Corpora and Discourse: A Three-Pronged Approach to Analyzing Linguistic Data

Monika Bednarek
University of Sydney

1. Introduction

The three-pronged framework to the analysis of discourse described in this paper was first suggested in Bednarek (2008a, b) but was not developed further there. In this paper I want to outline it in more detail, focusing in particular on those aspects of the framework that involve corpus methodology. In summary, the three-pronged approach involves a. large-scale computerized corpus analysis, b. semi-automated small-scale corpus analysis, and c. manual analysis of individual texts. As such, this is an approach that incorporates macro- (large-scale quantitative analysis), meso- (small-scale quantitative analysis), and micro- (individual text analysis) levels.

There are in fact a number of existing studies in linguistics that involve both corpus and discourse analysis, such as corpus-based discourse analyses (Baker, 2006), recent sociolinguistic research (e.g., Holmes & Schnurr, 2005), studies in critical discourse analysis (Fairclough, 2000; Barker & Galasiński, 2001, p. 26; Mautner, 2008), and Matthiessen's (2006) "two-pronged approach" (Matthiessen, 2006, p. 110). However, Baker (2006) suggests that "while there are a small number of researchers who are already applying corpus methodologies in discourse analysis, this is still a cross-disciplinary field which is somewhat under-subscribed, and appears to be subject to some resistance" (Baker, 2006, p. 6). Thus, the three-pronged approach introduced here is compatible with but extends the few existing studies in the area of corpus-based discourse analysis.

Dörnyei (2007, p. 42-46) gives an overview of researchers in social science using 'mixed methods' approaches involving both quantitative and qualitative research. This "has been endorsed by some of the most influential methodologists in the social sciences" (Dörnyei, 2007, p. 42). As he also notes, "most studies in which some sort of method mixing has taken place have not actually foregrounded the mixed methods approach and hardly any published papers have treated mixed methodology in a principled way" (Dörnyei, 2007, p. 44). Taking up Dörnyei's call, the following sections outline each of the three methodologies incorporated in the three-pronged approach in more detail.

2. Continuum of Discourse Data

With respect to the use of spoken and written text/discourse data, we can identify a continuum in terms of the size of linguistic data, with analyses ranging from the individual analysis of one or just a few texts, to the use of small-scale corpora consisting of a range of texts, to the recourse to large-scale corpora of hundreds of thousands or millions of words (Figure 1).

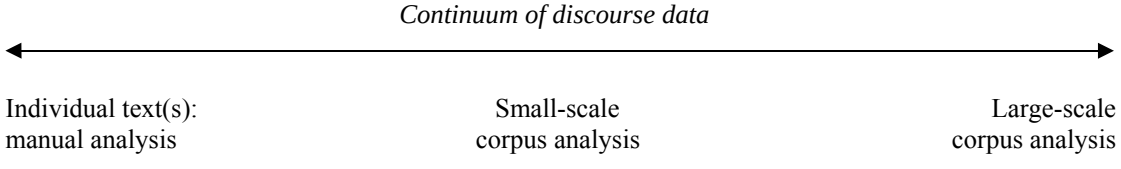

Figure 1. Continuum of discourse data.

The three-fold distinction between large-scale corpus analysis, small-scale corpus analysis, and manual analysis of individual texts is thus a simplification to a certain extent, and degrees of gradience exist between the approaches.

3. A Three-Pronged Approach

3.1. Large-Scale Corpus Analysis

Large-scale corpus linguistics is an approach to the linguistic analysis of data that uses large computerized collections of text (corpora) and appropriate software to analyze them. The material that is contained in corpora is usually said to be more or less representative of the variety of language for which it was designed, and there are many different types of corpora (see Hunston, 2002; Meyer, 2002 on corpora and corpus linguistics). Depending on the use made of corpora and corpus software, researchers can find out many different things, for instance (a) the frequency with which every word in the corpus occurs, (b) words that are unusually (in)frequent when compared with a reference corpus, (c) all occurrences of a particular word, (d) recurring larger structures (n-grams, clusters, phrases), (e) grammatical frames, (f) collocations, (g) occurrences of parts of speech and their combinations, et cetera. Corpus linguistic applications include discourse analysis, lexicography, stylistics, forensic linguistics, language variation studies, and language teaching (Baker, 2006, p. 2-3).

The epistemological advantage of this methodology (large-scale corpus linguistics) is that the data are highly representative, and that it involves the use of empirical, systematic, and sophisticated evidence. (Nevertheless, the interpretation of this empirical evidence can be subjective, depending on the focus of analysis.) Corpus searches and studies are often reliable and replicable, especially when they concentrate on formally defined items. Another advantage is that such studies can uncover features of language that are inaccessible to intuition or that cannot be discovered through the analysis of one or a few texts. This concerns patterning, typicality of usage, and quantification (e.g., type-token ratios, frequency lists, lexical density) and particular kinds of semantic-pragmatic meanings such as semantic prosody (Louw, 1993; Bednarek, 2008c). Corpus data also give researchers access to shared connotations (Coffin & O'Halloran, 2005) and to the experience of language users (Thompson & Hunston, 2000, p. 15). Because the focus of large-scale corpus linguistics is on large amounts of actually occurring discourse, it allows the study of typicality, quantitative norm, and across-text, or *inter*textual, patterning. User-friendly software enables the easy application of tests of statistical significance. Thus, hypotheses can be verified, falsified, or modified, and new language features can be uncovered with the help of large scale corpora.

However, there are also some drawbacks to the analysis of large amounts of data. For example, on account of the size of the data, researchers are able to execute searches only for formally defined items. If the corpus has previously been annotated, additional searches (e.g., for semantic meanings) become possible, but such annotation can be time-consuming, in particular if it is done manually or semi-automatically. Another point of critique is that corpus linguistic approaches often treat social categories such as gender or age as stable variables rather than as discursively construed (Butler, 1999). With respect to discourse analysis, it can further be argued that many (though not all) corpus analytic studies do not take into account reception, the dynamic nature of discourse, or its context or textual structure. Beyond its consideration of syntagms, large-scale corpus linguistics usually has less to say about context, and the unfolding of meaning in texts (e.g., *intra*textual patterning). As Stubbs (2008, p. 5) notes with respect to corpus linguistic keyword analysis, "since the texts have been ripped apart into lists of individual words and/or n-grams, the patterns ignore text segmentation. They are a feature of global textual cohesion, but not textual structure."

With respect to semiotic systems other than language, it must also be pointed out that the majority of existing corpora are monomodal (i.e., they include only linguistic material). With respect to multimodality and corpus linguistics, both Baldry and Thibault (2006) and Carter and Adolphs (2008) have recently emphasized the importance of using multimodal corpora in semiotic analysis while noting that their design is "very much in its infancy" (Bauldry & Thibault, 2006, p. 181). Also, Johnston (2008) has persuasively argued for the need for corpora in sign language research.

Another issue concerns the focus of large-scale corpus linguistics on repeated occurrence and its concurrent interest in the typical and the frequent, rather than the individual and outstanding (but see,

e.g., Hoey, 2005 on creativity). This emphasis neglects the importance of outstanding singular texts that might have an impact on phylogenesis and ontogenesis that goes beyond the impact of 'ordinary' singular texts that achieve their impact through repetition. For example, some texts are "highly valued in the community or […] [have] special significance in some domain such as history or politics. Here we treat texts as artefacts – objects of study in their own right" (Matthiessen, 2006, p. 108). A relatively recent example is the Al Gore produced and presented documentary *An Inconvenient Truth* which has been widely credited with causing a seismic shift in the public awareness of environmental sustainability. Indeed, its impact was recognized in the awarding of the Nobel Peace Prize to Gore in 2007 for "efforts to build up and disseminate greater knowledge about man-made climate change, and to lay the foundations for the measures that are needed to counteract such change." (http://nobelprize.org/nobel_prizes/peace/laureates/2007/). The impact of such extraordinary texts (politicians' speeches are another prime example), which are arguably more 'powerful' than normal texts produced countless times by speakers over time, presumably depends on diverse factors such as the profile of the author/speaker, the context of reception, and, in the case of *An Inconvenient Truth*, the sophisticated use of semiotic resources (Bednarek, Caple, & Hood, 2008). Studying such unique texts, then, also provides important insights, and is an argument for qualitative analyses of individual texts. With respect to large-scale corpus linguistics it needs to be acknowledged that powerful unique texts become 'lost' in the corpus where they have only as much weight as any other text:

> [F]requent patterns of language do not always necessarily imply underlying hegemonic discourses. Or rather, the 'power' of individual texts or speakers may not be evenly distributed. A corpus which contains a single (unrepresentative) speech by the leader of a country or religious group, newspaper editor or CEO may carry more weight discursively than hundreds of similar texts which were produced by 'ordinary' people. (Baker, 2006, p. 19)

Other 'special' texts are those that are perhaps not very 'sophisticated', but are considerably more read than others, and thus may have a greater impact on phylogenesis and ontogenesis: "According to claims, the most likely document that an ordinary English citizen will cast his or her eyes over is *The Sun* newspaper" (Sinclair, 2005, p. 7). In fact, many texts in existing corpora such as the British National Corpus differ widely in circulation status (size of readership/circulation level) (Lee, 2001, p. 68). This has implications for corpus design: For instance, Sinclair (2005) asks if we should include more texts from *The Sun* than from other newspapers in a corpus of British English on account of its importance. Sinclair concludes that issues of representativeness and balance in corpus design are far from resolved at present, and recommends that corpus design "should be documented fully with information about the contents and arguments in justification of the decisions taken" (Sinclair, 2005, p. 8). Other problematic issues in the design of corpora include the thorny notion of representativeness (Mahlberg, 2004), procedures of sampling (Mautner, 2008), and practicalities of coding and inputting as well as technological exigencies (Wynne, 2005). Depending on the discourse phenomenon in which the researcher is interested, it may also be difficult to find a corpus that has many naturally occurring instances of this particular phenomenon, and not all discourse phenomena can be formally defined. This necessitates the use of small-scale corpora at least in a pilot study or as complementary data. Section 3.2 describes small-scale corpus analysis in more detail.

3.2. Small-Scale Corpus Analysis

O'Donnell (2007) distinguishes between two main types of corpus studies: (1) automatic studies using computer software (as outlined in Section 3.1), and (2) computer assisted manual annotation (CAMA), "where a human annotates the text in terms of patterns that generally computers cannot recognize." The latter is often the case when smaller corpora are analyzed, as such annotation can take considerable time and effort. For this kind of analysis, we can use the term *small-scale corpus linguistics*. More specifically, by this methodology (small-scale corpus analysis), we mean the manual analysis of small-scale corpora which is ideally (but need not be) computer-assisted, and which makes use of quantitative and qualitative analysis. The corpora that are used in such studies should be small

enough to allow manual, context-sensitive analysis (including the annotation of semantic or pragmatic meaning), but large enough to show at least some patterns, exhibit a certain degree of representativeness, and enable some generalizibility. The specific size of the respective corpus depends on the nature of the investigated linguistic features; that is, the corpus needs to be bigger if lexis is the focus than if grammar is the focus. Analysis, coding, and annotation can be supported through computer software such as the UAM CorpusTool (http://www.wagsoft.com/ CorpusTool/; O'Donnell, 2007) and others.

Examples of existing small-scale corpus studies are those conducted by Semino and Short (2004), who analyze reported speech and thought in 'small' corpora of 80,000 words each of prose fiction, news reports, and (auto)biography; Bednarek (2006b), who studies the distribution of evaluative meaning in a 70,000 word corpus of hard news stories; Martin and White (2005, p. 165), who analyze appraisal in 85 items from journalistic discourse from a systemic functional linguistic perspective; and Bednarek (2008a), who explores the distribution of emotion terms in an 85,000-word corpus comprising conversation, news reportage, fiction, and academic discourse. Generally speaking, conversation analyses have discovered regularities of turn-taking and sequencing using collections of spoken data. (In addition, there may be other linguistic studies of 'small' corpora or collections of text, particularly in pragmatics/discourse analysis.)

In analogy to a common distinction made in large-scale corpus research, a distinction can be made between text-based and text-driven studies in small-scale corpus research (Bednarek, 2006a): With respect to large-scale corpus linguistics, Tognini-Bonelli (2001) distinguishes *corpus-driven* from *corpus-based* linguistics. She employs the term *corpus-based* "to refer to a methodology that avails itself of the corpus mainly to expound, test or exemplify theories and descriptions that were formulated before large corpora became available to inform language study" (Tognini-Bonelli, 2001, p. 65), and the term *corpus-driven* to refer to an approach where "the commitment of the linguist is to the integrity of the data as a whole, and descriptions aim to be comprehensive with respect to corpus evidence" (2001, p. 84). Similarly, we can employ the term *text-driven* to refer to a methodology where texts are analyzed without many (or indeed any) a priori theoretical assumptions, and the term *text-based* to refer to an approach where texts are analyzed using a previously established theory. In contrast to corpus-based/corpus-driven research, text-based/text-driven research is based on the manual analysis of small-scale text corpora rather than on automated large-scale corpus analyses. However, what Halliday (2004) notes with respect to corpus-based versus corpus-driven linguistics is also true for text-based and text-driven linguistics: The two methodologies are located on a continuum rather than representing a strict dichotomy, and many linguistic analyses are situated somewhere in between.

While analyses of small-scale corpora are more representative than analyses of just one or a few texts, they can be just as subjective depending on the features analyzed and the method of analysis. In addition to the advantages of manual analysis, allowing to take into account, for example, pragmatic and cotextual meaning, they offer an observation of recurring regularities, practices or patterns, at least to some extent.

3.3. Individual Text Analysis

The third methodological approach to discourse consists of the manual analysis of one or a few individual texts. Because the focus of this paper is on corpus linguistics not much will be said on this third 'prong'. In fact, analyses of individual texts are a prevalent methodology in critical discourse analysis, systemic functional linguistics, and other types of discourse analysis. Because it deals with small data (individual texts) the analysis is not very representative. However, this limitation in data size enables the analyst to take into account sociocultural context and interaction, cotext and context, and coarticulated meanings. Such analysis allows researchers to focus on the localized construal of social phenomena such as identity in particular contexts, often resulting in a complex, rich, interpretive, dynamic, and flexible analysis of microcontexts. and capturing the dynamic and negotiatory nature of much language use. If video data are used (Heath, 2004) the multimodal context can be taken into account and attention can be paid not only to intonation and prosody but also to gesture, posture, body movement, and facial expression. However, while this type of analysis enables us to study significant

or important texts (compare Section 3.1 above) and events, it 'loses out' on generalizability, replicability, and representativeness. To what extent are statements made about one particular text transferable to other texts and contexts? Such analysis is also frequently very time-consuming.

4. Conclusion: Triangulating Discourse Analysis

In view of the above observations it seems that each of the three 'prongs' provides certain insights and perspectives, without necessarily being able to capture the complexity of discourse in its entirety. It thus seems reasonable to argue for a 'mixed-methods' approach (Dörnyei, 2007) in analyzing discourse, combining quantitative and qualitative research. Triangulation, or the use of a variety of methods, allows researchers "to cross-check and verify the reliability of a particular research tool and the validity of data collected" (McNeill & Chapman, 2005, p. 23). Researchers who similarly argue for a combination of methods include Holmes (1997), Taylor (2001), Holmes and Schnurr (2005), and Wodak and Krzyżanowski (2008). Baker also concludes that triangulation is the best way to go, as "it facilitates validity checks of hypotheses, it anchors findings in more robust interpretations and explanations, and it allows researchers to respond flexibly to unforeseen problems and aspects of their research" (Baker, 2006, p. 16, citing Layder, 1993, p. 128). Dörnyei (2007, p. 45-46) notes the following strengths of mixed methods research:

- Increasing strengths and eliminating weakness of qualitative/quantitative research
- Providing an analysis of complex issues at multiple levels
- Improving the validity of research findings
- Accessible to diverse audiences

Thus, it is hoped that the three-pronged approach to discourse analysis introduced in this paper will be useful to researchers from across linguistic subdisciplines, and that its application will provide further illumination of the complex phenomenon that is language.

References

Baker, Paul. (2006). *Using corpora in discourse analysis.* London/New York: Continuum.

Baldry, Anthony, & Paul Thibault. (2006). Multimodal corpus linguistics. In Geoff Thompson & Susan Hunston (Eds.), *System and corpus. Exploring connections* (pp. 164-183). London: Equinox.

Barker, Chris, & Dariusz Galasińksi. (2001). *Cultural studies and discourse analysis: A dialogue on language and identity.* Thousand Oaks: Sage.

Bednarek, Monika. (2006a). Epistemological positioning and evidentiality in English news discourse: A text-driven approach. *Text and Talk, 26,* 635-660.

Bednarek, Monika. (2006b). *Evaluation in media discourse.* London/New York: Continuum.

Bednarek, Monika. (2008a). *Emotion talk across corpora.* Basingstoke/New York: Palgrave Macmillan.

Bednarek, Monika. (2008b). "What the hell is wrong with you?" A corpus perspective on evaluation and emotion in contemporary American pop culture. In Ahmar Mahboob & Naomi Knight (Eds.), *Questioning linguistics* (pp. 95-126). Newcastle: Cambridge Scholars Press.

Bednarek, Monika. (2008c). Semantic preference and semantic prosody re-examined. *Corpus Linguistics and Linguistic Theory, 4,* 119-139.

Bednarek, Monika, Helen Caple, & Sue Hood. (2008). *Communicating a sustainable Australia: Discourses of persuasion.* Unpublished manuscript, University of Technology, Sydney.

Butler, Judith. (1999) *Gender trouble: Feminism and the subversion of identity.* New York/New York: Routledge.

Carter, Ronald, & Svenja Adolphs. (2008). Linking the verbal and visual: New direction for corpus linguistics. In Andrea Gerbig & Oliver Mason (Eds.), *Language, people, numbers. Corpus linguistics and society* (pp. 275-291). Amsterdam/New York: Rodopi.

Coffin, Caroline, & Kieran O'Halloran. (2005). Finding the global groove: Theorising and analysing dynamic reader positioning using appraisal, corpus, and a concordancer. *Critical Discourse Studies, 2,* 143-163.

Dörnyei, Zoltán. (2007). *Research methods in applied linguistics.* Oxford: Oxford University Press.

Fairclough, Norman. (2000). *New labour, new language*? London: Routledge.

Halliday, Michael. (2004). The spoken language corpus: A foundation for grammatical theory. In Karin Aijmer & Bengt Altenberg (Eds.), *Advances in corpus linguistics. Papers from the 23rd International Conference on English Language Research on Computerized Corpora* (pp. 11-39). Amsterdam/New York: Rodopi.

Heath, Christian. (2004). Analysing face-to-face interaction: Video, the visual and material. In David Silverman (Ed.), *Qualitative research: Theory, method and practice* (pp. 266-282). London: Sage.

Hoey, Michael. (2005). *Lexical priming: A new theory of words and language.* London/New York: Routledge.

Holmes, Janet. (1997). Women, language and identity. *Journal of Sociolinguistics, 1,* 195-223.

Holmes, Janet, & Stephanie Schnurr. (2005). Politeness, humor and gender in the workplace: Negotiating norms and identifying contestation. *Journal of Politeness Research, 1,* 121-149.

Hunston, Susan. (2002). *Corpora in applied linguistics.* Cambridge: Cambridge University Press.

Johnston, Trevor. (2008). *Signed Languages and Corpus Linguistics*. Paper presented at the 2nd Free Linguistics conference, University of Sydney, Australia.

Lee, David. (2001). Genres, registers, text types, domains, and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning & Technology, 5,* 37-72.

Louw, Bill. (1993). Irony in the text or insincerity in the writer? In Mona Baker, Gill Francis, & Elena Tognini-Bonelli (Eds.), *Text and technology. In honour of John Sinclair* (pp. 157-176). Amsterdam/Philadelphia: John Benjamins.

Mahlberg, Michaela. (2004). Corpus design and the words in a dictionary. *Lexicographica, 20,* 114-29.

Martin, Jim R., & Peter R. R. White. (2005). *The language of evaluation: Appraisal in English.* Basingstoke/New York: Palgrave Macmillan.

Matthiessen, Christian M. I. M. (2006). Frequency profiles of some basic grammatical systems: An interim report. In Susan Hunston & Geoff Thompson (Eds.), *System and corpus: Exploring connections* (pp. 103-142). London: Equinox.

Mautner, Gerlinde. (2008). Analyzing newspapers, magazines and other print media. In Ruth Wodak & Michal Krzyżanowski (Eds.), *Qualitative discourse analysis for the social sciences* (pp. 30-53). Basingstoke/New York: Palgrave Macmillan.

McNeill, Patrick, & Steve Chapman. (2005). *Research methods* (3rd ed.). London: Routledge.

Meyer, Charles F. (2002). *English corpus linguistics: An introduction.* Cambridge: Cambridge University Press.

O'Donnell, Mick. (2007). *The UAM Corpus Tool*. Paper presented at the 28th ICAME conference; Stratford-upon-Avon, UK.

Semino, Elena, & Mick Short. (2004). *Corpus stylistics: A corpus-based study of speech, thought and writing in a corpus of English writing.* London: Routledge.

Sinclair, John. (2005). Corpus and text: Basic principles. In Martin Wynne (Ed.), *Developing linguistic corpora: A guide to good practice* (pp. 1-16). Oxford: Oxbow.

Stubbs, Michael. (2008). *Three concepts of keywords.* Revised version of paper presented at Keyness in Text, Certosa di Pontignano, University of Siena, 2007, available at http://www.uni-trier.de/fileadmin/fb2/ANG/Linguistik/Stubbs/stubbs-2008-keywords.pdf.

Taylor, Stephanie. (2001). Evaluating and applying discourse analytic research. In Margaret Wetherell et al., (Eds.), *Discourse as data: A guide for analysis* (pp. 311-330). London: Sage in association with The Open University.

Thompson, Geoff, & Susan Hunston. (2000). Evaluation: An introduction. In Susan Hunston & Geoff Thompson (Eds.), *Evaluation in text: Authorial stance and the construction of discourse* (pp. 1-27). Oxford: Oxford University Press.

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

Wodak, Ruth, & Michal Krzyżanowski (Eds.). (2008). *Qualitative discourse analysis for the social sciences.* Basingstoke/New York: Palgrave Macmillan.

Wynne, Martin (Ed.). (2005). *Developing linguistic corpora: A guide to good practice.* Oxford: Oxbow.

Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus: Mustering Languages

edited by Michael Haugh, Kate Burridge, Jean Mulder, and Pam Peters

Cascadilla Proceedings Project Somerville, MA 2009

Copyright information

ISBN 978-1-57473-435-5 library binding

Ordering information

Orders for the library binding edition are handled by Cascadilla Press.
To place an order, go to www.lingref.com or contact:

Cascadilla Press, P.O. Box 440355, Somerville, MA 02144, USA
phone: 1-617-776-2370, fax: 1-617-776-2271, e-mail: sales@cascadilla.com

Web access and citation information

This entire proceedings can also be viewed on the web at www.lingref.com. Each paper has a unique document # which can be added to citations to facilitate access. The document # should not replace the full citation.

This paper can be cited as:

Bednarek, Monika. 2009. Corpora and Discourse: A Three-Pronged Approach to Analyzing Linguistic Data. In *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus*, ed. Michael Haugh et al., 19-24. Somerville, MA: Cascadilla Proceedings Project.

or:

Bednarek, Monika. 2009. Corpora and Discourse: A Three-Pronged Approach to Analyzing Linguistic Data. In *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus*, ed. Michael Haugh et al., 19-24. Somerville, MA: Cascadilla Proceedings Project. www.lingref.com, document #2283.