# Taming Textual Data: The Contribution of Corpus Linguistics to Computer-Aided Text Analysis

**Pre-print version of:**
Pollach, I. (2012). Taming Textual Data: The contribution of Corpus Linguistics to Computer-aided Text Analysis, *Organizational Research Methods*, 15(2), 263-287.

## INTRODUCTION

A lot of research in management and organization studies is conducted with textual data. Such data can originate from organizations in the form of annual reports, letters to shareholders, proxy statements, press releases, mission statements, values statements, or in-house magazines. In addition, textual data can be collected outside the organizational domain in the form of news reporting, or can be solicited in the form of interviews. Organizational documents as naturally occurring materials are particularly rich and valuable data. They can provide insights into managerial cognitions, organizational values, culture, or identity which surveys or interviews cannot provide in the same manner. Because of the recurring nature of some of these documents, they are particularly suited for longitudinal studies of events, changes or developments in organizations (Caska et al., 1992; Jablin & Putnam, 2001; Stanton & Rogelberg, 2002). Contingent on one's epistemological positioning, these documents are either accounts of what an organization does or cultural artifacts constitutive of an organization (Taylor & Van Every, 2010, p. 92).

For both positions, the advantages of using software for the analysis of textual data are obvious and given the ample availability of textual data in digital format, the question is no longer whether or not to use computer-aided text analysis, as it was in its beginning (Wolfe et al., 1993), but which approach is the most insightful for a given dataset. Although text analysis is an established method in the social sciences (e.g. Bailey, 1994; Roberts, 1997; Bernard & Ryan, 1998), management researchers analyzing textual data qualitatively or quantitatively are essentially entering linguistic terrain. The cooperation between linguistics and the social sciences with regard to text analysis has always been meager (Markoff et al., 1974; Roberts, 1989; Bernard & Ryan, 1998; Popping, 2000), which hinders the advancement of text analysis in the social sciences, including computer-aided approaches. When examining the reference lists of seminal work in computer-aided texts analysis (Gephart & Wolfe, 1989; Roberts, 1989; Gephart, 1993; Wolfe et al., 1993; Kabanoff & Holt, 1996; Kabanoff, 1997; Kabanoff & Brown, 2008; Janasik et al., 2009), one cannot fail to notice that they quote primarily the content analysis literature and literature on qualitative methods in the social sciences, but largely ignore literature from the field of linguistics. Corpus linguistics, which is a branch of linguistics that conducts computer-aided analyses of language, can contribute insights into the analysis of textual data in the social sciences. Corpus linguistics studies real-life language use on the basis of a text corpus. A corpus is defined as "a body of text which is carefully sampled to be maximally representative of a

language or language variety" (McEnery & Wilson, 2001, p. 2). Corpus linguistics encompasses a number of analysis techniques that can be applied as needed rather than according to a particular protocol. Although there is a strong focus on quantitative language analysis techniques in corpus linguistics, an essential part of any corpus-linguistic study is always the qualitative examination and interpretation of quantitative results (Biber et al., 1998).

Corpus linguistics can be both a meaningful addition and an alternative to existing computer-aided text analysis methods, including both quantitative, positivist analyses of content as well as qualitative, interpretive analyses of discourse. This paper seeks to advance organizational research methods and in particular computer-aided text analysis by introducing developments from the field of corpus linguistics to the field of computer-aided text analysis. More specifically, the contribution of this paper is threefold: First, two different approaches within computer-aided text analysis are discussed and compared to corpus linguistics in order to highlight where corpus linguistics can provide new insights. Second, the paper introduces resources and analysis techniques from corpus linguistics that are currently not used or not fully exploited in computer-aided text analysis. Third, the paper presents an exemplary analysis of letters to shareholders based on techniques from corpus linguistics in order to demonstrate in a hands-on manner how corpus linguistics can be used by non-linguists. The paper concludes with a discussion of the value and limitations of corpus linguistics for management and organization studies.

## COMPUTER-AIDED APPROACHES FOR THE ANALYSIS OF TEXTUAL DATA

The analysis of textual data can focus on the manifest content of texts, which is transmitted through explicit vocabularies, or latent content, which denotes the implicit meaning in text (Merton, 1957; Phillips et al., 2008). While quantitative-oriented forms of text analysis, e.g. classic content analysis (Krippendorff, 1980), can produce indices of manifest content, text analyses following the interpretive tradition study both manifest and latent content, as the socio-cultural framework in which the text has been produced is an integral part of the analysis, e.g. as in grounded theory or discourse analysis. Both forms of text analysis can be supported by computer software. The following sections explain and compare two different computer-aided approaches: Corpus linguistics and computer-aided text analysis, the latter of which is divided into computer-aided content analysis and computer-aided interpretive textual analysis. These three approaches are summarized in Table 1, which provides an overview of the main characteristics of each approach.

---

Table 1

---

### Corpus Linguistics

In parallel to the computer-aided text analysis in the social sciences, a computation-oriented approach to text analysis has evolved in linguistics, which is called corpus linguistics. Seminal work in corpus linguistics goes back to the 1960s (e.g. Kucera & Francis, 1967), while modern corpus linguistics using large computer-based corpora emerged only in the 1990s (e.g. Sinclair, 1991; Leech, 1992; Svartvik, 1992; Biber, 1996; McEnery & Wilson, 1996; Stubbs, 1996; Biber et al., 1998), stimulated by the rise in available computing power. Since the 1990s, corpus linguistics has become a

dynamic branch of linguistics, out of which research groups, associations, conferences, journals, monographs, and book series have emerged (Anderson, 2008). Definitions of corpus linguistics are many and varied. While it is seen as a discipline by some (e.g. Tognini-Bonelli, 2001; Teubert, 2005) and as a methodology by others (e.g. McEnery & Wilson, 2001, p. 2), yet others see it as just "an approach" (Mahlberg, 2005), "a methodological basis" (Leech, 1992, p. 105), "a bundle of methods, procedures and resources" (Lüdeling & Kytö, 2008) or a "toolbox of techniques" (Lee, 2008, p. 87). As a discipline, corpus linguistics provides methodological innovations for computer-based analyses of texts. As a methodology, corpus linguistics is used by researchers from other fields who conduct computer-based language analyses (Mukherjee, 2010).

There is neither a "unifying theory binding together corpus linguistics analyses" (Barlow, 2011) nor are there uniform practices or techniques about how one 'does' corpus linguistics. Corpus linguistics can be used in a primarily qualitative design to generate theory inductively or in a primarily quantitative, positivist design to test theory on language (McEnery & Wilson, 2001, p. 110). McEnery and Gabrielatos (2006) argue that the role of theory in corpus linguistics should be viewed as a continuum between testing theory as one end point and developing theory inductively as the other. The lack of a stringent methodology and the flexibility this entails mean that researchers employ an eclectic mix of techniques that are combined as needed for a particular inquiry, guided by their research questions (McEnery & Gabrielatos, 2006). This flexibility allows researchers to use corpus linguistics for a wide variety of inquiries. However, this flexibility has also given rise to criticism against corpus linguistics, because one cannot rule out that an analysis is driven by mere intuition or the capabilities of the software rather than by a research question (Tognini-Bonelli, 2001; McEnery et al., 2006).

Corpus linguistics always analyzes corpus data both quantitatively and qualitatively in order to explain and interpret patterns rather than just count them (Biber et al., 1998). The techniques within corpus linguistics range from tools for the identification of meanings to descriptive, quantitative indices of textual data (Baker et al., 2008). The most central analysis techniques within corpus linguistics include (1) word frequencies and keyword-in-context (KWIC) searches, which display all instances of a given word in its immediate textual surroundings and help the researcher to connect words of potential interest to their context (Wood, 1974); (2) the comparison of corpora (McEnery & Wilson, 1996); (3) collocations, which denote the co-occurrence of two or more words (Sinclair, 1991); and (4) statistical procedures for the assessment of word frequencies (e.g. Lebart et al., 1998; Oakes, 1998; Manning & Schütze, 1999).

The research questions corpus linguistics can answer address the association of textual patterns with either other textual patterns or with contextual patterns (Biber et al., 1998). For example, it can study language use in a particular text genre, in a particular kind of discourse, or in a particular social or communicative context (Barlow, 2011). Linguists of all persuasions have used these techniques in qualitative or quantitative designs, ranging from semantic studies of near-synonyms (Liu, 2010) to sociolinguistic studies of language variation (Kachru, 2008) and language change (Baker, 2009). Linguists have also paid attention to the integration of corpus linguistics into discourse analysis with a view to reducing the subjectivity inherent in discourse analysis (e.g. Baker, 2006). This integration can take the form of a corpus-informed discourse analysis, which is purely qualitative in nature. Alternatively, a corpus-supported discourse analysis or a corpus-driven discourse analysis can be conducted, both of which are qualitative and quantitative in nature. The difference between the latter two is that a corpus-supported analysis starts out from a theoretical framework, whereas a researcher conducting a corpus-driven analysis approaches the data with only very few preconceptions (Lee, 2008).

A full-text search for "corpus linguistics" in management journals in the EBSCO database yields only one article that conducts a corpus-based linguistic analysis. This article by Cornelissen (2008) contributes to the literature on the discursive construction of organizations by studying how people use metonymies (part-whole or whole-part relations between words) when they talk about organizations. It draws on the British National Corpus, a large, publicly available corpus of naturally occurring texts, such as books and newspaper articles, but does not report the use of any specific corpus-linguistic technique. The study concludes that people use conventionalized metonymic patterns when they talk about organizations.

## Computer-Aided Text Analysis

Computer-aided approaches to text analysis in the social sciences are generally referred to as computer-aided text analysis (e.g. Gephart & Wolfe, 1989; Wolfe, 1990; Dowling & Kabanoff, 1996; Kabanoff, 1997; Duriau et al., 2007; Short et al., 2010) or computer-aided content analysis (e.g. Dowling & Kabanoff, 1996; Duriau et al., 2007; Short & Palmer, 2008). Several scholars have argued that the two terms can be used interchangeably (Dowling & Kabanoff, 1996; Popping, 2000; Duriau et al., 2007). Krippendorff (2004, p. 261) and Short et al. (2010) see computer-aided text analysis as an approach to content analysis, while Kabanoff (1997) argues that computer-aided text analysis should be the broader term, because it goes beyond the goals and possibilities of content analysis. This nomenclatural confusion may have its origin in the terminological inconsistency present at the level of traditional text and content analysis. Neuendorf (2002), for example, in her book on content analysis, considers text analysis to be "a part of content analysis research" (p. 25). Meanwhile, Titscher et al. (2000, p. 55), Bauer (2000, p. 132), and Bernard (1998, p. 437) see classic content analysis as a form of text analysis, along with discourse analysis or narrative analysis. In this article, the latter position is maintained, arguing that content analysis is one approach to text analysis, which implies that computer-aided content analysis is an approach within computer-aided text analysis. Parallel to computer-aided content analysis, computer-aided interpretive textual analysis has evolved out of the qualitative, interpretive research tradition. Computer-aided text analysis (CATA) thus represents the encompassing term under which both computer-aided content analysis and computer-aided interpretive textual analysis fall.

### *Classic and Computer-Aided Content Analysis*

Content analysis has its origin in communication studies, where it was first used in the beginning of the 20th century for comparative analyses of newspaper content (Krippendorff, 2004). The goal of content analysis is to make inferences from texts to context in an objective and systematic manner (e.g. Holsti, 1968; Krippendorff, 1980; Neuendorf, 2002). It follows a quantitative, positivist research approach that tries to produce thematic or semantic indices of observable and countable features of text on the basis of pre-defined categories. Research questions answerable with content analysis focus on the presence of concepts in texts, positive and negative sentiments in texts, or the co-occurrence of two concepts in texts (Krippendorff, 2004).

Content analysis began to use computer assistance as early as in the 1960s, where attempts were made to analyze the content of texts with word-count text analysis methods (Popping, 2000), based on the premise that the frequency of particular words and concepts in a text is a measure of importance, attention or emphasis (Krippendorff, 1980). This has led to the sub-field of computer-aided content analysis, which uses content dictionaries to study the frequency and prominence of particular concepts. Two approaches to computer-aided content analysis can be distinguished, depending on the

source of these dictionaries: Researchers can either make use of existing dictionaries or compile a dictionary specifically for a particular study. One such existing dictionary is the General Inquirer by Stone et al. (1966, 2000), which was at the time of its inception the first major attempt to create universal content dictionaries and apply them to texts with the help of computers. Other sets of content dictionaries and software tools include the Regressive Imagery Dictionary by Martindale (1975, 1990), the LIWC by Pennebaker et al. (2003), and DICTION (cf. Short & Palmer, 2008). Those studies for which researchers construct their own dictionaries resemble classic content analysis more closely, because the researchers essentially compile a coding frame based on the analytical constructs emerging from the research question and apply the coding frame to the texts. Self-constructed dictionaries can be derived deductively from theory or inductively from the corpus or from both in a combined approach (Short et al., 2010).

Researchers in management and organization studies have used both self-constructed and existing dictionaries. For example, the former have been used to study the extent of downsizing content in annual reports (Palmer et al., 1997), to investigate justifications for CEO compensation in proxy statements (Wade et al., 1997; Porac et al., 1999), to explore metaphors for teamwork used in interviews on teamwork (Gibson & Zellmer-Bruhn, 2001), and to identify the distinctive lexicon of strategic management from abstracts of articles published in major journals (Nag et al., 2007). Other scholars in management and organization studies have applied existing dictionaries to company documents, for example to study categories of organizational values (Kabanoff et al., 1995), changes in organizational values (Kabanoff & Holt, 1996), or word choice in dispute resolutions in a study on online dispute settlements (Brett et al., 2007). Further, researchers have drawn on existing dictionaries to explore news coverage, for example the coverage of quality circles in a study on the lifecycle of management fads (Abrahamson & Fairchild, 1999), the coverage of substance abuse as one variable in a study on workplace substance abuse (Spell & Blum, 2005), or positive coverage of companies in a study on firm reputation and celebrity (Pfarrer et al., 2010).

*Computer-Aided Interpretive Textual Analysis*

When studying textual data, management scholars are not necessarily interested in measuring concept frequencies, but may also be interested in understanding meanings and interpretations, following the qualitative, interpretive research tradition. The research interests of organizational researchers conducting such studies may, for example, lie in sensemaking (e.g. Gephart, 1993; Gephart, 1997), impression management (e.g. Snell & Wong, 2007), organizational conflicts (e.g. Doucet & Jehn, 1997), legitimation (e.g. Vaara & Tienari, 2008), or attributions of organizational outcomes (e.g. Tsang, 2002). Since interpretative analyses are always inductive and iterative in nature, computer assistance for this kind of analysis is first and foremost valuable for mechanical tasks associated with handling data, locating themes, organizing them and linking them (Kelle, 1995). The software used for such analyses therefore needs to be capable of organizing data, coding text segments, and examining words in their context, rather than counting frequencies and reporting statistics. Examples of such software tools include NVivo, QDA Miner, and ATLAS.ti.

Advances in text retrieval possibilities have led to the emergence of computer-aided interpretive textual analysis, which relies on computer assistance to uncover themes, meanings and interpretations of events in the hermeneutic tradition. The research questions for which this approach is most suited for focus on the role of language in constructing reality. Thus, it focuses on the use of particular concepts, on the way meaning is created in texts, and the way meaning is shaped by relations between texts (Gephart, 1993). Analysis techniques include mainly word frequency lists and KWIC searches. Further, collocations can be explored to uncover meanings constructed by a particular word combination. Computer-aided interpretive textual analysis can involve textual statistics, but only to

identify important words and search for meanings rather than to produce quantitative results (Gephart, 1993, 1997, 2003). Examples of such studies include Gephart's (1993, 1997) work on organizational sensemaking about hazards, which is based on an iterative process of KWIC searches for potentially interesting words and collocations, supplemented with expansion analysis.

## THE TECHNIQUES OF CORPUS LINGUISTICS

Corpus linguistics is similar to CATA in that it can be used for both quantitative and qualitative inquiries. The main differences between corpus linguistics and CATA are that corpus linguistics (1) focuses on lexical patterns rather than categories or meanings, (2) consists of a set of techniques without a methodological protocol, and (3) always involves a combination of quantitative and qualitative analysis. This section will explain those analysis techniques and resources from corpus linguistics that can be relevant for a study in the social sciences. These pertain to word dispersion measures, corpus comparison and keywords, collocations, and the construction of dictionaries. All analysis techniques outlined in this section are performed or supported by commercially available corpus linguistics software or spreadsheet software. The appendix to this paper contains an overview of various software tools, indicating which tool is capable of what.

### Word Dispersion

Textual data used in management and organization studies typically consist of a collection of texts rather than just one text, for example a sample of annual reports, mission statements, corporate self-descriptions, or interview transcripts. Apart from looking at total word frequencies, a researcher might want to know whether words of interest cluster in one of the texts or are evenly distributed among the texts in the corpus. To assess the dispersion of a word, three analytical possibilities exist. First, the range (Leech et al., 2001) or spread (Gabrielatos et al., 2010) of a word can be determined, which is the percentage of texts that contain a particular word, irrespective of how frequently it is used in these texts. A more precise measure is Juilland's D1 (Oakes, 1998; Leech et al., 2001, p. 18), which is a ratio between 0 and 1, with 1 indicating the perfectly equal dispersion of a word among the various texts studied and 0 indicating a highly unequal dispersion. Juilland's D, therefore, does not indicate where a word clusters but only to what extent it clusters. Ultimately, word dispersion can be studied in distribution plots, which visualize where in a set of documents a word clusters (Scott, 1999).

### Corpus Comparison and Keywords

Frequency information is always most informative when corpora from different sources or different times are compared (Hunston, 2006). When comparing two corpora, words that occur significantly more frequently (positive keywords2) or infrequently (negative keywords) in one corpus compared to the other one can be identified (Scott, 1999; Xiao & McEnery, 2005; Baker et al., 2008). This comparison should be made on the basis of a log-likelihood test1, since word frequencies in a text are not normally distributed (Dunning, 1993). The keywords identified can indicate the saliency of certain text features, such as the 'aboutness' of a text, stylistic characteristics or descriptors of text genres. While these keywords provide quantitative evidence of observations and therefore reduce

[1]See Table 3 for details

[2] The term 'keyword' is not to be confused with the term 'keyword' in the context of keyword-in-context (KWIC) searches. The keywords in KWIC searches are not the outcome of a statistical test of significant differences in word frequencies between two corpora. Rather, these keywords are selected by the researcher as potentially interesting words.

researcher bias, these observations only provide interdictors of patterns, which must be interpreted by the researcher with the help of keyword-in-context (KWIC) searches (Baker, 2004).

One can compare one's own corpora with each other or with a very large, publicly available corpus as a reference corpus to establish what is "normal" and what is not. Such public corpora include, for example, the 100m-word British National Corpus3, the American National Corpus4, Collins Wordbanks Online English5, Cambridge International Corpus6, or more specialized corpora such as the Corpus of Professional English7 and the Corpus of Professional Spoken American English8. Xiao and McEnery (2005), for example, have compared the latter corpus with the British National Corpus to compare the genres of conversation, speech, and academic prose in American English on the basis of positive and negative keywords. Similarly, Johnson et al. (2003) have used keywords to compare three newspapers over a 5-year period against the British National Corpus in a discourse analysis of political correctness in British newspapers. Both studies have used keywords to identify words of potential interest, which were then examined qualitatively with KWIC searches. External corpora can thus be of relevance in a business and management context, when a researcher wants to study either a particular genre of organizational texts or a very specific concept.

## Collocations

Collocations denote "the above chance co-occurrence of two word forms" (Sinclair, 1991). The analysis of collocations is considered to be "a natural extension of frequency lists" (Gries, 2009, p. 14), in that collocations capture multi-word expressions rather than individual words only. Collocations can indicate a semantic preference for certain constructions or can uncover meaning imbued in words by those words they collocate with (Stubbs, 2001) and thus give insights into the mental lexicon of the text producer (Mollin, 2009). The frequency of certain collocations in language often leads to established, conventionalized expressions which language users choose instead of creating their own combinations of words (Sinclair, 1991, p. 170). Because of this repeated use, collocations can become carriers of cultural meanings or domain-specific meanings (Bartsch, 2004, p. 12). A collocation analysis therefore reveals discourse patterns and meanings that are neither evident form frequency lists of individual words nor from the readings of larger volumes of text in a manual analysis (Baker et al., 2008).

Corpus software can identify collocations in two different ways: First, the search for collocations can be open such that the software returns the most frequent word combinations within a pre-determined word span. Second, when one word in particular is examined for other words it co-occurs with, the former is referred to as node and the latter as collocate (Sinclair et al., 2004). The search for collocations then begins with the specification of a node and the corpus software finds all collocates within a predetermined word span, usually 3 to 5 words on each side of the node word (Bartsch, 2004). Collocations as such are not new to computer-aided text analysis, as they have also been used in computer-aided interpretive textual analysis (Gephart, 1993, 1997), although software at that time was not advanced enough to return frequent collocations without researcher input, but required the researcher to define node words. In order to judge how noteworthy a collocation is, there are multiple ways of determining their strength. This is necessary, as the mere frequency of a collocation is biased by the frequency of the two words making up the collocation, i.e. more frequent words are more likely

[3] http://www.natcorp.ox.ac.uk/
[4] http://www.americannationalcorpus.org/
[5] http://www.collins.co.uk/Corpus/Corpussearch.aspx/
[6] http://www.cup.cam.ac.uk/elt/corpus/cancode.htm
[7] http://www.perc21.org/
[8] http://www.athel.com/cspa.html

to appear as part of a collocation than less frequent words. Therefore, statistical measures of collocation strength are needed in order to account for the likelihood of two words to co-occur (Biber et al., 1998).

The strength of a collocation between two words can be measured by the mutual information (MI) score of these two words[9] (Church et al., 1994), with a score higher than 3 being considered a strong collocation (Baker, 2006, p. 120). However, the MI score assigns higher scores to rare words that produce unique collocations than to collocations containing frequent words. An alternative is z-scores[9], which favor high-frequency words (Lindquist, 2009) and assume normally distributed data, which is, however, not the case (Dunning, 1993). They have also been used in computer-aided interpretive textual analysis (Gephart, 1993). To address the trade-off between saliency and frequency, a log-likelihood test[9] can be used as a compromise between MI scores and z-scores. Corpus linguists argue that the best approach to determining collocation strength is to calculate the results with all three algorithms for several collocations, rank the collocations according to the scores obtained for each algorithm, and then draw conclusions based on a comparison of these rankings (Baker, 2006; Lindquist, 2009). Despite these measurement problems, collocation strength is still an important element of a collocation analysis, as a purely manual analysis may miss strong collocations or include weak ones (Baker et al., 2008). Examples of studies drawing on collocations include Koteyko's (2010) qualitative study of the discourse of climate change based on institutionalized compound words as well as new coinages. Hamilton et al. (2007) examined the meanings of 'risk' in Collins Wordbanks Online English and the Cambridge International Corpus, using both KWIC searches and strength measures. Caldas-Coulthard and Moon (2010) conducted a critical discourse analysis of the construction of gender in tabloid and broadsheet newspapers. They compared the use of collocations containing selected adjectives in these two corpora on the basis of KWIC searches and quantitative measures of collocation strength and concluded that the two types of newspapers label and categorize men and women differently.

## Dictionaries and WordNet

Although the construction of dictionaries is not a typical task in corpus linguistics, its resources can nevertheless be used for the construction of dictionaries that represent certain themes. Researchers in sociolinguistics and language acquisition have developed collections of words and expressions that fulfill certain communicative functions in texts. Table 2[10] contains a selection of such collections. Precht (2000), for example, studied English stance markers based on the Longman Corpus of Spoken and Written English and compiled lists of, inter alia, markers of doubt and certainty. Further, Flowerdew (1998) developed a list of cause/effect markers to study how learners of English use these expressions. Management researchers can benefit from considering such wordlists in order to determine whether or not particular communicative functions are present in a text. This can be important when a study is grounded in sensemaking, impression management or identity construction to gain more insights into word choices. It is worth noting that the lists in Table 2 overlap to some extent. It may thus be meaningful to combine some of these lists, for example hedges with downtoners, or emphatics and amplifiers with expressions of certainty.

[9] See Table 3 for details

[10] The table is by no means exhaustive, but only represents word collections known to the author.

---

Table 2

---

Researchers using computer-aided text analysis will sometimes have to compile their own dictionaries when no existing dictionary captures what they seek to study. Self-constructed dictionaries are typically compiled from words in the corpus (e.g. Kabanoff & Holt, 1996; Palmer et al., 1997; Gibson & Zellmer-Bruhn, 2001), which entails that the dictionary is biased towards corpus words. In addition to the strategies for improving the validity of self-constructed dictionaries contributed by Short et al. (2010), another strategy to reduce this bias towards corpus words is to consult WordNet (2010), a lexical database for English containing over 150,000 nouns, verbs, adjectives, and adverbs. WordNet places the conceptual-semantic relationships between these words in a hierarchical, tree-like network (Fellbaum, 1998). WordNet is freely available to the public and can be used online or as a stand-alone application. In order to discover words related to those that have already been chosen for the dictionary, one can look up any given word in the WordNet database, which then returns all words that have lexical sense relations with the search word, including synonyms, antonyms, meronyms (part-to-whole relations), holonyms (whole-to-part relations), hyperonyms (type-to-subtype relations), or hyponyms (subtype-to-type relations). These relations can point to relevant dictionary words in addition to those the researcher has already noted, including both more general and more specific words.

This section has introduced keywords, word dispersion, collocations, linguistic dictionaries, and WordNet. These will be applied in the demonstration example below. Table 3 provides a summary of the calculations and formulae mentioned above.

---

Table 3

---

## DEMONSTRATION EXAMPLE

This section illustrates how the resources of corpus linguistics can be applied to textual data. Letters to shareholders have been chosen as a text corpus for this demonstration. They are a relatively standardized component of annual reports (Bettman & Weitz, 1983) and therefore well-suited for comparative analyses across companies or in longitudinal designs (Daly et al., 2004). Although there is no doubt that a company's communication department plays an active role in the drafting of this letter, the senior management team still has to find its content and language acceptable. Therefore, letters to shareholders can be seen as a reflection of the senior management team's shared cognitions (Abrahamson & Hambrick, 1997). At the same time, most senior management communication can be characterized as "carefully crafted discursive performances" (Ng & De Cock, 2002) and is thus subject to impression management.

Numerous studies in management and organization journals have focused on letters to shareholders, using manual content coding, self-compiled dictionaries, existing dictionaries, or qualitative analysis. These studies have looked at attributions of organizational performance (Staw et al., 1982), causal reasoning patterns to explain organizational performance (Bettman & Weitz, 1983; Tsang, 2002), impression management (Fiol, 1995), senior management attentional foci (D'Aveni & MacMillan, 1990; Yadav et al., 2007), attentional homogeneity within industries (Abrahamson &

Hambrick, 1997), cognitive mental models that characterize strategic groups (Osborne et al., 2001), concealment of poor results (Abrahamson & Park, 1994), CEO commitment to the status quo (McClelland et al., 2010), and espoused values (Daly et al., 2004). This illustrative study uses the techniques of corpus linguistics to examine the discourse in which poor financial results are embedded in shareholder letters. More specifically, the study tries to identify recurring discursive themes in the shareholder letters that serve as frames for the poor results. Themes in a corpus-linguistic study of discourse can comprise both content themes and linguistic themes, e.g. functional aspects of language or semantic themes (Wood & Kroger, 2000; Conrad, 2002).

For this illustrative analysis, a total of 155 letters to shareholders were collected from the largest European and US banks listed on the Forbes 2000 list from 2009. Letters were collected from the years 2008, which was an extremely difficult year for banks, and the year 2006, which represents a normal financial year, i.e. before the financial markets were hit by the financial crisis. Since one tenet of corpus linguistics is that frequencies can only be meaningfully interpreted when compared to other frequencies (Hunston, 2006), the 2008 shareholder letters were studied on the basis of a comparison to letters from 2006. Overall, the sample contains 80 letters from the year 2006 and 75 letters from 2008. The corpus of letters from 2008 contains 130,337 words, while the 2006 corpus contains 119,840 words. The average length of these letters is 1,498 words in 2006 and 1,737 words in 2008.

The first step in the analysis is to identify positive and negative keywords in the 2008 letters by comparing them to the 2006 letters based on Dunning's log-likelihood test. These keywords can point to themes and attentional foci that are dominant in one corpus of letters, but not in the other. This analysis was performed with *WordSmith Tools*. Table 4 shows the top 25 positive and negative keywords with the biggest differences in a 2008-2006 comparison. These differences are all highly significant ($p<.001$). In the 2008 letters, words revolving around the financial crisis dominate the top keyword list: *crisi*, loss*, reces*, difficult*, downturn, declin*, deterior*, reduc*,* and *unpreced**. The top 25 negative keywords of 2008 (i.e. positive keywords of 2006) are dominated by word stems with positive and dynamic connotations among the top ten (*growth, improve*, expand*, success*, achiev**) and more general business terms afterwards. Knowing the context out of which these letters have emerged, the results are fully plausible.

---

Table 4

---

In order to identify the presence and absence of particular themes in the 2008 letters, all keywords identified for the two years with a significance level of $p <.01$ were examined, which included a little over 500 words. Apart from the two obvious themes, i.e. crisis words in 2008 and success words in 2006, a number of other words were identified as potentially indicative of themes in the negative-results shareholder letters (see Table 5). Juilland's D, which indicates the distribution of a word across multiple texts, was calculated to rule out that the words were keywords, only because they were used excessively by a few companies. The calculation of the D values was performed with a spreadsheet package. All these keywords have D values of close to or above 0.8 in the year for which they were identified as keywords, which can be considered evenly distributed. The only exceptions are the words *system* and *could*. Upon closer inspection, it turned out that almost 50% of all instances of *could* and 33% of all instances of *system* were used by one particular company, which is why their D values are so low. They can thus not be considered keywords of the 2008 shareholder letters.

---

Table 5

---

After an examination of the above keywords using KWIC searchers, the keywords identified were grouped into five different themes, with the exception of *could*, *much*, and *however*. The keywords assigned to these five themes were then used as seed words for a search for lexically related words in WordNet's online search interface in order to identify additional related words. Table 6 shows the five themes that were identified, the words they include, the frequency of the themes, and the magnitude of the difference between 2006 and 2008. The WordNet search contributed the words *hazard*, cautio*,* and *conservative* to the theme 'Risk management' as well as the words *context* and *events* to the theme 'Environment' and the words *assur*, reassure* and hope*/hoping* to the theme 'Reassurance'. Additional words identified for the themes 'Strategy' and 'People' include *goal, direction, focus, clients,* and *employees*. The frequency of the themes differs significantly ($p<0.001$) between the 2006 and the 2008 letters. The high D values of the themes indicate that the themes are sufficiently equally distributed among the letters.

---

Table 6

---

The first theme identified based on keywords in the 2008 letters include words related to the environment, which serves as a justifying frame for the poor results. The exemplary sentences in Table 7 illustrate how words belonging to this theme are used in the 2008 shareholder letters. In addition, the 2008 shareholder letters contain risk management as a significant theme, with the banks stressing their increased focus on risk. Another theme is reassurance, which contains lexical items that express confidence and optimism about the future. In addition to the presence of environment, risk management, and reassurance, the 2008 letters are also characterized by an absence of discourse about people and strategy relative to the letters from the year 2006. These themes apparently had to give way to themes that are related to the negative results communicated in the letters. Overall, the keyword comparison and the subsequent identification of themes with the help of KWIC searches and WordNet have revealed three themes that received relatively more attention and two themes that received relatively less attention in the 2008 shareholders letters compared to the year 2006.

---

Table 7

---

As a second lexical exploration, an open search for collocations was conducted with WordStat, without the specification of a node word. This collocation search was performed for the entire corpus of shareholder letters as well as for the 2008 letters and the 2006 letters individually. These searches returned a large number of word combinations containing only function words (e.g. *in which we*) or financial phrases (e.g. *in the fourth quarter*). A number of collocations seemed worth examining further, though. They are listed in Table 8, together with their total frequency in the two corpora, their D values as well as their z-scores, MI scores, and log-likelihood test statistics $G^2$, all of which indicate collocation strength. The z-scores, MI scores, and $G^2$ were calculated with a spreadsheet package. The

results of the three measures are not unanimous, as expected. When ranking the collocations identified according to the three different scores, the results produced by MI scores and z-scores are consistent, while the log-likelihood test produces different results in particular for the top four collocations. The eight collocations identified differ substantially in terms of strength, but can be classified as strong collocations, apart from *all of*, which is the least strong collocation according to all three algorithms and has substantially lower scores than the other seven collocations.

---

Table 8

---

Two collocations were found to differ significantly in terms of frequency between 2006 and 2008 letters: *we believe (that/the/we)* and *many of (our/the)* were found more frequently in 2008 letters. While *we believe (that/the/we)* is a strong collocation, *many of (our/the)* is not, but is still relevant because of the significant differences. These two collocations were then explored qualitatively with KWIC searches in WordStat. The other strong collocations (*around the world, would like, will continue*) seem to be standard features of shareholder letters in general, as they do not differ in terms of frequency between the two years. The KWIC search revealed that 65% of all instances of *we believe* in the 2008 letters denote a form of reassurance and trust restoration. For example:

- *Based on what we know today,* ***we believe we****'ll have the opportunity to earn back a substantial portion of these write-downs*
- ***We believe that*** *we have the right business model to benefit from these changes*
- ***We believe we*** *have corrected for the underwriting mistakes of the past.*

This, together with the fact that this collocation occurs significantly more often in the 2008 letters, makes this collocation another facet of the theme 'Reassurance', which was identified earlier.

A KWIC search for *many of* revealed that the expression is used for comparisons with other companies in a self-congratulatory or a justifying manner, in addition to simply denoting a quantity. For example:

- *We believe we are better positioned than* ***many of our competitors*** *(2006)*
- ***Unlike many of*** *our competitors, in the financial services industry, we are well-apitalized. (2008)*
- ***Along with many of*** *our peers, Lincoln National Corporation ("Lincoln") faced elevated investment losses*

After reading the KWIC results of the statements in which companies made comparisons to other companies, 'Comparison' was added as a theme. To study this theme in more detail, words used in those statements were added to the theme, including: *like*, *unlike*, *along with*, *most of*, and *position**. This theme is found significantly more often ($G^2$=32.67, $p<0.001$) in the 2008 letters than in the 2006 letters and well distributed (D= 0.87). Overall, the collocation search expanded the theme 'Reassurance' and contributed the theme 'Comparison'.

The third inquiry compared the corpus of 2008 shareholder letters against some of the linguistic wordlists presented in Table 2. In view of the poor results that were communicated in the 2008 shareholder letters, a number of language features can be expected to be found in the negative-results letters. First, expressions of reason/cause and results/effects (Flowerdew, 1998) explain causalities and

may be relevant, given that poor results are generally attributed to outside forces and external events rather than to oneself (Schlenker, 1980), which has also been found in annual reports (Aerts, 1994). Second, extreme-case formulations are typically used for defenses and justifications, when one's legitimacy is challenged (Edwards, 2000). Therefore, they can be expected to be found more frequently in 2008 shareholder letters than in 2006 shareholder letters. Third, speakers/writers can increase the intensity of a statement and express confidence with markers of certainty (including amplifiers and emphatics). This is expected to be an important feature of shareholder letters commenting on poor results and seeking to provide reassurance. Ultimately, downtoners (including hedges) can soften the impact of a statement but also indicate a lack of confidence (Holmes, 1982; Hinkel, 2003a). Downtoners are expected to be used in the 2008 shareholder letters to make poor results seem less poor. However, since letters to shareholders are carefully crafted documents, it is not expected that downtoners are used in a manner that expresses a lack of confidence.

Following these hypotheses, five dictionaries were built, based on the corresponding wordlists presented in Table 2: *Cause*, *Certainty*, *Downtoners*, *Extreme Case*, and *Results*. A log-likelihood test conducted with WordStat indicated that all dictionaries occur significantly more frequently in the 2008 letters than in the 2006 letters (see Table 9). The corresponding D values, which were calculated with a spreadsheet package, are close to or above 0.8, with the exception of *Downtoners.* These results suggest that the need to comment on disastrous financial results in shareholder letters is connected with the use of all of the above types of linguistic features. With the smallest difference and the lowest D value, downtoners seem to be the least characteristic, whereas expressions of reason/cause and markers of certainty seem to be the most prominent ones in the 2008 letters to shareholders.

---

Table 9

---

In order to study particularly prominent words from the dictionaries in more detail, the entries of each dictionary were examined individually and those words on which the 2008 letters and the 2006 letters differ significantly were identified. They all occur significantly more frequently in the 2008 letters. Table 10 shows these words together with the magnitude of the difference ($G^2$) and the distribution among the 2008 letters (D). Since *caus**, *certain/ly* and *could* (see results of keyword analysis above) are not well distributed, they can – by themselves – not be seen as characteristic words of the 2008 shareholder letters, but only as part of the dictionaries. The other words were examined more closely in KWIC searches. The certainty markers *even, much*, and *never,* as well as the extreme-case formulations *everything* (e.g. *everything we do, everything possible, everything in our power*) and *no* (e.g. *no question, no exception, in no way*) are characteristic of 2008 shareholder letters both individually and as part of the dictionaries. The dictionary of certainty markers and the dictionary of extreme-case formulations contribute to the theme 'Reassurance', because they communicate confidence and seek to eliminate doubts. The dictionary of reason/cause and the dictionary of results/effect form a theme of their own entitled 'Attribution', which includes all those words and expressions needed to explain how the poor results came about.

---

Table 10

---

Overall, the corpus-linguistic analysis has revealed seven themes, five of which are more prominent in the 2008 shareholder letters (Environment, Risk Management, Reassurance, Comparison, Attribution) compared to the 2006 letters, and two of which are less prominent (Strategy, People). Figure 1 summarizes the steps taken to arrive at these themes. First, keywords were identified based on a log-likelihood test, followed by an examination of the dispersion of the keywords, KWIC searches, and WordNet searches. The collocation analysis began with an open search for collocations, followed by the calculation of their strength. Then their frequencies were compared in the two corpora together with their dispersion measures. This was followed by KWIC searches with the most noteworthy ones. Lastly, linguistic word collections were used as dictionaries. Again, frequencies were compared, word dispersion measures were examined, and the most prominent words were examined with KWIC searches. Within each of these three strands, both qualitative and quantitative explorations were used to arrive at valid findings. The findings (i.e. the themes) were derived both inductively and deductively, including content themes and language themes. Together, they give a rich picture of the negative-results discourse of the 2008 letters to shareholders.

---

Figure 1: Summary of Analytical Steps

---

The above approach represents one possible way of combining the techniques of corpus linguistics into an inquiry. The demonstration example does not specifically draw on computer-aided content analysis or computer-aided interpretive textual analysis in order to show how a pure corpus-linguistic analysis is conducted. But, clearly, there are overlaps. First, some of the themes identified are content themes, resembling the content dictionaries employed in computer-aided content analysis. However, the corpus linguistic procedure identifies noteworthy themes rather than check the presence of existing dictionaries such as DICTION. Second, collocations are also used in computer-aided interpretive textual analysis but only in qualitative explorations of particular node words. In the corpus-linguistic approach presented in this paper, an open search for collocations is made in order to identify recurring and noteworthy patterns in text. Third, the linguistic wordlists resemble a content-analytical methodology with an existing dictionary, but contain words that fulfill meta-communicative purposes rather than content words.

## DISCUSSION

In view of the lack of interaction between linguistics and the social sciences regarding text analysis (Markoff et al., 1974; Roberts, 1989; Bernard & Ryan, 1998; Popping, 2000), this paper has set out to demonstrate how the resources of corpus linguistics can be meaningfully applied in the social sciences. This exemplary analysis of letters to shareholders has demonstrated how the use of corpus-linguistic analysis techniques can provide insights that computer-aided content analysis or computer-aided interpretive textual analysis alone would not provide. More specifically, these pertain to the comparison of corpora by means of keywords, the dispersion of words within a set of corpora, the identification of strong collocations, and the enhancement of self-constructed dictionaries with WordNet, all of which were employed in the demonstration example. Thus, corpus linguistics makes a contribution to organizational research methods in four areas: First, corpus linguistics has the techniques to identify and quantify recurring patterns in textual data. Second, corpus linguistics highlights the importance of examining collocations and multi-word expressions rather than looking

at individual words only. Third, corpus linguistics can provide techniques for the comparison of one's own corpus with a large public corpus as well as for the comparison of different texts within a corpus. Ultimately, corpus linguistics contributes methodological innovations in the form of new or improved tools and resources for exploring and handling textual data.

Corpus linguistics can be of value for two types of studies in the field of management and organization. First, corpus-linguistic techniques can be more insightful than content analysis for quantitative, positivist studies drawing on large samples of texts from the same genre, e.g. mission statements, letters to shareholders, proxy statements, annual reports, CSR reports, corporate self-presentations on websites, executive speeches, e-mails, press releases or news articles. In these studies, the focus is on manifest content and surface features of texts either in a snapshot analysis or a longitudinal design. With corpus-linguistic techniques, lexical patterns can be identified, quantified and compared across large samples to find commonalities. Alternatively, such studies can use large general-language corpora when general texts about organizations are needed (c.f. Cornelissen, 2008) or more specialized corpora of professional English either as the main data source or as a reference corpus. A second stream of research to which corpus linguistics can add value is discursive or narrative studies on organizations, for example on sensemaking and sensegiving, framing, emotions, or impression management. In addition to what computer-aided interpretive textual analysis can provide to such studies, corpus-linguistic techniques can add more elaborate measures of identifying interesting themes through collocations and their strength, the identification of keywords, and the calculation of word dispersion measures. Both positivist and interpretive studies can draw on the resources provided by researchers in linguistics, including WordNet for the identification of potentially interesting words or linguistic wordlists such as those used in the demonstration example. These techniques and resources can open up additional inquiry opportunities or refine existing ones.

The absence of a stringent methodology behind corpus linguistics is a strength when it comes to incorporating its techniques into other methodologies. Because it consists of a loose bundle of analysis techniques, corpus linguistics is flexible enough to be embedded into CATA or can be used as an alternative to CATA altogether. For example, one could strengthen a content analysis based on existing dictionaries with word dispersion measures. Further, a content analysis based on self-constructed dictionaries could be enriched with WordNet searches. An interpretive analysis could be enhanced with open collocation searches and collocation measures as well as linguistic wordlists. However, corpus linguistics is not without limitations. First, some of the techniques presented in this paper can only be applied to English text corpora, including linguistic wordlists, publicly available corpora and WordNet. Second, applying corpus linguistics requires the researcher to have a good understanding of language and its irregularities, in particular spelling variants and words with multiple meanings, both of which can severely distort one's findings, if they are not accounted for in the analyses. Third, a researcher's subjectivity is an inevitable element of any corpus-linguistic analysis, not only because of its qualitative elements, but also because the researcher has to take decisions about corpus building, the selection of analysis steps, the construction of dictionaries, and the amount of validation work (cf. Baker et al., 2008). Subjectivity is also inherent in the interpretation of results, when researcher input is required for setting cut-off points for keywords or for the values of dispersion measures, as no firmly established standards exist yet. Therefore, constant checking, reflecting, critiquing, contextualizing, refining and adapting have to be integral parts of any corpus study in order to minimize subjectivity and ambiguities. Then only can corpus linguistics provide management scholars with powerful methodological resources.

## CONCLUSION

This paper has introduced corpus linguistics as an enhancement of or an alternative to computer-aided text analysis. Based on an exemplary analysis of letters to shareholders, the paper has demonstrated what the resources of corpus linguistics can contribute to organizational research methods. Given that corpus linguistics has been developed by scholars in the field of language studies, drawing on their expertise when it comes to exploring textual data can only be beneficial in future studies in the field of management and organization. With its methodological innovations for the identification of recurring lexical patterns, the comparison of corpora, and the enhancement of dictionaries, the field of corpus linguistics can fertilize the field of computer-aided text analysis, if researchers are willing to broaden their methodological repertoire with its techniques.

## APPENDIX

**CORPUS LINGUISTICS SOFTWARE TOOLS**

| | *Morph Adorner* | *R corpus* | *SCP* | *Textpack* | *WMatrix* | *Word Cruncher* | *WordList Creator* | *WordSmith Tools* | *WordStat* |
|---|---|---|---|---|---|---|---|---|---|
| Frequency List | | x | | x | x | x | x | x | x |
| KWIC | | x | x | x | x | x | | x | x |
| Keywords | | | | x | x | | | x | x |
| Dictionary | | | | x | | x | | | x |
| Collocation | | x | | | x | x | | | x |
| Lemmatization | x | | | | | | | | x |
| Statistics | | x | | | | | | | x |
| Word dispersion | | | | | | | | | x |
| Dispersion plot | | | | | | | | x | |

*Morph Adorner*: http://morphadorner.northwestern.edu/morphadorner/download/
*R corpus*: http://cran.r-project.org/web/packages/corpora/
*SCP:* http://www.textworld.com/scp/
*Textpack:* http://www.gesis.org/en/services/methods/software/textpack/
*WMatrix:* http://ucrel.lancs.ac.uk/wmatrix/
*Word Cruncher:* http://www.wordcruncher.com
*WordList Creator:* http://www.safe-install.com/programs/word-list-creator.html
*WordSmith Tools:* http://www.lexically.net/wordsmith/
*WordStat:* http://www.provalisresearch.com/wordstat/Wordstat.html

Table 1: Comparison of the three Approaches

| | **Computer-Aided Content Analysis** | | **Computer-aided interpretive textual analysis** | **Corpus Linguistics** | |
|---|---|---|---|---|---|
| **Epistemological assumptions** | Positivist | | Interpretive | Positivist | Interpretive |
| **Main focus** | Concepts | | Meanings | Lexical patterns and themes | |
| **Inferences** | Inductive/ deductive | Deductive | Inductive | Deductive | Inductive |
| **Main techniques** | Self-constructed dictionary | Existing dictionary | • KWIC<br>• Collocations<br>• Self-constructed dictionary | • KWIC<br>• Collocations<br>• Word distribution<br>• Corpus comparisons<br>• WordNet | |
| **Research questions** | • Presence of concepts<br>• Positive/negative sentiment in texts<br>• Co-occurrence of concepts in texts | | • Language in the construction of reality<br>• Meaning creation in texts<br>• Relations between texts | • Comparison of textual patterns with other textual patterns in the same corpus or in other corpora<br>• Comparison of textual patterns with contextual patterns | |
| **Software** | e.g. Diction 5.0, General Inquirer | | e.g. NVivo, QDA Miner | e.g. WordStat, WordSmith Tools | |

Table 2: Linguistic Wordlists

| **Language Phenomenon** | **Examples** | **Sources** |
|---|---|---|
| Expressions of reason/cause | *lead to, due to* | (Flowerdew, 1998) |
| Expressions of results/effects | *arise from, therefore* | (Flowerdew, 1998) |
| Hedges (Presence of uncertainty) | *more or less, almost* | (Biber, 1991; Precht, 2000) |
| Downtoners (Degree of uncertainty) | *nearly, partly, slightly* | (Biber, 1991; Hinkel, 2003a; Rizomilioti, 2006) |
| Emphatics (Presence of certainty) | *for sure, a lot, really* | (Biber, 1991; Hinkel, 2003a) |
| Amplifiers (Degree of certainty) | *absolutely, completely* | (Biber, 1991; Hinkel, 2003a) |
| Expressions of certainty | *clearly, undoubtedly* | (Precht, 2000; Rizomilioti, 2006) |
| Expressions of importance | *very, highly, really* | (Precht, 2000) |
| Extreme-case formulations | *all, none, best* | (Pomerantz, 1986; Edwards, 2000; Norrick, 2004) |
| Public verbs (Observable actions) | *assert, claim, say* | (Quirk et al., 1985; Hinkel, 2003b) |
| Private verbs (Mental states) | *know, think, believe* | (Quirk et al., 1985; Hinkel, 2003b) |
| Nouns indicating abstraction | **dom*, **ity*, **ness* | (Mergenthaler, 1996) |

Table 3: Summary of Calculations

<table>
<tr><th colspan="2">Measure</th><th>Calculation</th><th>Components</th><th>Purpose</th></tr>
<tr><td rowspan="2">Word dispersion</td><td>Spread</td><td>$\frac{w}{N}$</td><td>w = Number of texts that contain a word<br>N = Total number of texts</td><td rowspan="2">Examining how the instances of a word are distributed among the texts in the corpus</td></tr>
<tr><td>Juilland's D</td><td>$1-\frac{s}{x\sqrt{N-1}}$</td><td>N = Total number of texts<br>x = Mean frequency of a word across N texts<br>s = Standard deviation of a word in N texts</td></tr>
<tr><td>Keyword identification</td><td>Log-likelihood test statistic $G^2$</td><td colspan="2">Log likelihood calculator (Excel file):<br>http://ucrel.lancs.ac.uk/people/paul/LL.xls</td><td>Identifying words that are salient in texts compared to other texts</td></tr>
<tr><td rowspan="3">Collocation strength</td><td>MI score</td><td>$\log_2(\frac{n\times f_{nc}}{f_{node}\times f_{coll}})$</td><td rowspan="2">$f_{node}$ = Frequency of the node<br>$f_{coll}$ = Frequency of the collocate<br>$f_{nc}$ = Frequency of the collocation between node and collocate<br>n = Total number of words<br>$E = p\times f_{node}$<br>$p=\frac{f_{coll}}{n-f_{node}}$</td><td rowspan="3">Examining lexical units rather than individual words for the study of semantic preferences and characteristic discourse features in texts</td></tr>
<tr><td>z-score</td><td>$\frac{f_{nc}-E}{\sqrt{E\times(1-p)}}$</td></tr>
<tr><td>Log-likelihood test statistic $G^2$</td><td colspan="2">Log likelihood calculator (Excel file):<br>http://ucrel.lancs.ac.uk/people/paul/LL.xls</td></tr>
</table>

Table 4: Top 25 Keywords in Shareholder Letters

| **2008 Keywords (word stems only)** | **2006 Keywords (word stems only)** |
|---|---|
| CRISI | GROWTH |
| LOSS | IMPROV |
| CAPIT | TARGET |
| LIQUID | COMMERC |
| ECONOM | EXPAND |
| RECES | DEVELOP |
| DIFFICULT | SUCCESS |
| FINANCI | BRAND |
| CONDITION | ACHIEV |
| LOAN | PARTNER |
| NOT | PLAN |
| DOWNTURN | OPPORTUN |
| SEVER | COMPETI |
| DECLIN | PRODUCT |
| RESERV | YIELD |
| DETERIOR | STRATEGI |
| CREDIT | LAW |
| REDUC | OBJECT |
| UNPRECED | RETURN |
| CAUS | PROMOTION |
| RISK | CONSOLID |
| SHEET | SALE |
| RESILI | GROUP |
| UNCERTAINTI | COMPLIANC |
| TIME | INNOV |

Table 5: Selected Keywords in Shareholder Letters

| **Keyword** | **Frequency 2006** | **Frequency 2008** | **Log-likelihood G2** | **$D_{2008}$** | **$D_{2006}$** |
|---|---|---|---|---|---|
| System | 17 | **112** | 84.54 *** | 0.62 | |
| Conditions | 39 | **129** | 61.80 *** | 0.85 | |
| Uncertain/ty/ties | 8 | **66** | 55.71 *** | 0.85 | |
| Confidence | 31 | **98** | 41.02 *** | 0.79 | |
| Time/s | 215 | **336** | 35.17 *** | 0.82 | |
| Risk | 169 | **277** | 33.85 *** | 0.77 | |
| Environment | 119 | **205** | 28.99 *** | 0.87 | |
| Circumstances | 4 | **31** | 25.42 *** | 0.80 | |
| Exposure | 8 | **35** | 20.08 *** | 0.79 | |
| Could | 45 | **83** | 14.04 ** | 0.46 | |
| Industry | 172 | **226** | 11.25 ** | 0.83 | |
| Much | 61 | **95** | 9.83 ** | 0.75 | |
| However | 66 | **101** | 9.82 ** | 0.81 | |
| Strateg/y/ies/ic | **324** | 159 | 47.37 *** | | 0.89 |
| Customer/s | **635** | 416 | 32.93 *** | | 0.81 |
| Objective/s | **64** | 18 | 24.48 *** | | 0.76 |
| Serv/e/ing/ed | **143** | 85 | 11.41 ** | | 0.86 |

Table 6: Themes in Shareholder Letters

| **Theme** | **Words** | **2006** | **2008** | **$G^2$** | **$D_{2008}$** | **$D_{2006}$** |
|---|---|---|---|---|---|---|
| *1 Environment* | System<br>Industry<br>Environment<br>Conditions<br>Circumstances<br>Context<br>Events<br>Uncertain/ty/ties<br>Time/s | 412 | **961** | 262.64 *** | 0.88 | |
| *2 Risk management* | Risk<br>Exposure<br>Conservative<br>Hazard/s/ous<br>Cauti/on/ous/ousness | 264 | **470** | 72.66 *** | 0.78 | |
| *3 Reassurance* | Confiden/t/ce<br>Assur/e/ed/ing/ance<br>Reassur/e/ed/ance/ing<br>Hop/e/ing/ed/eful | 83 | **165** | 33.18*** | 0.79 | |
| *4 Strategy* | Strateg/y/ies/ic<br>Goal/s<br>Focus<br>Objective/s<br>Direction | **592** | 331 | 58.94*** | | 0.90 |
| *5 People* | Customers<br>Clients<br>Employees<br>Serv/e/ed/ing | **1201** | 829 | 49.77 *** | | 0.88 |

Table 7: Exemplary Statements for the Themes

| **Theme** | **Exemplary statements** |
|---|---|
| Environment | – *Under the **circumstances**, this is an excellent performance.*<br>– *The reality is that **conditions** were the worst we have seen in many years, and our results were disappointing.*<br>– *In that **context** we are proud of, but not nearly satisfied by, our progress in creating value*<br>– *We were not, however, immune to the **environment** in 2008.*<br>– *... repair a balance sheet that might have been stretched by unanticipated market **events***<br>– *The financial services **industry** has undergone transformative, wrenching change*<br>– *But when the panic started, it was too much for the **system***<br>– *The banking industry is experiencing very difficult **times**.*<br>– *As we began 2009, we continued to face **uncertainty**.* |
| Risk management | – *We are maintaining a **cautious** stance toward lending*<br>– *We are being very **conservative** with our capital*<br>– *But, we monitored our **exposure** carefully and reduced it aggressively*<br>– *We also responded proactively to address the **hazards** in the capital markets.*<br>– *However, the spending and credit disruptions we experienced caused us to refine our **risk** controls* |
| Reassurance | – *I hope, after reading this letter, you will share my **confidence** in our ability to build a stronger, more vibrant company for the future*<br>– *I can **assure** you, we are committed to restoring Citi to profitability as quickly as possible.*<br>– *Finally, one key indicator **reassures** us particularly.*<br>– *We **hope** to attain a sustainable 15% by the year 2010.* |

Table 8: Collocations in Shareholder Letters

| **Collocation** | **Total frequency** | **D** | **Collocation strength (Rank in brackets)** | | |
|---|---|---|---|---|---|
| | | | **z score** | **MI score** | **$G^2$** |
| *around the world* | 70 | 0.71 | 188.99 (1) | 9.00 (1) | 818.91 (3) |
| *would like (to)* | 62 | 0.75 | 140.41 (2) | 8.32 (2) | 653.15 (4) |
| *(we) will continue (to)* | 185 | 0.85 | 121.13 (3) | 6.35 (3) | 1470.73 (2) |
| *we believe (that/the/we)* | 214 * | 0.79 | 79.99 (4) | 5.02 (4) | 1543.55 (1) |
| *one of (the)* | 174 | 0.81 | 36.90 (5) | 3.25 (6) | 526.53 (5) |
| *our ability (to)* | 75 | 0.79 | 35.99 (6) | 4.30 (5) | 400.03 (6) |
| *many of (our/the)* | 93 * | 0.76 | 22.70 (7) | 2.85 (7) | 227.28 (7) |
| *all of (our)* | 54 | 0.73 | 6.72 (8) | 1.26 (8) | 33.20 (8) |

* = significant difference between 2006 and 2008 shareholder letters

Table 9: Linguistic Wordlists in Shareholder Letters

| **Wordlists** | **Examples** | **Frequency 2006** | **Frequency 2008** | **$G^2$** | **D** |
|---|---|---|---|---|---|
| Cause | *caus*, because, trigge*, due to, underlying, le(a)d/ing to* | 269 | 411 | 39.75*** | 0.78 |
| Certainty | *certain/ly, even, much, never, extreme/ly, clearly, sharp/ly* | 1,599 | 1,801 | 28.59*** | 0.84 |
| Result | *effect/s, then, so that, thus, as a result, resulting, therefore* | 181 | 240 | 12.54** | 0.84 |
| Extreme case | *everything, no, always, best, everyone, ever, nothing, none* | 897 | 972 | 9.80** | 0.84 |
| Downtoners | *could, relatively, probably, likely, appear*, slightly, possible, partly* | 851 | 903 | 6.72* | 0.75 |

Table 10: Words from Linguistic Dictionaries

| **Words** | **Dictionary** | **Frequency 2008** | **$G^2$** | **$D_{2008}$** |
|---|---|---|---|---|
| *caus** | Cause | 61 | 44.07 *** | 0.45 |
| *certain/ly* | Certainty | 70 | 20.37 *** | 0.69 |
| *could* | Downtoners | 83 | 14.04 ** | 0.46 |
| *even* | Certainty | 139 | 11.89 ** | 0.80 |
| *everything* | Extreme Case | 22 | 10.78 ** | 0.73 |
| *effect/s* | Results | 54 | 10.42 ** | 0.70 |
| *much* | Certainty | 95 | 9.83 ** | 0.75 |
| *no* | Extreme Case | 120 | 8.63 * | 0.80 |
| *never* | Certainty | 34 | 6.93 * | 0.67 |

## REFERENCES


Abrahamson, E., & Fairchild, G. (1999). Management fashion: Lifecycles, triggers, and collective learning processes. *Administrative Science Quarterly, 44*, 708-740.

Abrahamson, E., & Hambrick, D. C. (1997). Attentional homogeneity in industries: The effect of discretion. *Journal of Organizational Behavior, 18*, 513-532.

Abrahamson, E., & Park, C. (1994). Concealment of negative organizational outcomes: An agency theory perspective. *Academy of Management Journal, 37*(5), 1302-1334.

Aerts, W. (1994). On the use of accounting logic as an explanatory category in narrative accounting disclosures. *Accounting, Organizations and Society, 19*(4-5), 337-353.

Anderson, W. (2008). Corpus linguistics in the UK: Resources for sociolinguistic research. *Language and Linguistics Compass, 2*(2), 352-371.

Bailey, K. D. (1994). *Methods of social research*. New York: Free Press.

Baker, P. (2004). Querying keywords. Questions of difference, frequency, and sense in keywords analysis. *Journal of English Linguistics, 32*(4), 346-359.

Baker, P. (2006). *Using corpora in discourse analysis*. London: Continuum.

Baker, P. (2009). The BE06 Corpus of British English and recent language change. *International Journal of Corpus Linguistics, 14*(3), 312-337.

Baker, P., Gabrielatos, C., Khosravinik, M., Kryzanowski, M., McEnery, T., & Wodak, R. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourse of refugees and asylum seekers in the UK press. *Discourse & Society, 19*(3), 273-306.

Barlow, M. (2011). Corpus linguistics and theoretical linguistics. *International Journal of Corpus Linguistics, 16*(1), 3-44.

Bartsch, S. (2004). *Structural and functional properties of collocations in English: A corpus study of lexical and pragmatic constraints on lexical co-occurrence*. Tübingen: Gunter Narr.

Bauer, M. W. (2000). Classical content analysis: A review. In M. W. Bauer & G. Gaskell (Eds.), *Qualitative researching with text, image and sound* (pp. 131-151). London: Sage.

Bernard, H. R., & Ryan, G. (1998). Text analysis: Qualitative and quantitative methods. In H. R. Bernard (Ed.), *Handbook of methods in cultural anthropology* (pp. 595-645). London: Sage.

Bettman, J. R., & Weitz, B. A. (1983). Attributions in the board room: Causal reasoning in corporate annual reports. *Administrative Science Quarterly, 28*, 165-183.

Biber, D. (1991). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D. (1996). Investigating language use through corpus-based analyses of association patterns. *International Journal of Corpus Linguistics, 1*(2), 171-197.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Brett, J. M., Olekalns, M., Friedman, R., Goates, N., & Anderson, C. (2007). Sticks and stones: Language, face, and online dispute resolution. *Academy of Management Journal, 50*(1), 85-99.

Caldas-Coulthard, C. R., & Moon, R. (2010). 'Curvy, hunky, kinky': Using corpora as tools for critical analysis. *Discourse & Society, 21*(2), 99-133.

Caska, B. A., Kelley, K., & Christensen, E. W. (1992). Organizational needs assessment. In K. Kelley (Ed.), *Issues, theory, and research in industrial/organizational psychology* (pp. 229-256). Amsterdam: North Holland.

Church, K. W., Gale, W., Hanks, P., Hindle, D., & Moon, R. (1994). Lexical substitutability. In B. T. S. Atkins & A. Zampolli (Eds.), *Computational approaches to the lexicon* (pp. 153-180). Oxford: Oxford University Press.

Conrad, S. (2002). Corpus linguistic approaches for discourse analysis. *Annual Review of Applied Linguistics, 22*, 75-95.

Cornelissen, J. (2008). Metonymy in language about organizations: A corpus-based study of company names. *Journal of Management Studies, 45*(1), 79-99.

D'Aveni, R. A., & MacMillan, I. C. (1990). Crisis and the content of managerial communications: A study of the focus of attention of top managers in surviving and failing firms. *Administrative Science Quarterly, 35*, 634-657.

Daly, J. P., Pouder, R. W., & Kabanoff, B. (2004). The effects of initial differences in firms' espoused values on their postmerger performance. *Journal of Applied Behavioral Science, 40*, 323-343.

Doucet, L., & Jehn, K. A. (1997). Analyzing harsh words in a sensitive setting: American expatriates in communist China. *Journal of Organizational Behavior, 18*, 559-582.

Dowling, G. R., & Kabanoff, B. (1996). Computer-aided content analysis: What do 240 advertising slogans have in common? *Marketing Letters, 7*(1), 63-75.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics, 19*(1), 61-74.

Duriau, V. J., Reger, R. K., & Pfarrer, M. D. (2007). A content analysis of the content analysis literature in organization studies. Research themes, data sources, and methodological refinements. *Organizational Research Methods, 10*(1), 5-34.

Edwards, D. (2000). Extreme case formulations: Softeners, investment, and doing nonliteral. *Research on Language and Social Interaction, 33*(4), 347-373.

Fellbaum, C. (Ed.). (1998). *WordNet. An electronic lexical database*. Cambridge: MIT Press.

Fiol, C. M. (1995). Corporate communications: Comparing executives' private and public statements. *Academy of Management Journal, 38*(2), 522-536.

Flowerdew, L. (1998). Integrating 'expert' and 'interlanguage' computer corpora findings on causality: Discoveries for teachers and students. *English for Specific Purposes, 17*(4), 329-345.

Gabrielatos, C., Torgersen, E. N., Hoffmann, S., & Fox, S. (2010). A corpus-based sociolinguistic study of indefinite article forms in London English. *Journal of English Linguistics, 38*(4), 297-334.

Gephart, R. P. (1993). The textual approach: Risk and blame in disaster sensemaking. *Academy of Management Journal, 36*(6), 1465-1514.

Gephart, R. P. (1997). Hazardous measures: An interpretative textual analysis of quantitative sensemaking during crises. *Journal of Organizational Behavior, 18*, 583-622.

Gephart, R. P. (2003). Grounded theory and the integration of qualitative and quantitative research. In F. Dansereau & F. J. Zammarino (Eds.), *Multi-level issues in organizational behavior and strategy* (pp. 113-125). Oxford: Elsevier.

Gephart, R. P., & Wolfe, R. A. (1989). Qualitative data analysis: Three microcomputer-supported approaches. *Academy of Management Best Paper Proceedings*, 382-386.

Gibson, C. B., & Zellmer-Bruhn, M. E. (2001). Metaphors and meaning: An intercultural analysis of the concept of teamwork. *Administrative Science Quarterly, 46*, 274-303.

Gries, S. T. (2009). *Quantitative corpus linguistics with R: A practical introduction*. New York: Routledge.

Hamilton, C., Adolphs, S., & Nerlich, B. (2007). The meanings of 'risk': A view from corpus linguistics. *Discourse & Society, 18*(2), 163-181.

Hinkel, E. (2003a). Adverbial markers and tone in L1 and L2 students' writing. *Journal of Pragmatics, 35*, 1049-1068.

Hinkel, E. (2003b). Simplicity without elegance: Features of sentences in L1 and L2 academic texts. *TESOL Quarterly, 37*(2), 275-301.

Holmes, J. (1982). Expressing doubt and certainty in English. *RELC Journal, 13*(2), 9-28.

Holsti, O. R. (1968). Content analysis. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology* (pp. 596-692). Reading, MA: Addison-Wesley.

Hunston, S. (2006). Corpus linguistics. In K. Brown (Ed.), *The encyclopedia of language and linguistics* (pp. 234-248): Elsevier.

Jablin, F. M., & Putnam, L. L. (Eds.). (2001). *The new handbook of organizational communication*. Thousand Oaks: Sage.

Janasik, N., Honkela, T., & Bruun, H. (2009). Text mining in qualitative research: Application of an unsupervised learning method. *Organizational Research Methods, 12*(3), 436-460.

Johnson, S., Culpeper, J., & Suhr, S. (2003). From 'politically correct councillors' to 'Blairite nonsense': Discourses of 'political correctness' in three British newspapers. *Discourse & Society, 14*(1), 29-47.

Kabanoff, B. (1997). Computers can read as well as count: Computer-aided text analysis in organizational research. *Journal of Organizational Behavior, 18*, 507-511.

Kabanoff, B., & Brown, S. (2008). Knowledge structures of prospectors, analyzers, and defenders: Content, structure, stability, and performance *Strategic Management Journal, 29*, 149-171.

Kabanoff, B., & Holt, J. (1996). Changes in the espoused values of Australian organizations 1986-1990. *Journal of Organizational Behavior, 17*, 201-219.

Kabanoff, B., Waldersee, R., & Cohen, M. (1995). Espoused values and organizational change themes. *Academy of Management Journal, 38*(4), 1075-1104.

Kachru, Y. (2008). Language variation and corpus linguistics. *World Englishes, 27*(1), 1-8.

Kelle, U. (1995). An overview of computer-aided methods in qualitative research. In U. Kelle (Ed.), *Computer-aided qualitative data analysis. Theory, methods and practice* (pp. 1-18). London: Sage.

Koteyko, N. (2010). Mining the internet for linguistic and social data: An analysis of 'carbon compounds' in Web feeds. *Discourse & Society, 21*(6), 665-674.

Krippendorff, K. (1980). *Content analysis. An introduction to its methodology*. Beverly Hills: Sage.

Krippendorff, K. (2004). *Content analysis. An introduction to its methodology* (2nd ed.). Beverly Hills, CA: Sage.

Kucera, H., & Francis, W. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Lebart, L., Salem, A., & Berry, L. (1998). *Exploring textual data*. Dordrecht: Kluwer.

Lee, D. Y. W. (2008). Corpora and discourse analysis: New ways of doing old things. In V. K. Bhatia, L. Flowerdew & R. H. Jones (Eds.), *Advances in discourse studies* (pp. 86-99). London: Routledge.

Leech, G. (1992). Corpora and theories of linguistic performance. In J. Svartvik (Ed.), *Directions in corpus linguistics* (pp. 105-122). Berlin: Mouton.

Leech, G., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English based on the British National Corpus*. Harlow, UK: Pearson.

Lindquist, H. (2009). *Corpus linguistics and the description of English*. Edinburgh: Edinburgh University Press.

Liu, D. (2010). Is it a chief, main, major, primary, or principal concern? A corpus-based behavioral profile study of the near-synonyms. *International Journal of Corpus Linguistics, 15*(1), 56-87.

Lüdeling, A., & Kytö, M. (2008). *Corpus linguistics: An international handbook*. Berlin: Mouton de Gruyter.

Mahlberg, M. (2005). *English general nouns: A corpus theoretical approach*. Amsterdam: John Benjamins.

Manning, C., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Markoff, J., Shapiro, G., & Weitman, S. R. (1974). Toward the integration of content analysis and general methodology. In D. R. Heise (Ed.), *Sociological methodology*. San Francisco: Jossey-Bass.

Martindale, C. (1975). *Romantic progression: The psychology of literary history*. Washington, D.C.: Hemisphere.

Martindale, C. (1990). *The clockwork muse: The predictability of artistic change*. New York: Basic Books.

McClelland, P. L., Liang, X., & Barker, V. L. (2010). CEO commitment to the status quo: Replication and extension using content analysis. *Journal of Management, 36*, 1251-1277.

McEnery, T., & Gabrielatos, C. (2006). English corpus linguistics. In B. Aarts & A. McMahon (Eds.), *The handbook of English linguistics* (pp. 33-70). New York: Wiley-Blackwell.

McEnery, T., & Wilson, A. (1996). *Corpus linguistics*. Edinburgh: Edinburgh University Press.

McEnery, T., & Wilson, A. (2001). *Corpus linguistics* (2nd ed.). Edinburgh: Edinburgh University Press.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. London: Routledge.

Mergenthaler, E. (1996). Emotion-abstraction patterns in verbatim protocols: A new way of describing psychotherapeutic processes. *Journal of Consulting and Clinical Psychology, 64*, 1306-1315.

Merton, R. K. (1957). *Social theory and social structure*. Glencoe, IL: Free Press.

Mollin, S. (2009). Combining corpus-linguistics and psychological data on word co-occurrences: Corpus collocates versus word associations. *Corpus Linguistics and Linguistic Theory, 5*(2), 175-200.

Mukherjee, J. (2010). Corpus linguistics versus corpus dogmatism. *International Journal of Corpus Linguistics, 15*(3), 370-378.

Nag, R., Hambrick, D. C., & Chen, M.-J. (2007). What is strategic management, really? Inductive derivation of a consensus definition of the field. *Strategic Management Journal, 28*, 935-955.

Neuendorf, K. A. (2002). *The content analysis guidebook*. Thousand Oaks, CA: Sage.

Ng, W., & De Cock, C. (2002). Battle in the boardroom: A discursive perspective. *Journal of Management Studies, 39*(1), 23-49.

Norrick, N. R. (2004). Hyperbole, extreme case formulation. *Journal of Pragmatics, 36*(9), 1727-1739.

Oakes, M. (1998). *Statistics for corpus linguistics*. Edinburgh: Edinburgh University Press.

Osborne, J. D., Stubbart, C. I., & Ramaprasad, A. (2001). Strategic groups and competitive enactment: A study of dynamic relationships between mental models and performance. *Strategic Management Journal, 22*, 435-454.

Palmer, I., Kabanoff, B., & Dunford, R. (1997). Managerial accounts of downsizing. *Journal of Organizational Behavior, 18*, 623-639.

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology, 54*, 547-577.

Pfarrer, M. D., Pollock, T. G., & Rindova, V. P. (2010). A tale of two assets: The effects of firm reputation and celebrity on earnings surprises and investors' reactions. *Academy of Management Journal, 53*(5), 1131-1152.

Phillips, N., Sewell, G., & Jaynes, S. (2008). Applying critical discourse analysis in strategic management research. *Organizational Research Methods, 11*(4), 770-789.

Pomerantz, A. (1986). Extreme case formulations: A way of legitimizing claims. *Human Studies, 9*(2-3), 219-229.

Popping, R. (2000). *Computer-assisted text analysis*. London: Sage.

Porac, J. F., Wade, J. B., & Pollock, T. G. (1999). Industry categories and the politics of the comparable firm in CEO compensation. *Administrative Science Quarterly, 44*, 112-144.

Precht, K. (2000). *Patterns of stance in English.* Northern Arizona University

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language*. New York: Longman.

Rizomilioti, V. (2006). Exploring epistemic modality in academic discourse using corpora. In E. Arnó Macià, A. Soler Cervera & C. Rueda Ramos (Eds.), *Information technology in languages for specific purposes* (pp. 53-71). New York: Springer.

Roberts, C. W. (1989). Other than counting words: A linguistic approach to content analysis. *Social Forces, 68*(1), 147-177.

Roberts, C. W. (1997). *Text analysis for the social sciences: Methods for drawing statistical inferences from texts and transcripts*. Mahwah: Lawrence Erlbaum.

Schlenker, B. R. (1980). *Impression management: The self-concept, social identity, and interpersonal relations*. Monterey, CA: Brooks/Cole.

Scott, M. (1999). *WordSmith Tools 4*: Oxford University Press.

Short, J. C., Broberg, J. C., Cogliser, C. C., & Brigham, K. H. (2010). Construct validation using computer-aided text analysis (CATA). An illustration using entrepreneurial orientation. *Organizational Research Methods, 13*(2), 320-347.

Short, J. C., & Palmer, T. B. (2008). The application of DICTION to content analysis research in strategic management. *Organizational Research Methods, 11*(4), 727-752.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Sinclair, J. M., Jones, S., Daley, R., & Krishnamurthy, R. (2004). *English collocation studies*. London: Continuum International.

Snell, R. S., & Wong, Y. L. (2007). Differentiating good soldiers from good actors. *Journal of Management Studies, 44*(6), 883-909.

Spell, C. S., & Blum, T. C. (2005). Adoption of workplace substance abuse prevention programs: Strategic choice and institutional perspectives. *Academy of Management Journal, 48*(8), 1125-1142.

Stanton, J. M., & Rogelberg, S. G. (2002). Beyond online surveys: Internet research opportunities for industrial-organizational psychology. In S. G. Rogelberg (Ed.), *Handbook of research methods in industrial and organizational psychology*. Malden: Blackwell.

Staw, B. M., McKechnie, P. I., & Puffer, S. M. (1982). The justification of organizational performance. *Administrative Science Quarterly, 28*, 582-600.

Stone, P. J., et al. (1966). *The General Inquirer: A computer approach to content analysis*. Cambridge: MIT Press.

Stone, P. J., et al. (2000). *The General Inquirer*. http://www.wjh.harvard.edu/~inquirer/.

Stubbs, M. (1996). *Text and corpus analysis. Computer-assisted studies of language and culture*. Oxford: Blackwell.

Stubbs, M. (2001). *Words and phrases: Corpus studies of lexical semantics*. Oxford: Blackwell.

Svartvik, J. (Ed.). (1992). *Directions in corpus linguistics: Proceedings of Nobel Symposium 82 Stockholm 1991*. Berlin: Mouton.

Taylor, J., & Van Every, E. J. (2010). *The situated organization. Case studies in the pragmatics of communication research*. New York: Routledge.

Teubert, W. (2005). My version of corpus linguistics. *International Journal of Corpus Linguistics, 10*(1), 1-13.

Titscher, S., Meyer, M., Wodak, R., & Vetter, E. (2000). *Methods of text and discourse analysis*. London: Sage.

Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. Amsterdam: John Benjamins.

Tsang, E. (2002). Self-serving attributions in corporate annual reports: A replicated study. *Journal of Management Studies, 39*(1), 51-65.

Vaara, E., & Tienari, J. (2008). A discursive perspective on legitimation strategies in multinational corporations. *Academy of Management Review, 33*(4), 985-993.

Wade, J. B., Porac, J. F., & Pollock, T. G. (1997). Worth, words, and the justification of executive pay. *Journal of Organizational Behavior, 18*, 641-664.

Wolfe, R. A., Gephart, R. P., & Johnson, T. E. (1993). Computer-facilitated qualitative data analysis: Potential contributions to management research. *Journal of Management, 19*(3), 637-660.

Wolfe, T. (1990). Using AGENDA for qualitative data analysis. *Journal of Business and Psychology, 5*(2), 261-274.

Wood, L. A., & Kroger, R. O. (2000). *Doing discourse analysis: Methods for studying action in talk and text*. Thousand Oaks: Sage.

Wood, M. (1974). Using key-word-in-context concordance programs for qualitative and quantitative social research. *The Journal of Applied Behavioral Science, 20*(3), 289-297.

WordNet (2010). http://wordnet.princeton.edu/

Xiao, Z., & McEnery, A. (2005). Two approaches to genre analysis: Three genres in modern American English. *Journal of English Linguistics, 33*, 62-82.

Yadav, M. S., Prabhu, J. C., & Chandy, R. K. (2007). Managing the future: CEO attention and innovation outcomes. *Journal of Marketing, 71*, 84-101.