Corpus analysis

Amelie Kutter

Introduction

The past decade has seen a surge in digitisation. Web-based access to digital sources has widened, coupled with advances in general-purpose technologies for the automated processing of digital documents. Searches in web archives, networking in social media, marketing targeting web users, virtual financial transactions, or international cyber intelligence and policing: contemporary social practices rely on digital documents and their automated processing. This is also true for politics and the involvement of language in political struggles. Politics is more radically being subjected to the logics and technological trends of mass mediation. It comes as no surprise then that corpus analysis, with its potential to work with large numbers of digital texts, has become popular beyond the confines of computational and corpus linguistics. Notably, in linguistic discourse studies, using collections of digital texts and specialised analytical software has become something of a must. Pushed by new technologies, traditions of automated textual analysis, such as computer-aided content analysis and lexicometrics, have been revived in the social sciences, too. Corpus analysis, more precisely: the semi-automated analysis of numerical-linguistic patterns in a specified text collection, promises to link these traditions to current technological and social trends.1

The popularity of corpus-analytical techniques has revived controversies about the advantages and drawbacks of computer-aided textual analysis that have structured the field ever since digital documents were first processed by computers. Current debates centre on the opposition between ‘corpus-driven’ and ‘corpus-based’ approaches, or ‘unsupervised’ and ‘supervised’ approaches. Corpus-driven approaches entrust lexicometric or corpus linguistic software with the task of revealing patterns of language use that may be associated with social meaning, while corpus-based approaches insist that such investigations should be theory-driven, rather than technology- or data-driven (Gür-Şeker 2014). Unsupervised approaches expect recent advances in text mining to generate the artificial intelligence necessary for automated text comprehension, while supervised approaches doubt that any meaningful interpretation can be made based on text mining unless this is consistently controlled by human interpretive effort (Scharkow 2012 in Wiedemann 2013). These labels,
apparently relating to the use of technology, involve tacit positioning alongside established divides in the philosophy of science (see the sub-section on ‘Dilemmas of conducting corpus research’ on p. 172). They show that corpus analysis has become subject to struggles for scholarly authority.

The objective of the present contribution is to give an overview of the field of corpus analysis so that entering the field and navigating field-specific controversies become easier. Readers will get to know the plurality of approaches to corpus analysis, familiarise themselves with the specificities of text-statistical methods and learn about possible applications, all illustrated by examples from a study on crisis discourse (Kutter 2013a). The contribution argues that corpus analysis is a valid addition to the tool-box of the discourse researcher, in particular, when used as an explorative technique for heuristic and reflexive purposes. Corpus analysis enables a specific macro-view of text material and supports the development of questions and hypotheses (heuristic potential). It also productively disturbs our intuitive and preconceived views of the subject studied and, thereby, contributes to a multi-layered interpretation of language and meaning (reflexive potential). In the following sections, the reader will learn how to exploit these potentials for political discourse studies. Information is provided on approaches and dilemmas of corpus research, specificities of lexicometrics and corpus linguistics, and the development of a discourse-analytical strategy involving corpus analysis.

**Corpus analysis: a multi-disciplinary and difficult field or research**

Corpus analysis has been emerging since the 1960s, in parallel with the digitisation of documents and computation. It is now an independent field of research that deals with the **semi-automated statistical analysis of patterns of language use** that reveal themselves in a specific corpus. In general terms, a corpus is a collection of semiotic artefacts, whether of linguistic, or other quality, that are the subject of interpretive enquiry. In the more specific terms of corpus analysis, a corpus is a collection of digital texts that has been compiled and annotated according to specific criteria with a view to statistically analysing linguistic, or other, characteristics of the whole collection.

**Approaches to corpus analysis**

Corpora are used in a growing number of sub-disciplines in linguistics, the humanities and the social sciences. Corpus linguists construct corpora in order to investigate patterns of language use in a natural, specialised, or group-specific language. They build large corpora that serve as reference corpora of a specific natural language, design technologies for automatically compiling and annotating digital texts with linguistic information, and develop formulae and software for the statistical analysis of distributional patterns in corpora. The steps of analysis usually comprise semi-automated annotation, automated statistical analysis of distributional patterns and further manual analysis of the patterns revealed. Applications include the construction of digital dictionaries and thesauruses, or the study of linguistic varieties, translation and language acquisition (McEnery & Wilson 2001, Tognini-Bonelli 2001). Computational linguists and computer scientists use corpora as resources for Natural Language Processing (NPL), for developing and testing the automated classification and processing of digital data with regard to their linguistic characteristics, including morphological, syntactic, semantic or pragmatic aspects. They develop algorithms that are based on formal models of knowledge of language and recognise
Corpus analysis

character strings or sound bites as graphemes and morphemes; parse the grammatical structure of a sentence, recognise parts of speech or named entities; and infer speech acts and semantic categories from this information. Increasingly, such processing of linguistic data is based on probabilities and machine learning. Algorithms learn from existing annotated corpora how to process and classify large samples of digital texts and employ statistical models to project such information onto un-annotated text samples (Clark, Fox & Lappin 2013; Jurafsky & Martin 2009).

While corpus and computational linguists seek to construct a corpus in a way that shows language use in (a representative segment of) a natural language, discourse linguists focus on (a segment of) the totality of texts that, through semantic, functional, intertextual and pragmatic relations, connect with the issue being investigated (Busse & Teubert 1994, p. 14f). Such issue-specific corpora have been used in two ways in discourse linguistics. On the one hand, they facilitate a description of language use in a specific socio-cultural domain that draws on traditions of semantics, stylistics and pragmatics in corpus linguistics (Baker 2006, Stubbs 1996, Stubbs 2001). On the other hand, issue-specific corpora assist in the analysis of culturally dense language use that borrows from historical semantics and cultural studies (Busse & Teubert 2013; Wengeler 2006) and focuses on the corpus linguistic exploration of categories such as topoi or semantic frames, which are rooted in traditions of philological-hermeneutic text critique (Bubenhöfer 2009; Gür-Şeker 2012; Spitzmüller & Warnke 2011; Ziem, Scholz & Römer 2013).

Issue-specific corpora are also used in Critical Discourse Analysis (CDA), which highlights the discursive constitution of symbolic power and social exclusion, linking categories from socio-, text- or systemic-functional linguistics to social theory and theories of discourse (Fairclough & Wodak 1997; Forchtner & Wodak in this volume). Corpus analysis is here employed as an ancillary tool, to better contextualise small-scale studies of detailed discourse analysis. For instance, a query of words regularly co-occurring with the search word ‘unemployed’ in a newspaper corpus was used to obtain a rough idea of how that specific newspaper lexicalised unemployment – here relating to either structural disadvantage or individuals’ failure (Mautner 2009). The distributional-semantic patterns revealed by corpus linguistic procedures also facilitate the systematic focusing and selection of individual texts for an otherwise unaltered CDA methodology (Baker et al. 2008; Kutter 2013b).

An ancillary approach towards issue-specific corpora is also applied by content analysts in the social sciences. They use computational and corpus linguistic tools to either automatically retrieve lexical-semantic information from large amounts of texts, or automatically assign content-related information to them that matters in relation to a specific social-science research question. The objective is thus to extend the analysis beyond the few texts that can be handled manually by hermeneutic methods. Following the example of The General Inquirer, the first influential content-analytical software, such analysis is usually based on an abductively generated dictionary (or ‘code book’) that defines how single terms, specific word clusters or passages of text shall be read in terms of the social-science categories investigated. The annotations are then queried with regard to salience, co-occurrence, relations to meta-information, and so on. Recent developments in NPL, in particular topic models, word-sense disambiguation, or named entity recognition, promise to replace dictionary-based annotation with machine-learning algorithms (Wiedemann 2013), or inform and prepare smaller-case hermeneutic content analysis (Kutter & Kantner 2012).

Lexicométrie (henceforth lexicometrics) is yet another approach to the analysis of issue-specific corpora. It is rooted in French lexical statistics and has, since the 1960s, been continuously developed as a method for tracing lexical shifts in political discourse over a
longer period of time (Lebart, Salem & Berry 1998). Statistical procedures, implemented by specialised software, are geared towards comparing different time periods, speakers or text genres. The objective is to establish the overall structure and dynamic of an issue- or domain-related discourse in a systematic manner (Scholz & Mattisek 2014). Recently, lexicometrics has gained resonance beyond its pioneer hubs in French political history, in cultural geography and discourse semantics (Glasze 2013, Scholz & Ziem 2013).

**Dilemmas of conducting corpus research**

The surge in digitisation, automated text processing, Big Data, and funding for digital humanities has enhanced the aforementioned uses of corpus analysis. Scholars in the social sciences and discourse studies hoped to boost the basis and plausibility of their research, while computational and corpus linguistics expected their techniques to be straightforwardly applicable in other disciplines. But pioneers soon discovered that the tools provided little more than an overview of the distribution of lexis in texts, while the time needed to develop an interpretive approach towards these revelations had gone on the establishment of data and technical infrastructure, the development of tools and interdisciplinary co-ordination. Conducting corpus analysis in sustainable ways requires taking decisions early on with regard to corpus size, tools, middle-range theories and the philosophy of science. The appropriate size of text collections depends on the research purpose: general statements about a natural language require extra-large reference corpora of several billion words; the study of distinct vocabularies of a specific debate or institution will draw on a smaller issue-specific corpus of a million words or so, and the systematic comparison of historical varieties on a few texts. However, given that corpus analysis relies on statistics, a critical mass of ‘tokens’, that is, of computable (linguistic) items, is necessary to ensure that distributional features can actually be calculated and regular patterns represented. The compilation, cleaning and pre-processing of larger corpora is labour-intensive, though, and sustainable only if corpora are (re-)used in long-term projects.

The choice of tools is a related problem: tools in NPL, corpus linguistics and lexicometrics require some form of prior or dynamic linguistic annotation, time-consuming pre-processing and, occasionally, support from computational linguists (see section on ‘Focus and procedures’). NPL tools are particularly suited for data management and pre-classification. For example, a tailored web crawler can compile a corpus automatically from the web; duplicates and sampling errors can be discovered by text mining, and topic models or supervised vector machines can identify texts containing similar content (Kantner et al. 2011; Wiedemann 2013). Using corpus linguistic and lexicometric tools is advisable if there is interest and skill in linguistic and statistical interpretation. Scholars focusing on the distribution of content or interpretive categories such as ‘frames’ might be well advised to use content-analytical software, such as MAXQDA or Atlas.ti, instead.

What is important to keep in mind is that corpus analysis operates on linguistic categories: it is telling with regard to lexical, semantic, syntactic, grammatical and pragmatic characteristics of texts, insofar as these reveal themselves in the distribution and clustering of words. Taken by themselves, these insights are not necessarily relevant to social and discourse researchers. The further research questions go beyond linguistics proper, the more urgent is, therefore, the adoption of middle-range theories that specify how the issue being investigated is likely to be constructed discursively and how this can be studied in terms of the distributional features of language use of a specific corpus (see section on ‘Analytical strategy…’ on p. 179).
Along with the question of how to approach linguistic information comes the question of how to conduct corpus analysis scientifically. Should we take sides with those discourse linguists who suggest that corpus-driven approaches, because they proceed inductively, are more scientific than corpus-based ‘intuition-led’ approaches (Bubenhofer 2009)? Should we align ourselves with discourse scholars who, mirroring this (mis)representation, reject corpus-driven approaches as a-theoretical (Fairclough 2014) or incompatible with a social-studies understanding of text collections (Angermüller 2014)? Or, should we focus on the operational level of method triangulation, instead (Baker & Levon 2015)? Rather than providing a fair portrayal of alternative routes into corpus analysis, these stances suggest an alignment with specific positions in the ‘boot-camp debate’ in corpus linguistics (Worlock Pope 2010) and the schism between the quantitative (nomothetic-deductive) and qualitative (interpretive) paradigms in the social sciences. A more productive way of resolving research-philosophical issues is to acknowledge that purist corpus-driven approaches, too, start from some conception of language (McEnery & Gabrielatos 2006), that the coupling of data-driven assessment with introspection and of automated-quantitative assessment with manual-interpretive analysis is an established practice in contemporary corpus analysis (Fillmore 1992), and that the critical appropriation of corpus analysis, rather than its rejection, is an appropriate way of dealing with the new dispositive of text processing in qualitative social research.

The crucial decision concerning what philosophy of science should guide one’s corpus-analytical project, then, no longer relates to the choice between corpus-driven and corpus-based approaches. Instead, the question is whether corpus-driven and corpus-based research should be conducted from the perspective of a positivist or post-positivist philosophy of science. Adopting a positivist stance means accepting as scientific only those insights that are generated by applying quasi-experimental methods to data sets, which are representative of the subject studied. Corpus-analytical expertise, then, centres on constructing corpora in line with principles of representativeness and reliability, so that text-statistical analysis will yield patterns that can be regarded as evidence of a ‘general truth’ about language use (Gabrielatos 2007; Gür-Șeker 2014; Kantner, Kutter, Hildebrandt & Püttcher 2011). When adopting a post-positivist stance, one will accept as scientific only those insights that are produced by reflexive methods and facilitate a complex reading of the issue investigated, which extends, instead of confirms, conventional ways of interpretation (Andersen Åkerstrøm 2003). While corpora used for exploration have to be constructed in a way that plausibly supports the research objective, the main emphasis is on revealing the analytical strategy applied (see section on ‘Analytical strategy…’, p. 179).

In short, as with any other method, corpus analysis comes with a range of technical issues and disputed research-philosophical assumptions. They need careful consideration before proceeding to implementation. The following section will provide a basis on which to decide whether or not corpus analysis is the way to go, more particularly, it highlights the specificities and limitations of corpus analysis and descriptive-statistical applications in lexicometrics and corpus linguistics. The examples given below were produced with the help of the corpus linguistic tool Wordsmith and are taken from a German-language corpus of financial commentary that was part of a larger study on crisis discourse (Kutter 2013a).

**Focus and procedures of corpus analysis**

Corpus analysis starts from the *word as the basic unit of analysis*. Words are taken as principal entities embodying content (in content analysis and lexicometrics) or lexical-syntactic and pragmatic meaning (in NPL, corpus and discourse linguistics and corpus-based CDA). Words
function as the primary unit even when analysis extends to word clusters, phrases, texts or hypertexts: ‘the word is the peg that everything else is hung on’ (Mautner 2009, p. 124).

The focus on words and their distributional characteristics within a given corpus corresponds to a certain conception of language and meaning. With the exception of content-analytical applications in the social sciences, where a theory of language is missing, corpus analyses usually draw on linguistic structuralism and its pragmatic turn. They assume the meaning of the analysed words to be given by their relational positioning, which can be characterised in paradigmatic (logical-hierarchical) or syntagmatic (sequential) terms. These relations are seen as being given by the ‘language in use’ manifest in concrete artefacts of written and spoken speech, rather than by a universal linguistic structure. Lexicometrics, corpus linguistics and NPL further stress that these relations can be partially revealed by computing the distribution of words in a text collection. Specific to corpus linguistics is the Firthian assumption that relations recurring in the immediate neighbourhood of a word determine the meaning it tends to have in a specific text collection (McEnery & Gabrielatos 2006; Tognini-Bonelli 2001).

The conceptual focus of corpus analysis is reflected in the view of computer-aided ‘word statistics’. Statistical analysis organises ungrouped numerical data that are characterised by either a quantitative measure, such as score or weight, or some quality, such as a common property. Performing such analysis on texts with the help of computers means transforming texts into numerical data. The basic unit, ‘word’, becomes a set of bytes or a string of characters that is more or less different from other character strings in numerical terms. These ‘tokens’ can be computed with regard to frequency, distribution and co-occurrence and subjected to descriptive or inferential statistics. Their recognition by corpus-analytical software as linguistic units does, however, presuppose pre-processing, in the course of which information modelled on linguistic theory is read from or assigned to character strings. The recognition of bytes as *graphemes* (as units of a writing system of a natural language) depends on standardised character encoding, such as UNICODE or UTF 8, whereas the recognition of various graphemes as forms of a *lexeme* (a lexical-morphological unit of a natural language) or as *parts of speech* presupposes lexical and syntactic parsing. In the course of such parsing, character strings are matched with and classified as matching given information on lemmas and parts of speech of a natural language. In addition, texts may be annotated with (and subsequently queried with regard to) meta-information, such as information on text production or text reception.

What procedures of statistical analysis are then applied, and how the results of computation are represented, largely depends on the chosen school and tool of corpus analysis. Lexicometric and corpus linguistic tools operate on descriptive statistics; they focus on properties of distribution in a given static corpus. Occurrences of a word or word type are counted in terms of absolute or normalised frequencies and these numbers are subsequently subjected to automated statistical procedures. One standard procedure is to calculate the extent to which individual words, pairs of words or patterns of word distribution in one corpus or corpus partition deviate from those in another corpus or corpus partition and from a central tendency (contrasting). Another common procedure is to measure the probability according to which two or more words co-occur in a given corpus (concordancing).

**Contrasting in lexicometrics**

Lexicometric software, such as *Alceste*, *Lexico3* or *Hyperbase*, is particularly strong in corpus-driven contrasting. It focuses on revealing lexical macro structures of a discourse or
topic in a given corpus by comparing partitions of an issue- or actor-specific corpus with each other. These partitions are defined by the researcher beforehand, following criteria that are external to the texts, such as sequences of the period under investigation, authors, speakers or geographical origin. The objective of such contrasting is to reveal, in a systematic manner, whether and how lexical preferences converge or differ between the subdivisions of a corpus both across time (diachronically) and across segments and types of discourse (synchronically). Standard procedures include specificity tests, which establish what words are statistically significantly over-represented in a corpus division when compared to another corpus division, or the overall corpus, and factor analyses that compute latent interdependencies and joint variations in the total of tokens of a corpus before then calculating statistically significant degrees of convergence between these summarising factors. Lexicometric software also provides for a co-occurrence analysis, which establishes what words frequently co-occur in a text passage of specified length, either in terms of a simple co-occurrence (A–B), poly-occurrence (A–B~C~D), or reciprocal co-occurrence (A–B, B–A). Using specificity tests, it is possible, for instance, to calculate and then display in a bar graph how the salience of words such as ‘junior professor’, ‘BA’ or ‘strikes’ changes over time in texts on the Bologna reform of Higher Education in Germany. Employing factor analysis, it is possible to show, in a four-dimensional dispersion plot, which speaker aligns with what group in terms of wording and lexical preferences. Using co-occurrence analysis, it is possible to draw a semantic web or tree graph, from the same text material, which highlight the proximity of expressions such as ‘area of Higher Education’ with ‘European’, ‘unified’ or ‘mobility’ (Scholz & Mattisek 2014).

**Concordancing in corpus linguistics**

Corpus linguistic tools, such as AntConc® or Wordsmith®, are particularly suited to a type of concordancing that focuses on selected search words and their co-text. Co-occurring words are shown in the concordance lines of a keyword-in-context (KWIC) display and listed according to absolute or normalised frequencies, or according to the number of texts in which they co-occur. The fact of co-occurrence can be further assessed by computing different types and measures of probable co-occurrence while considering the different degrees of closeness or distance that the search word and co-occurring words tend to have in a corpus. Depending on the formula chosen, different features of collocation are highlighted. For instance, a $T$ score will foreground those collocates that frequently occur next to the search word, while a Mutual Information (MI) score will rank highest those collocates that may not be particularly frequent but, compared to other collocates, tend to co-occur only in close proximity to the search word, not further away (Hoffmann et al. 2008, pp. 139–158; McEnery & Wilson 2001, p. 86f). These measures may be combined with tests of statistical significance, such as log-likelihood (LL), which establishes the level of confidence that the revealed feature is not due to chance (Gabrielatos & Baker 2008). The Collocates thus calculated may be assessed with regard to the frequency with which they occur in specific positions to the left or right of the search word within a specified span of words, using the Patterns display, or as part of multiword units that are based on n-grams, using the Clusters display of the concordancing tool in Wordsmith.

Table 11.1 shows the co-occurrences of Staat (‘state’) highlighted by $T$ score vs MI score as they occurred in the investigated German-language corpus of financial commentary. The $T$ score foregrounds conjunctions (if, then, so that) and modal verbs (has to, ought, should, may), financial means (money, billions) and quantifiers (more, many, billions), or specific actors
Table 11.1 Collocates of Staat according to T Score and MI Score

<table>
<thead>
<tr>
<th>Word 2</th>
<th>T Score</th>
<th>MI Score</th>
<th>Word 2</th>
<th>Joint</th>
<th>MI</th>
</tr>
</thead>
<tbody>
<tr>
<td>dass (so that, in order to)</td>
<td>167</td>
<td>12,13</td>
<td>schlanker (lean)</td>
<td>5</td>
<td>9,62</td>
</tr>
<tr>
<td>wenn (if)</td>
<td>103</td>
<td>9,63</td>
<td>zurückzieht (withdraws)</td>
<td>3</td>
<td>9,15</td>
</tr>
<tr>
<td>muss (has to)</td>
<td>85</td>
<td>8,99</td>
<td>erpressen (blackmail)</td>
<td>4</td>
<td>8,71</td>
</tr>
<tr>
<td>von (by, from)</td>
<td>74</td>
<td>8,47</td>
<td>Vater (father, nanny)</td>
<td>10</td>
<td>8,50</td>
</tr>
<tr>
<td>kann (can)</td>
<td>73</td>
<td>8,22</td>
<td>mischt (meddles)</td>
<td>3</td>
<td>8,47</td>
</tr>
<tr>
<td>mehr (more)</td>
<td>74</td>
<td>8,06</td>
<td>subventioniert (subsidises)</td>
<td>3</td>
<td>8,30</td>
</tr>
<tr>
<td>Geld (money)</td>
<td>61</td>
<td>7,56</td>
<td>regelt (regulates)</td>
<td>4</td>
<td>8,08</td>
</tr>
<tr>
<td>Banken (banks)</td>
<td>55</td>
<td>7,00</td>
<td>heraushalten (step back)</td>
<td>3</td>
<td>8,01</td>
</tr>
<tr>
<td>soll (ought)</td>
<td>43</td>
<td>6,46</td>
<td>bürgt (guarantees)</td>
<td>3</td>
<td>8,01</td>
</tr>
<tr>
<td>solle (should)</td>
<td>38</td>
<td>5,93</td>
<td>zurückziehen (withdraw)</td>
<td>6</td>
<td>8,01</td>
</tr>
<tr>
<td>Milliarden (billions)</td>
<td>37</td>
<td>5,91</td>
<td>einspringen (jump in)</td>
<td>13</td>
<td>8,00</td>
</tr>
<tr>
<td>viel (much, many)</td>
<td>27</td>
<td>4,84</td>
<td>übernimmt (takes over)</td>
<td>9</td>
<td>7,56</td>
</tr>
<tr>
<td>helfen (help [verb])</td>
<td>23</td>
<td>4,73</td>
<td>ruft (calls [verb])</td>
<td>4</td>
<td>7,30</td>
</tr>
<tr>
<td>dann (then)</td>
<td>27</td>
<td>4,72</td>
<td>Großaktionär (main investor)</td>
<td>3</td>
<td>7,30</td>
</tr>
<tr>
<td>Wirtschaft (economy, business)</td>
<td>25</td>
<td>4,66</td>
<td>starker (strong [nominative])</td>
<td>6</td>
<td>7,22</td>
</tr>
<tr>
<td>Unternehmen (firms)</td>
<td>22</td>
<td>4,37</td>
<td>eingreifen (intervene)</td>
<td>8</td>
<td>7,10</td>
</tr>
<tr>
<td>darf (may)</td>
<td>20</td>
<td>4,37</td>
<td>sichert (secures)</td>
<td>3</td>
<td>6,88</td>
</tr>
<tr>
<td>müssen (have to)</td>
<td>22</td>
<td>4,32</td>
<td>rufen (call [verb])</td>
<td>7</td>
<td>6,86</td>
</tr>
<tr>
<td>starken (strong)</td>
<td>19</td>
<td>4,32</td>
<td>behält (retains)</td>
<td>3</td>
<td>6,82</td>
</tr>
<tr>
<td>Markt (market)</td>
<td>19</td>
<td>4,25</td>
<td>starken (strong [accusative])</td>
<td>19</td>
<td>6,77</td>
</tr>
<tr>
<td>Bürger (citizen/s)</td>
<td>19</td>
<td>4,24</td>
<td>gerettet (rescued)</td>
<td>12</td>
<td>6,73</td>
</tr>
<tr>
<td>damit (so that, in order to)</td>
<td>21</td>
<td>4,23</td>
<td>Notfall (emergency)</td>
<td>4</td>
<td>6,71</td>
</tr>
<tr>
<td>können (can)</td>
<td>21</td>
<td>4,17</td>
<td>Anteile (shares [noun])</td>
<td>3</td>
<td>6,71</td>
</tr>
<tr>
<td>ob (whether)</td>
<td>20</td>
<td>4,17</td>
<td>Ruf (call [noun])</td>
<td>9</td>
<td>6,70</td>
</tr>
</tbody>
</table>

Note: The table shows the frequency with which the investigated pair of words occurred within the entire span of words (Joint) and the density of relationship according to T score or MI score.

(business, firms, citizens) as collocates of Staat. In contrast, the MI score highlights as collocates of Staat words associating with state intervention (subsidies, regulate, intervene, rescue, guarantee, Nanny) and state retreat (withdraw, step back, retreat) and related adjectives (strong vs lean) (see Table 11.1). The two measures, when studied across the different displays of Wordsmith’s concordancing tool, facilitate a syntagmatic analysis of the semantic field of Staat in the corpus investigated. The modal verbs revealed by the T score turn out to be predicates, and financial resources and actors turn out to be objects of clauses, in which Staat appears as the grammatical subject. The state thus seems to be regularly portrayed by financial journalists as an agent that is expected to act upon redistribution and dispose of capacities to do so vis-à-vis specific recipients. The collocates highlighted by the MI score suggest that financial journalists conceive of the state not only primarily as a regulatory-redistributive agent that is expected to act upon economic problems (rather than, for instance, a polity), but also relate that conception to catchwords of habitual debates on regulatory policy and idiomatic phrases related to liberal state critique. Such insights into patterns of word use at the macro level of the entire corpus may inform further explorations in a corpus-based content analysis (Kutter & Kantner 2012) or a corpus-based conceptual history (Kutter 2013b).
Contrasting in corpus linguistics

Along with concordance analyses of word co-occurrences, corpus linguistic software also provides procedures for contrasting corpora that are similar to specificity tests in lexicometrics: *keyword analysis* and *consistency analysis*. Keywords are those words in the corpus being investigated that are considered to be distinct or ‘key’ in that corpus because they are over-represented or under-represented when compared to the distribution of the same words in another corpus (the reference corpus). In order to identify keywords, corpus linguists generate *word lists*, which rank the words in a corpus according to frequency. Keywords are then calculated by cross-tabulating the frequency of a single item and the total of tokens in the wordlist of one corpus and the frequency of a single item and the total of tokens in the wordlist of the other corpus. These numbers are subjected to statistical tests, such as *chi-square* or *log-likelihood*, to estimate what difference in frequency is likely to be stronger than random and, hence, points to words that are ‘key’ (McEnery & Wilson 2001, p. 84). Keyword analyses are often used to establish the distinct lexical features of an issue- or realm-specific corpus in comparison to a general language corpus, such as the BNC. But they can also be applied in comparisons of partitions of a corpus where each comprises texts associated with a specific time span, genre or speaker. Figure 11.1 shows which words were key in German financial commentary during the period in which the financial crisis emerged and escalated (P₂: February 2007 to August 2008) when compared to the preceding period of relative calm (P₁: January 2006 to January 2007, reference corpus). The figure ranks the words according to positive and negative keyness, which is the score that measures the degree to which words are statistically significantly over- or under-represented when compared to words in the reference corpus (Figure 11.1; see also Figures 11.2 and 11.3). The method of consistency analysis compares the frequencies of individual words across text files, or across word lists and the keyword lists of different corpora. The procedure is usually applied in order to determine the lexical, stylistic, and so forth, variance of (historical) versions or types of texts. It may also, however, be used to establish consistent word use across different partitions of a corpus where each comprises texts associated with a specific time span, genre, or speaker (see section on ‘Methods…’, p. 181).

Implications of the corpus-analytical view

The given examples illustrate that corpus analysis yields a very specific view of the investigated texts. The complexity of texts is radically reduced: first, through quantitative abstraction that highlights distributional features of words; and second, by abstracting search words and their immediate co-text from the rest of the texts. This condensed representation of texts does not result from a process of interpretation, but from statistical procedures applied to linguistic units according to algorithms modelled on theories of natural language. It is conceptually selective as it highlights the unit ‘word’ as carrier of meaning, rather than, for instance, structures that ensure the consistency and coherence of texts, and conceives of meaning constitution in syntagmatic terms only insofar as syntagmatic relations can be revealed by statistical methods. Moreover, context is primarily conceived in terms of the statistically defined co-text of a specific word (the words regularly co-occurring in the neighbourhood of a search word). Of course, the larger co-text can be zoomed in on by additionally consulting the embedding text using a KWIC display, or by jumping to the full text. The historically specific context of text production and comprehension may be integrated, too, by relating information on word distribution to additionally collected...
Figure 11.1  Keywords of Period 2 when compared to Period 1

Source: Kutter 2013a (log-likelihood; p-value: 0.000001; stop list applied)
meta-information. It may also be considered by designing an issue-specific corpus in a way that already reflects the specificity of the context of utterance, for example, by selecting a text genre, such as a column on financial pages that enacts the institutionalised practice of financial journalism, and so on. However, the very analytical procedures of corpus analysis are meant to abstract from an individual text and its specific conditions of emergence. Instead, they draw our attention to regular, statistically significant patterns of language use in an entire text collection. In what sense the results of a corpus analysis are meaningful to a particular discourse-analytical project, or question of political studies, depends on the analytical strategy that has been chosen to arrive at an informed interpretation.

**Analytical strategy: the example of crisis discourse**

An analytical strategy constructs the researcher’s perspective on the subject of investigation in an explicit way; it specifies ‘how the [researcher] will construct the observations of others to be the object of his or her own observations’ (Andersen Åkerstrøm 2003, p. xiii). It determines from what research-philosophical viewpoint and macro-theoretical assumptions the study starts, and how these inform the conceptualisation of the research subject. Specific to a discourse-analytical strategy is to additionally point out what *theory of discourse* guides the study, which defines how reality is expected to be (co-)constituted through acts of uttering and signification. Based on these general decisions, it is necessary to adopt *middle-range theories* about the discursive construction of the subject of investigation that are concrete enough to inform *categories and methods of discourse analysis* and abstract enough to translate insights back into the macro-theoretical framework. A discourse study employing corpus analysis will further specify how insights generated by corpus-based or corpus-driven methods will be read in light of general assumptions and middle-range concepts.

The study on crisis discourse in financial commentary illustrates how such an analytical strategy involving corpus analysis can be formed. It was part of a larger project called ‘Cultural political economy of crisis and crisis management’, which aimed to include the perspective of culture (here understood as semiotic practice) into a political-economic examination of the conditions for the emergence and resolution of the North-Atlantic financial crisis (Jessop 2013). A more specific objective of the discourse study was to explore how professional producers of economic discourse, such as financial journalists, represented and explained the crisis, and whether they adjusted conceptions of economic policy and theory. To this end, we employed the discourse-analytical repertoire of CDA, supported by corpus analysis. The corpus comprised commentaries on macroeconomic policy published by reputable financial columnists between 2006 and 2010. They were manually compiled by three people from papers ascribed priming authority in public debates on economic policy, including: *The Financial Times, The Economist* and *The Guardian* in the UK; *The Wall Street Journal, The New York Times* and *The Washington Post* in the US; and *Handelsblatt, Financial Times Deutschland* and *Süddeutsche Zeitung* in Germany. The German sample used for illustration here contained 2,025 commentaries and amounted to 1,172,111 tokens (Kutter 2013a).

The overall project was motivated by a critical-realist philosophy of knowledge and science, which posits that objective mechanisms exist outside subjective perceptions and can partially be disclosed by humans provided they adopt scientific methods that account for the mutual constitution of human agency and social structure (philosophy of knowledge and science). This research-philosophical stance suggested using reflexive methods and
open-ended empirical-theoretical exploration. It also implied a moderate constructionist ontology, presuming that economies are constituted through both collective imagination and structuration (ontology). Using theories from heterodox political economy and interpretive strands therein, this ontological perspective was substantiated in conceptions of economy and crisis (macro theory). Economies were seen to be historically specific constellations of economic structures, institutions and technologies of economic governance, actors’ strategic coalitions and prevalent discourses that, together, temporarily contain inherent contradictions of capitalist accumulation (Jessop 2013). A crisis in such constellations, when mobilised as a decisive moment of political intervention in public debate, may challenge fortifying discourses. New conceptions of economy may emerge, depending on whether collective actors perceive a specific conjuncture of crisis as opening up horizons for action and whether structural-economic conditions lend plausibility to new complexity reduction (Kutter & Jessop 2014).

The framework implied focusing the discourse study on complexity reduction in financial commentary, while considering the conjunctural conditioning of this selective process. The theory of discourse developed in CDA, which posits that intersubjective meaning emerges from correspondence between linguistic interaction and the context of expression, helped to align these assumptions with a theoretically grounded discourse analysis (theory of discourse) (Kutter & Jessop 2014). The challenge consisted, however, of developing concepts that would grasp shifts in the discursive construction of crisis and inform the choice of categories and methods (middle-range theories). We drew inspiration from the concepts of ‘periodisation’ and ‘crisis narratives’. Periodisation conceptualises social change in line with conjunctural analysis. In contrast to a chronology, which sorts events by putting them on a linear time axis in a simple narrative of succession and simultaneity, periodisation constructs a sequence of periods according to specific criteria, which are relevant to the practical and theoretical topic in question and account for the specific temporality of the phenomenon under investigation, such as economic cycles, discourse events, or longue durée (Jessop 2002). In line with this assumption and drawing on secondary analyses and economic data, we divided the period of investigation (2006–2010) into six phases, with each being characterised by new crisis developments and new sets of priorities in crisis management and, therefore, likely to mark crucial temporal horizons in financial commentary. This included: a period of relative calm and moderate growth or stagnation (P1; January 2006 to February 2007); a period of emerging crisis (P2; March 2007 to August 2008); a period of panic (P3; September to November 2008); a period of normalisation in finances (P4; December 2008 to March 2009); recession and crisis in real economy (P5; April 2009 to February 2010); and the emerging Eurozone crisis (P6; March to December 2010).

The concept of crisis narratives is based on the assumption, developed in narrative policy analysis, that humans structure experiences and define problems through narration. The construction of a crisis as a moment of specific decisive political intervention is seen to rely upon narration and its mobilisation in mediatised public debate. In the course of this debate, individual stories about the emergence and specific quality of the crisis are mapped into a more general template of interpretation (a meta-narrative) that recruits only specific events and only specific attributions of causation and responsibility to the public agenda (Hay 1999). The construction of crisis could, consequently, be text-analysed by establishing what events and topics were selected as newsworthy in financial commentary, what stories about causation and responsibility were developed, and which of them became most persuasive through moves of generalising abstraction (Kutter 2014).
Methods: the example of intra-corpus comparison

Corpus analysis, supported by the corpus linguistic tool *Wordsmith*, promised to capture efficiently the lexical-semantic dimensions of the categories that we had established with the help of middle-range theories. We expected keyword analysis to reveal the salient lexis of each crisis period, hinting at events and topics to us that were virulent during these periods. Via concordance analysis, we hoped to gain an overview of the connotations of fundamental concepts of macroeconomic policy, such as ‘state’. As a result of corpus exploration, three methods could be consolidated, and these suggest themselves for use beyond the study of financial commentary: corpus-based conceptual history, corpus-based content analysis and intra-corpus comparison (Kutter 2013b; Kutter & Kantner 2012). Intra-corpus comparison draws on earlier work (Koller & Farrelly 2010) and will be presented here as an example of corpus-driven methods.

**Periodised intra-corpus comparison**

Intra-corpus comparison, when applied to the category of time, helps to establish how salient lexis changes over the period of investigation and how the use of words broadens or narrows down. It is corpus-driven because it trusts corpus-analytical software to produce relevant insights. Basic corpus linguistic procedures used in this method are keyword analysis and consistency analysis (see section on ‘Focus and procedures…’). However, instead of comparing the language use in the corpus being investigated with the language use in a general corpus of a natural language, following the procedure that corpus linguists usually apply, one compares the periodized partitions of the corpus being investigated with each other. The partitions cover periods of time during which changes in discursive practice, which are relevant from the perspective of the research question, are likely to have occurred (periodisation, see the previous section). For the study on crisis discourse, we divided the period of investigation (2006–2010) into six phases, each being marked by different crisis developments, and divided the corpus in corresponding partitions (see the previous section). The steps of exploration included:

- a comparison of individual period-specific corpus partitions to the overall corpus, which reveals word use that distinguishes the respective crisis period from the rest of the period of investigation;
- a comparison of succeeding periods to preceding periods (P₂ to P₁; P₃ to P₂ and so on), which reveals in what way word use in the succeeding periods differs from that of respective previous periods;
- a consistency analysis that establishes what words are more or less consistently used over the period of investigation.

Applied to the periodised corpus of financial commentary, the first step of exploration suggested that journalists did indeed shift their attention corresponding to those events and issues that we had considered relevant during periodisation. For instance, while using words in P₁ that related to adjustments to German fiscal policy and health-system reform agreed upon by the new federal coalition government, journalists employed words in P₂, which are associated with the US housing and subprime crisis, the successive defaults of US and German financial institutions, and rescue actions (see Figure 11.1; section on ‘Focus and procedures...’). The second step revealed, however, that despite these distinct features,
overall word use differed considerably only between \( P_1 \) and \( P_2 \): here, both the number of keywords and their keyness proved very high. In contrast, word use during subsequent periods (\( P_3 \) to \( P_5 \)) seemed to remain largely the same: only some words of little keyness were added and a few dropped out again, with the emerging Eurozone crisis (\( P_6 \)) deviating again from that pattern (see Figures 11.2 and 11.3). Scrutiny of actual keywords revealed that the lexis, which was salient during \( P_1 \) and mainly related to welfare-state adjustment and fiscal reform, was largely replaced by a rich variety of financial-market lexis in \( P_2 \) (see Figure 11.1; section on ‘Focus and procedures’). During subsequent periods, words entered and dropped out again, which pointed to new sites of trouble and new crisis managers. The financial-market lexis, however, seemed to remain; it did not surface according to any threshold of significant deviation. This also held for the period of the emerging Eurozone crisis \( P_6 \), which had seemed different when looking at the numbers.

The exploration of crisis labels (compound nouns built from the German word \textit{Krise}) in the third step further indicated that the consolidation of financial-market lexis corresponded to a narrowing of conceptions of crisis. During the period of emerging financial crisis, \( P_2 \), a variety of crisis labels was used, including housing crisis, mortgage crisis, subprime crisis, banking crisis, crisis of confidence, and so on. In the panic period, \( P_3 \), however, the term \textit{Finanzkrise} (financial crisis) became the default label. It was hitherto consistently used, complemented only by \textit{Wirtschafts- und Finanzkrise} (economic and financial crisis) during \( P_4 \) and \( P_5 \), and \textit{Griechenlandkrise} (Greek crisis) and \textit{Eurokrise} (Eurozone crisis) during \( P_6 \).

The revelations regarding the diachronic features of word distribution had great heuristic value for the study. Read in light of the interpretive scheme of crisis narratives, these revelations suggest that financial journalists developed a crisis narrative as early as 2007, which hitherto remained largely the same and which emphasised, above all, financial-sector issues. Commentators seemed to have consolidated their conception of the crisis as being a

![Figure 11.2](https://example.com/figure112.png)

\textit{Figure 11.2} Number of keywords per period when compared to the previous period

Source: Kutter 2013a (log-likelihood; \( p \)-value: 0.000001; stop list applied)

182
financial-system crisis no later than 2008. They applied this reading to subsequent crisis displacements, too, regardless of whether these related to fiscal squeeze, social-humanitarian calamities and political turmoil. The insights obtained by intra-corpus comparison were also valuable because they induced reflexivity. They challenged conclusions drawn from the content and discourse analysis of causal stories, which had suggested that crisis conceptions and lessons drawn changed during the period of investigation (Kutter 2014). More generally, the insights obtained by this method suggested that, when studying crisis-induced shifts in debates on economic policy, one has to take prevalent conceptions of crisis into account to gain a fair understanding of perpetuation or change.

**Conclusions: using corpus analysis in political discourse studies**

In this chapter, discourse researchers, who are considering using computer-aided textual analysis, can familiarise themselves with the field and method of corpus analysis, more precisely: the semi-automated statistical analysis of large collections of digital texts called ‘corpora’. The chapter has promoted the argument that the distinct view of corpus analysis, the sort of abstraction that it generates from large amounts of texts, is the particular value that corpus analysis adds to the discourse study of social and political issues. Corpus analysis does not present texts according to their linear or intertextual composition. Instead, it highlights patterns of regular use of single words or word clusters as they appear in a large collection of texts. Whether showing these words in situ, or in quantified abstraction, this reading provides us with a bird’s-eye view of the use of specific words across all the texts contained in a corpus. In addition, the display of word clusters reveals patterned semantic, syntactic, and so on, relationships between words. This provides us with an additional layer of interpretation, which cannot be gained otherwise. Corpus analysis may challenge not only our habitual way of linear reading, but also our preconceived views about the subject studied.
However, the results of such corpus exploration are often not meaningful in themselves. In order to relate to discourse studies of social and political issues, methods of corpus exploration have to be made part of an analytical strategy. In other words: big data and smart tools are nothing without theory. Neither should we expect procedures of corpus analysis to replace human-manual efforts to disentangle discourse (as many advocates of corpus-driven and unsupervised approaches do). Nor should we nourish the hope that corpus analysis will free us from the bias of small-scale studies and decisively help us in reconstructing the larger context of utterance (as corpus-based and supervised studies often do). Precisely because of its selective focus on the distributional properties of words, corpus analysis is not suited to reproduce the deep comprehension that hermeneutic-interpretive or detailed linguistic discourse analysis generates from texts and their contextual and interdiscursive conditioning. If discourse researchers take on the huge task of compiling, preparing and analysing corpora, they should rather use corpus analysis as an explorative technique. As such, it will yield distinct insights into texts and facilitate generating hypotheses and research questions for further investigation by other methods of textual analysis.

Notes
1 I wish to thank the two editors and Costas Gabrielatos for their helpful comments on an earlier version.
2 An exhaustive account of the growing field of corpus analysis is beyond the scope of this chapter; the review is meant to illustrate the plurality of the field, instead.
3 The following will elaborate on word-based languages only. For an equivalent approach to symbol-based languages, see Jurafsky and Martin (2009).
4 When using corpus linguistic or lexicometric software, this information has to be physically inscribed into texts using algorithmic processing tools, such as tree tagger or the tagging functionalities of analytical software. Text-mining tools, on the contrary, process texts dynamically. Note that standard character encoding, provided by standard corpus-analytical software, is usually enough for a rough analysis of lexis.
5 http://www.image-zafar.com/en/alceste-software
6 http://www.tal.univ-paris3.fr/lexico/
7 www.antlab.sci.waseda.ac.jp/software.html
8 www.lexically.net/wordsmith
10 This is the author’s ex-post reading of the project in light of an analytical strategy. It is based on a series of discussions with the principal investigator, Bob Jessop, but might not be identical with his views.
11 In lexicometrics, the procedure of specificity tests and factor analyses may yield similar insights.

References


