over vast amounts of digitized text data at the expense of quality. The quickly generated quantitative results were often of low semantic validity.¹¹

2.3 The Third Generation: The Swiss Army Knife

Grounded in a long-standing tradition of research experience, third generation studies are taking up the challenge of combining the assets of qualitative and quantitative text analysis. Past research has shown that they cannot build valid indicators for ‘soft concepts’ without interpretive procedures and manual annotation. Therefore, they opt for the systematic combination of quantitative and qualitative steps of analysis.

Third-generational text analysts combine, for example, quantitative corpus linguistic approaches with qualitative discourse analysis (e. g. Baker and McEnery 2005; Gabrielatos and Baker 2006; Kutter 2007; Bayley and Williams 2012). Compared to content-analytical studies, these analyses generally tend to have an inductive-exploratory character and do not attempt to operationalize complex theoretical concepts. Their aim is to identify typical linguistic patterns and to deconstruct them via in-depth discourse analysis. In comparison to such exploratory approaches, *semantic field analysis* (Kutter and Kantner 2012) is a theory driven approach. Quantitative analyses are systematically combined with qualitative steps of validation and disambiguation in order to identify relevant terms and word combinations representing complex theoretical concepts within large text corpora (for a more detailed description, see also 3.2).¹² We have now the chance to create platforms or workflows of tools that provide – like a Swiss-army-knife – various options for qualitative analysis (manual coding of codebook guided hermeneutic interpretation) as well as for quantitative, semi-automatic, or automatic categorization of text.

¹¹ For a critical stance in this regard, see Krippendorff (2018, pp. 215–276, 361–382) and Roberts (1989).

¹² Kutter and Kantner (2012) developed semantic field analysis for the examination of international discussions on wars and humanitarian military interventions within a carefully cleaned newspaper corpus of about 500 000 newspaper texts. They created semantic fields for a semi-automatic identification of discussions on international organization, or humanitarian military interventions. For further applications of *semantic field analysis* in media analysis, see Kantner, Kutter, and Renfordt (2008), Kantner (2011a,b, 2014), Overbeck (2014), and Kantner (2016).

The recent rapid computational developments have the potential to advance hermeneutic text analyses significantly. Supervised or unsupervised¹³ machine learning techniques and other algorithmic methods can be used and combined in many different ways (for a helpful overview, see Wiedemann 2016, pp. 49–54). They are currently used for example in text-clustering (Hall et al. 2008; Janasik et al. 2009; Grimmer 2010; Evans 2014), or in (semi-) automatic classification methods operating on the level of entire documents or specific text segments (see e. g. Stalpouskaya and Baden 2015).

The growing amount and increasing availability of digitized text on electronic text archives offers new opportunities for social scientists. Commercial software for text analysis is continuously improving and becoming more user-friendly and affordable (Alexa and Zuell 2000; Krippendorff 2018). Moreover, there are countless open-source applications, often created in the context of research projects or doctoral theses, or freely combinable computer-linguistic modules for self-programming in R or Python. National ministries for research and other research funding institutions are currently financing the establishment of comprehensive national and international platforms (e. g. DSPIN, CLARIN, ERIC, and DARIAH) as future research infrastructures, making cutting-edge corpus-analytic solutions permanently available for all scientific users.

Yet, why do scholars still make relatively little use of corpus analytic methods of the third generation? Why do first-generation research designs with small corpora or second generation ‘lawn mower’ approaches still dominate our discipline? Is it due to a deficiency in user friendliness, or to poor promotion?

3 Before Take-Off: Three Barriers and Prospects for Solutions

Social scientists today cannot yet exploit the full potential of third generation methodology due to three fundamental barriers:

1. Researchers tend to underestimate the considerable efforts that are necessary to create text corpora and preprocess large amounts of text. The more heterogeneous the texts and their sources (i. e., because multiple languages and

¹³ Whereas *supervised* classifications rely on training data, e. g. manually classified text segments, *unsupervised* text classifications process textual corpora automatically without training data (e. g. Pollak et al. 2011, p. 659).

- sources from different countries are compared), the more problems will occur. Moreover, there are no common software applications to support these tasks.
2. Social scientists want to learn about society, not language *per se*. Each researcher is embedded in a scientific debate in which different theoretical paradigms compete. Each researcher will certainly try to define her concepts with ‘hard’ arguments; however, those concepts will always be contested in the scientific community and insofar they will remain ‘soft’. Moreover, these abstract concepts are rarely expressed in direct terms in popular language. We cannot capture these ‘soft’ concepts by mere quantitative methods such as simple word counts or dictionary based tools.
 3. Many corpus-analytic applications and tools are limited in terms of their data output and visualization options. In order to connect corpus-analytic data to the discipline’s existing research (e. g. opinion polls or official statistics), tools must consider the specific needs for subsequent statistical analyses.

3.1 The First Barrier: Corpus Creation and Processing

The easy access to large text corpora often turns out to be a trap. The preprocessing of large corpora poses various problems before the actual analysis can even begin (Krippendorff 2018, p. 115). Yet even the best maintained electronic text archives offer only limited search and retrieval options. Keyword strategies that are based on selected search terms and simple Boolean operators inevitably contain semantic ambiguities. The common result is an abundance of *sampling errors* within the collected corpora.

Sampling errors are documents within the ‘raw corpus’¹⁴ which do not deal with the topic or the primary text genre of interest (Kantner 2016, pp. 65–67). To give an example from our project, news articles about completely off-topic issues such as sports events were erroneously included in our corpus because of their metaphorical usage of ‘war’-terms (e. g. “civil war like conditions reigned in the soccer stadium when the Bosnian team scored in the 56th minute”).¹⁵

A second source of sampling errors or “white noise” (Gabrielatos 2007, p. 6) in the raw sample is the occurrence of *duplicates*. Duplicates are fully or partly similar or identical documents that appear twice or several times in the raw corpus due to a) badly maintained archives, b) different versions of documents

¹⁴ A ‘raw corpus’ is the entirely unprocessed amount of text delivered by a certain keyword query in an electronic text archive.

¹⁵ These and other preprocessing challenges are described more in detail in Kantner, Kutter, Hildebrandt, et al. 2011 and Kantner 2016.

in the database that are not relevant to the research question, or c) complex Boolean keyword searches which have to be split into parts for searches in professional data archives like Nexis or Factiva (Kantner 2016, pp. 63–65).¹⁶ To keep these documents within a corpus would impair the empirical results' validity. No publicly available software is currently available to remove these documents. Conventional tools just assume that scholars possess a perfectly prepared and preprocessed corpus from the outset. Most researchers therefore just work with the 'raw corpus' or they write their own scripts for corpora storage, management, and preprocessing within individual databases. This limits the potential range of big data analysts to researchers who have either programming skills or are collaborating with computer scientists. The vast majority of social scientists would however welcome easily usable tools and shared best practices for corpus creation and preprocessing.

In our research, we developed innovative solutions for these recurring problems. For the removal of sampling errors, we developed an innovative procedure based on topic modeling and supervised text classification.¹⁷ There are various approaches to automatic corpora classification based on lexical features. In recent years, the so-called Latent Dirichlet Allocation method (LDA, Blei et al. 2003) has become one of the most popular usages for automatic document classification. The approach is based on the idea that each document contains a certain number of latent variables ('topics') that are themselves represented by co-occurring structures of words and phrases. Apart from determining some formal parameters, the detection of these latent variables is an entirely automatic procedure within LDA. Clustering an entire corpus into small sub collections has decisive advantages. It enables the researcher to become aware of systematic features and peculiarities within a large amount of textual data to an extent reaching beyond purely keyword-based procedures. However, the main difficulty within this approach is that the created document clusters do not represent 'topics' in the public debate's sense of the term.¹⁸ In our case, collections of names such as "*Merkel* or '*Blair*'", rubrics such as "*reviews*", as well as terms relating to places or events were among the topics.

16 Some online archives provide automatic duplicate-removal options; however, these options lack transparency since they do not define what counts as a duplicate (with respect to a specific research question) and how they automatically identify and remove them.

17 For more detailed information on our sampling cleaning process, see Blessing, Sonntag, et al. (2013) and Blessing, Kliche, et al. (2015b).

18 According to Wessler (1998, p. 666) topics or 'issues' are controversial objects of political communication that must be distinguishable from other issues, and on which the public forms opinions in the course of a debate.

Fig. 1: The e-Identity corpus cleaned issue cycle (article occurrences per month in absolute numbers, 1990–2012, N=460 917)

We first removed duplicates from the raw corpus within our *Exploration Workbench*.¹⁹ We then used LDA topic models to cluster each language-specific corpus (English, French, and German) into 200 sub collections. In a second step, trained coders manually classified²⁰ a large number of articles within each of the 200 topics as ‘relevant’ or ‘irrelevant’ according to our issue of interest (‘wars’ and ‘humanitarian military interventions’).²¹ We used these manual annotations as train-

¹⁹ The e-Identity *Exploration Workbench* has been developed for creating, managing and preprocessing large text corpora (see Kliche et al. 2014; Blessing, Kliche, et al. 2015a). It allows importation of texts of different document types, from a wide variety of sources, with diverse text and character formats, and in different languages. By selecting the relevant options from a menu, the *tool* also offers flexibility according to one’s research question. The workbench features selectable options for the duplicate removal according to individual research questions. For our project, for example, we considered identical texts as non-duplicates if they were published on different days or in different media. In order to measure the similarity of two newspaper articles, we used the methods of ‘fingerprinting’ and ‘shingling’ by Manning et al. (2008) (for more information, see Kliche et al. (2014, pp. 694–696).

²⁰ The quality, reliability, and transparency of the annotation process were ensured through a codebook with detailed annotation instructions and examples. We moreover held weekly coder trainings, and continuously verified intercoder reliability and respectively inter-annotator agreement.

²¹ For each topic, we selected articles from 20 different topic ranks (e. g. 1st, 2nd, 3rd order topic), and the whole range of topic related percentages (e. g. the whole article dealt to 70 % with a 1st topic, to 30 % with a 2nd topic etc.). With this stratified selection strategy, we expected a higher

ing data within a subsequent machine learning approach that operated on the compositions of topics within ‘relevant’ and ‘irrelevant’ articles (for more details, see Blessing, Kliche, et al. (2015a), Blessing, Sonntag, et al. (2013), and Dick et al. (2015). The quality of the semi-automatically labelled articles was then further manually validated. The final cleaned and fully preprocessed corpus of 460 917 articles (see Figure 1) was half as large as the original raw corpus (901 856 articles).

3.2 The Second Barrier: Operationalization of Complex Theoretical Concepts

Social scientists are mostly not interested in “hard facts” that could be measured by an objective uninvolved observer.²² Especially the most interesting and most intensively debated subjects of our research are, in the words of one of the French founding fathers of sociology, “*social facts*” (Durkheim 1984 [1895]). Concepts such as ‘identity’, ‘social values’, ‘the state’, ‘love’, ‘social justice’, ‘beauty’, ‘art’ and so forth do not exist outside of the practice of communication and interaction in which people think about these abstract objects, argue about them, so to speak ‘socially construct’ them, and (re-)interpret them collectively from generation to generation in discourse. In the medium of such discourses, cultures, societies, and communities reproduce themselves (Tietz 2002).

By analyzing text and text corpora, social scientists seek answers to questions about society that are relevant against the background of their scientific theories. They do not simply want to describe *how* something is communicated linguistically, but rather they observe ‘real-world’ communication processes from specific theoretical viewpoints with complex theoretical concepts in mind. Those complex concepts are usually *not* expressed directly in everyday language by non-scientists.

Since Thomas S. Kuhn (Kuhn 1962, 1970) and the post-analytical theory of science, we know that sciences do not approach truth asymptotically, but that we always engage in critique and counter-critiques of concepts and conceptual contents. This leads to a notorious pluralism of the concepts within the humanities,

chance to cover all ‘mixture’ combinations of topics as compared to a full random sampling. Rules for all coder decisions were defined in a codebook in advance.

²² Such ‘hard facts’ exist of course. They are generated via processes like demographic or socio-economic data collection and statistical analysis. However, in many cases social scientists are interested in associating ‘hard’ data (e. g., distribution of education, income, unemployment, etc.) or political preferences (e. g., survey results, election results) with their communicative context in order to understand why people act as they do.

social sciences and cultural studies, and in empirical terms, it leads to a pluralism of ways to operationalize these concepts: Terms that according to one theory might count as semantically valid indicators for, as an example, ‘national identity’ might be irrelevant or they might measure something different according to another theory.

Social scientists are usually conducting their research from a very specific theoretical point of view, or theoretical paradigms, within a scientific community characterized by theoretical pluralism. Taking a firm and innovative theoretical standpoint within these debates is one of the criteria for scientific originality in our field. This notorious pluralism needs to be translated into options for various operationalizations: a researcher studying collective ‘identity’ in a Habermasian perspective looks for different linguistic expressions than one who follows a Foucauldian perspective.

While linguists, computational linguists, or commercial operators tend towards identifying universal linguistic features within text, scholars in the humanities, social sciences, or cultural studies are instead interested in reconstructing complex social meaning through the analysis of manifest textual content (Krippendorff 2018).²³ They do not look for simple terms or term lists. The geographic concept of ‘Europe’ (as needed, e. g., for the purposes of navigation software for cars) might be operationalized by a list of words such as Paris, Rome, Helsinki, Rhine, Via Repubblica, Karpaty, Vltava, etc. Yet, for complex expressions of abstract concepts like ‘political Europe’ that are rarely expressed *directly* in everyday language we need to choose more hermeneutically sensitive approaches. Typical representations for the phenomena of interest for social scientists often are paraphrases (e. g., “our historical responsibility as Europeans”, “Brussels has to”, “Europe should”, “therefore we are obliged to”).

Moreover, one term might relate to the specific concept of interest only in a very specific context, whilst in other contexts it might not reach beyond its literal meaning. ‘National identity’ thus might be differently expressed in security policy, environmental policy, or a cultural discourse. From a social science perspective, it is therefore not practical to use standardized word lists or other prefabricated indicators.

In *elidentity*, we addressed this challenge and developed three workflows for identifying and exploring expressions of complex concepts in large corpora (see Figure 2). In the first workflow, we explore corpora by simple term or collocation

²³ Given these specific needs for analysis, the attempt by commercial software to standardize and integrate social science concepts within fixed lexica makes them intransparent and unusable (e. g. in the software *SPSS Clementine* that is very useful for many other purposes).

Fig. 2: Three different analysis workflows for sub-corpora creation

lists. This approach, although sharing some of the weaknesses of the second generation of text analysis, proves useful for an initial overview of the corpus. Simple and more complex search queries based, for example, on quantity or distance operators, can be linked to the selected search terms. Each search query can be manually checked and validated within its context in the full sentence. The results can be exported and sorted by metadata (e. g., by newspaper or data), and visually examined through processes like diachronic visualizations.

In the second workflow, we apply the methodological approach of *Semantic Field Analysis* (Kutter and Kantner 2012; Kantner 2016, pp. 67–70, 71–75) to create comprehensive lists of terms and collocations referring to a complex concept. This methodology follows a four-step approach: In the first step of concept *specification* we identify key terms and associated expressions via external sources such as encyclopedias and thesauri. In the second step of concept *identification* we trace, within the corpus of interest, lexical proxies of key terms and lexical variants. We moreover translate relevant terms from one language to another to identify additional potential terms until we reach a satisfying level of exhaustion. In the third step of concept *assignment* we apply the list to our corpus. In the fourth step of *quantitative analysis* we evaluate and statistically quantify our results.²⁴ The iden-

²⁴ Our *Complex Concept Builder* provides a tool for the creation and management of even sophisticated term lists by combining various query operators with simple or multiple terms and collocates (e. g. by connecting quantity- or distance operators with selected terms) (for more details, see Blessing, Kliche, et al. (2015a), Blessing, Sonntag, et al. (2013)). The tool moreover offers the possibility to examine every single query concerning its semantic validity in a sentence context (via KWIC-lists or whole sentences). In the data export, the results can be sorted by metadata (e. g. name of newspaper or date) and visualized in time series analyses.

tification and disambiguation of semantically valid indicators usually turns out to be an extremely difficult and time-consuming enterprise. Sometimes even the most sophisticated term-lists will not suffice to catch the exact meaning of a concept of interest – some terms are inherently ambiguous.

To deal with such semantically ambiguous terms, we therefore used, as a third workflow, the previously presented semi-automatic approach for removing sampling errors (see Section 3.1). This approach has the advantage of considering a large amount of semantic context. It proved suitable to create tailored subsamples of articles dealing explicitly with, for example, the role of religious identities in armed conflicts (Overbeck 2015).

The automatic clustering procedure of the overall corpus through LDA topics supports the difficult and ambitious task to identify the relevant semantic fields as indicators for the complex theoretical concepts.²⁵ The manual validation and classification of ‘relevant’ and ‘irrelevant’ articles subsequently helps to sort out articles that might contain religious keywords but do not explicitly address the role of religious identities in armed conflicts. The supervised machine learning procedure eventually helps to identify similar ‘relevant’ articles in the overall corpus of yet un-annotated articles.

In applying these three workflows, highly specific sub-corpora containing *only those texts* of the cleaned corpus *speaking to our theoretical research question* could be created and further be examined via qualitative content analysis, as practiced by the first generation.²⁶

3.3 The Third Barrier: Enabling Suitable Result Visualization for Social Scientists

Quantitative text analyses – in one way or another – result in data. Once the textual analysis is completed, social scientists require flexible data arrangement and visualization options organized along the units of analysis that are relevant for their research questions and their formulated hypotheses. Sometimes we make statements about the whole corpus, sometimes about different countries or actors or (sub-)issues; sometimes we compare countries or aggregated results along time

²⁵ We integrated filter options which helped to identify the most ‘relevant’ topics for one’s research question via key term searches.

²⁶ We coded, e. g., references to ‘multiple collective identities’ in a randomly drawn subsample of the cleaned corpus ($n = 6539$). Based on the experience of previous projects, we generated a detailed codebook and performed continuous intercoder reliability tests on every eleventh newspaper article (cf. Renfordt 2011, pp. 94–101).

periods. The data output of many, even widely used commercial software packages, however, is often based on linguistically relevant units of analysis (e. g., the whole corpus, specific words, collocates, phrases, sentences).

Such outputs however are rarely interpretable for social scientists, and they cannot be directly connected to other data in social sciences (e. g., event-statistics, socio-demographics, public opinions, election results, etc.). As shown in the previous parts, social scientists look for textual references of their relevant complex concepts and subsequently count them in relation to different units of analysis. Social scientists *describe* the social world (e. g., as mirrored in texts), but they also want to *explain* it. We want to make claims about relations between variables. The indicators for these variables are to be found in different sources such as meta-data (e. g. dates, geographic locations, names of newspapers), lists of terms or collocates as generated by second-generation quantitative approaches, or counts of occurrences (per texts, text passages, or other text units) of sophisticated measures for complex theoretical concepts.

During statistical analysis, these units of analysis must often be aggregated to larger dimensions (e. g. time periods, geographic locations, newspaper types). Yet the data output of many computational tools often does not allow compiling the data according to these needs. It is also often very difficult to switch from one data compilation (e. g., how often were religious identities discussed in different countries per year?) to another one (e. g., how big is the share of articles addressing religious identities in the European newspapers as compared to the American papers?).

By providing flexible data output options, tools should enable researchers to connect various variables with each other. Of course, the statistical analysis can be done with common statistics software; however, it should become easier to import text analytical data into these packages. Since every empirical sub-question together with its theoretically informed definition of dependent and independent variables requires a different kind of data arrangement, tools should provide very flexible options for data output and visualization.²⁷

²⁷ The *Complex Concept Builder* developed in *eIdentity* provides flexible data output (e. g., as CSV or Excel files which can be easily imported into SPSS or R for statistical analysis) and visualization options. Annotated data can be aggregated on different levels (e. g., temporal units such as days, months, years) and units of analysis (e. g., text, segment, sentence-level), and filtered by specific sub-samples (e. g., newspaper, language, country).

4 Conclusion

Because it seems so simple to ‘run’ innovative computational tools on millions of electronically available text data, it is tempting to use these procedures with great enthusiasm but little care. Especially in this take-off phase of big (textual) data research in our discipline, social scientists need to promote and defend the hard-won methodological standards of their scientific discipline. The analysis of large corpora will not, *per se*, lead to more knowledge or methodological progress. If the current boom of computer-linguistic methodology in the social sciences shall not turn into great disappointment soon, we must develop best practices of theory-guided empirical research that can significantly contribute to important scientific debates. For the application of computational tools and methods in the social sciences this necessarily involves time-intensive steps of manual annotation and qualitative validation, and a reflective selection of suitable corpora and appropriate (computational) methods.

In this contribution, we outlined the establishment of quality standards and best practices in reliable and valid text analysis in the social sciences since the 1990s. By focusing on the third, still ongoing phase of textual analysis, we argued that three hurdles impede the proliferation of quality-oriented corpus-analytic research in the social sciences. Flexible generic tools should standardize recurring tasks in order to make them much easier and more efficient, particularly during corpus creation, cleaning, and preprocessing. Our overall suggestion is to develop methods and workflows that systematically combine corpus-linguistic and qualitative steps of analysis to encourage the very individual theory driven operationalization of complex social scientific concepts in semantically valid ways. The better the new tools respond to the needs of flexible data-export and -output options for subsequent statistical analyses, the more these new methods will be able to prove themselves in discipline specific discourse and position themselves at eye level with established methods.

Funding: This article results from the collaborative interdisciplinary research project “Multiple Collective Identities in International Debates on War and Peace since the End of the Cold War” (e-Identity) that has been conducted by Prof. Dr. Cathleen Kantner, Prof. Dr. Jonas Kuhn, Prof. Dr. Manfred Stede and Prof. Dr. Ulrich Heid (funding period 2012–2015, funding code: 01UG1234A-C), as well as the “Center for Reflected Text Analytics” (CRETA, 2016–2020, funding code: 01UG1601). We would like to thank the Federal Ministry of Education and Research (BMBF) for the generous support of both projects. Last but not least, we want to thank Caitlin Reynolds for the meticulous and insightful language editing.

References

- Alexa, Melina and Cornelia Zuell (2000). "Text analysis software: Commonalities, differences and limitations: The results of a review". In: *Quality and Quantity* 34 (3), pp. 299–321.
- Baker, Paul and Tony McEnery (2005). "A corpus-based approach to discourses of refugees and asylum seekers in UN and newspaper texts". In: *Journal of Language and Politics* 4 (2), pp. 197–226.
- Bayley, Paul and Geoffrey Williams (2012). *European identity: what the media say*. Oxford; New York: Oxford University Press.
- Biber, Douglas (1993). "Representativeness in corpus design". In: *Literary and linguistic computing* 8.4, pp. 243–257.
- Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003). "Latent dirichlet allocation". In: *The Journal of Machine Learning Research* 3, pp. 993–1022.
- Blessing, André, Fritz Kliche, Ulrich Heid, Cathleen Kantner, and Jonas Kuhn (2015a). "Comput-erlinguistische Werkzeuge zur Erschließung und Exploration großer Textsammlungen aus der Perspektive fachspezifischer Theorie". In: *Grenzen und Möglichkeiten der Digital Humanities. Sonderband der Zeitschrift für digitale Geisteswissenschaften*. Ed. by Constanze Baum and Thomas Stäcker. Vol. 1. doi: 10.17175/sb001_013.
- Blessing, André, Fritz Kliche, Ulrich Heid, Cathleen Kantner, and Jonas Kuhn (2015b). "Die Exploration großer Textsammlungen in den Sozialwissenschaften". In: *CLARIN Newsletter* 2015 (8), pp. 17–20.
- Blessing, André, Jonathan Sonntag, Fritz Kliche, Ulrich Heid, Jonas Kuhn, and Manfred Stede (2013). "Towards a tool for interactive concept building for large scale analysis in the humanities". In: *Proceedings of the 7th Workshop on Language Technology for Cultural heritage, Social Sciences, and Humanities*, pp. 55–64.
- Dick, Melanie, André Blessing, and Ulrich Heid (2015). "Automatische Verfahren zur Bewertung der Relevanz von Dokumenten für geisteswissenschaftliche Forschungsfragen". In: *Abstracts der DHd: Von Daten zu Erkenntnissen*. Graz.
- Durkheim, Emile (1984 [1895]). *Die Regeln der soziologischen Methode*. Frankfurt: Suhrkamp.
- Evans, Michael (2014). "A computational approach to qualitative analysis in large textual datasets". In: *PloS one* 9 (2), pp. 1–10.
- Gabrielatos, Costas (2007). "Selecting query terms to build a specialised corpus from a restricted-access database". In: *ICAME Journal* 31, pp. 5–43.
- Gabrielatos, Costas and Paul Baker (2006). "Representation of refugees and asylum seekers in UK newspapers: Towards a corpus-based analysis". In: *Joint Annual Meeting of the British Association for Applied Linguistics and the Irish Association for Applied Linguistics (BAAL/IRAAL 2006): From Applied Linguistics to Linguistics Applied: Issues, Practices, Trends*.
- Grimmer, Justin (2010). "A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases". In: *Political Analysis* 18 (1), pp. 1–35.
- Hall, David, Daniel Jurafsky, and Christopher D. Manning (2008). "Studying the History of Ideas Using Topic Models". In: *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Honolulu, Hawaii: Association for Computational Linguistics, pp. 363–371. URL: <https://www.aclweb.org/anthology/D08-1038> (visited on June 1, 2020).

- Janasik, Nina, Timo Honkela, and Henrik Bruun (2009). "Text mining in qualitative research application of an unsupervised learning method". In: *Organizational Research Methods* 12 (3), pp. 436–460.
- Kantner, Cathleen (2011a). "Debating humanitarian military interventions in the European public sphere". In: *RECON Online Working Paper* 2011 (30), pp. 1–19.
- Kantner, Cathleen (2011b). "European Identity as *Commercium* and *Communio* in Transnational Debates on Wars and Humanitarian Military Interventions". In: *RECON Online Working Paper* 2011 (37), pp. 1–22.
- Kantner, Cathleen (2014). "The European public sphere and the debate about humanitarian military interventions". In: *European Security* 23 (4), pp. 409–429.
- Kantner, Cathleen (2016). *War and Intervention in the Transnational Public Sphere: Problem-solving and European identity-formation*. London: Routledge.
- Kantner, Cathleen, Amelie Kutter, Andreas Hildebrandt, and Mark Püttcher (2011). "How to get rid of the Noise in the Corpus: Cleaning Large Samples of Digital Newspaper Texts". In: *International Relations Online Working Paper* 2011 (2), pp. 1–23.
- Kantner, Cathleen, Amelie Kutter, and Swantje Renfordt (2008). "The Perception of the EU as an Emerging Security Actor in Media Debates on Humanitarian and Military Interventions (1990-2006)". In: *RECON Online Working Paper* 2008 (19), pp. 1–24.
- Kantner, Cathleen and Maximilian Overbeck (2018). "Die Analyse 'weicher' Konzepte mit 'harten' korpuslinguistischen Methoden". In: *Computational Social Science: Die Analyse von Big Data*. Ed. by Andreas Blätte, Joachim Behnke, Kai-Uwe Schnapp, and Claudius Wagemann. Baden-Baden: Nomos Verlag, pp. 163–189.
- Kliche, Fritz, André Blessing, Jonathan Sonntag, and Ulrich Heid (2014). "The e-identity exploration workbench". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC'14*, pp. 691–697.
- Klüver, Heike (2009). "Measuring interest group influence using quantitative text analysis". In: *European Union Politics* 10 (4), pp. 535–549.
- Krippendorff, Klaus (2018). *Content Analysis: An Introduction to Its Methodology (4th Edition)*. Thousand Oaks, CA: Sage.
- Kuhn, Thomas S. (1962). *Die Struktur wissenschaftlicher Revolutionen*. Frankfurt am Main: Suhrkamp.
- Kuhn, Thomas S. (1970). "Reflections on my critics". In: *Criticism and the Growth of Knowledge*. Ed. by Imre Lakatos and Alan Musgrave. Cambridge: Cambridge University Press, pp. 231–278.
- Kutter, Amelie (2007). "Petitioner or partner? Constructions of European integration in Polish print media debates on the EU Constitutional Treaty". In: *Discourse and Contemporary Social Change*. Ed. by Norman Fairclough, Giuseppina Cortese, and Patrizia Ardizzone. Bern: Peter Lang, pp. 433–457.
- Kutter, Amelie and Cathleen Kantner (2012). "Corpus-Based Content Analysis: A Method for Investigating News Coverage on War and Intervention". In: *International Relations Online Working Paper* 2012 (01), pp. 691–697.
- Laver, Michael, Kenneth Benoit, and John Garry (2003). "Extracting Policy Positions from Political Texts Using Words as Data". In: *American Political Science Review* 97 (2), pp. 311–331.
- Lemnitzer, Lothar and Heike Zinsmeister (2006). *Korpuslinguistik: Eine Einführung*. Tübingen: Gunter Narr.
- Lowe, Will (2008). "Understanding wordscores". In: *Political Analysis* 16 (4), pp. 356–371.

- Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze (2008). *Introduction to Information Retrieval*. New York, Cambridge: Cambridge University Press.
- Neuendorf, Kimberly A. and Paul D. Skalski (2009). "Quantitative Content Analysis and the Measurement of Collective Identity". In: *Measuring Identity: A Guide for Social Scientists*. Ed. by Rawi Abdelal, Yoshiko M. Herrea, Alastair I. Johnston, and Rose McDermott. Cambridge: Cambridge University Press, pp. 203–236.
- Overbeck, Maximilian (2014). "European debates during the Libya crisis of 2011: shared identity, divergent action". In: *European Security* 23 (4), pp. 583–600.
- Overbeck, Maximilian (2015). "Observers turning into participants: Shifting perspectives on religion and armed conflict in Western news coverage". In: *The Tocqueville Review/La revue Tocqueville* 36 (2), pp. 95–124.
- Pfetsch, Barbara and Annett Heft (2015). "Theorizing communication flows within a European public sphere". In: *European Public Spheres. Politics is Back*. Ed. by Thomas Risse. Cambridge: Cambridge University Press, pp. 29–52.
- Pollak, Senja, Roel Coesemans, Walter Daelemans, and Na-da Lavrač (2011). "Detecting contrast patterns in newspaper articles by combining discourse analysis and text mining". In: *Pragmatics*, pp. 674–683.
- Renfordt, Swantje (2011). *Framing the Use of Force: An International Rule of Law in Media Reporting. A Comparative Analysis of Western Debates about Military Interventions, 1990-2005*. Baden-Baden: Nomos.
- Roberts, Carl W. (1989). "Other than counting words: A linguistic approach to content analysis". In: *Social Forces* 68 (1), pp. 147–177.
- Rössler, Patrick (2010). *Inhaltsanalyse*. Stuttgart: UTB.
- Rucht, Dieter, Peter Hocke, and Thomas Ohlemacher (1992). *Dokumentation und Analyse von Protestereignissen in der Bundesrepublik Deutschland (Prodatt): Codebuch*. Berlin: WZB, Abt. Öffentlichkeit und Soziale Bewegung.
- Slapin, Jonathan B. and Sven-Oliver Proksch (2008). "A scaling model for estimating time-series party positions from texts". In: *American Journal of Political Science* 52 (3), pp. 705–722.
- Stalpouskaya, Katsiaryna and Christian Baden (2015). "To Do or Not to Do: the Role of Agendas for Action in Analyzing News Coverage of Violent Conflict". In: *Proceedings of the First Workshop on Computing News Storylines*. Beijing, China: Association for Computational Linguistics, pp. 21–29. doi: 10.18653/v1/W15-4504.
- Tietz, Udo (2002). *Die Grenzen des "Wir". Eine Theorie der Gemeinschaft*. Frankfurt am Main: Suhrkamp.
- Weber, Max (1910). "Geschäftsbericht auf dem ersten Deutschen Soziologentage in Frankfurt 1910". In: *Gesammelte Aufsätze zur Soziologie und Sozialpolitik*. Ed. by Marianne Weber. Tübingen: JCB Mohr, pp. 431–449.
- Wessler, Hartmut (1998). "Issue". In: *Politische Kommunikation in der demokratischen Gesellschaft. Ein Handbuch mit Lexikonteil*. Ed. by Otfried Jarren, Ulrich Sarcinelli, and Ulrich Saxer. Vol. 1. Opladen: Westdeutscher Verlag, p. 666.
- Wiedemann, Gregor (2016). *Text Mining for Qualitative Data Analysis in the Social Sciences*. Wiesbaden: VS Verlag.