Introduction to the special issue: Exploring corpus-informed approaches to writing research

Stephanie A. Schlitz

Bloomsburg University of Pennsylvania | USA

1. Introduction

Biber et al. differentiate two main approaches to language study: structural and usage-based (2004, p. 1). The structural approach, the foreground of linguistic study during much of the early to mid- twentieth century, seeks to identify and to describe the structural units (e.g. morphemes, words, phrases) employed in language and to explain the relationships between these units as they combine to form larger units of meaning. The usage- or corpus-based approach, the approach adopted by the articles included in this special issue, seeks to describe and to explain language using the "actual language used in naturally occurring texts" (p. 1). In the context of language research, a corpus is defined as a 'body' or collection of language texts. The collection may be small, consisting, for instance, of a set of classroom essays compiled by an individual compiler-analyst (e.g. Bloch) who aims to use the corpus for instructional application; it may be designed as a large reference resource consisting of millions of words (e.g. Deane and Quinlan; Sharpling); or it may be developed as a parallel corpus designed to correspond with an existing corpus resource and thus to enable comparative study. While corpora in general may comprise written, spoken, or other types of texts, the studies in this issue utilize corpora compiled exclusively from written sources or from written source data.

journal of WRITING RESEARCH

Schlitz, S.A. (2010). Introduction to special issue: Exploring corpus-informed approaches to writing research. *Journal of Writing Research, 2* (2), 91- 98. http://dx.doi.org/10.17239/jowr-2010.02.02.1
Contact and copyright: Earli | Stephanie A. Schlitz, Bloomsburg University of Pennsylvania, Department of English, 117B Bakeless Hall, Bloomsburg, PA 17815 | USA - sschlitz@bloomu.edu.

Regardless of size, purpose, or text type, a valid research corpus conforms to a set of common design criteria (Biber, 2004; Sinclair, 2005). A corpus should be *representative*; that is, it should comprise texts that represent a particular variety of language. It should be *machine-readable*; that is, compiled and stored on a computer so that it is accessible and analyzable in digital format. And while traditionally a corpus is *finite*, containing a fixed collection of texts that serve as a static reference compilation, some corpora (e.g. those used for lexicographical purposes) are open-ended, regularly adding texts to increase the total sample size (McEnery & Wilson, 2004, p. 30).

To examine the data contained in corpora, researchers utilize different types of text analysis software. The authors in this special issue employ freely available as well as proprietary software, including the freeware program AntConc (used by Henderson & Barr; Römer & Wulff) and the license-based programs MonoConc Pro (used by Bloch) and WordSmith Tools (used by Hüttner). Although these programs vary in functionality and performance, they share the essential properties of facilitating quantitative analysis and enabling analysts quickly and easily to collect, sort, and manipulate the linguistic data contained in a corpus. This is accomplished through the generation of, for instance, *keywords* (words with a high frequency when compared to words with average rates of occurrence); *frequency* lists (lists of all of the words in a corpus organized by frequency of occurrence or organized alphabetically with frequency information included); a *concordance* (a list of a particular search word or phrase in all of its contexts); *collocates* (words that co-occur with a search word or phrase); and *clusters/n-grams* (characteristic word sequences).

Analysis of these kinds of data, as Hüttner notes, aids researchers in identifying "typical patterns of language usage" that "frequently escape intuitions of native speakers and of teachers" (p. 200). Evidence from the studies in this issue indicates that such patterns are particularly important in understanding "the development and acquisition" of writing (Parr, p. 130) and in designing resources for writing instruction and assessment.

2. Toward Corpus-Informed Approaches to Writing Research

Corpus-based approaches have dominated linguistic study since the latter half of the twentieth century, and, today, leveraging language corpora and corpus-based methods to describe and to analyze written and spoken language is an established tradition within this broad, interdisciplinary field. Teachers and researchers from a variety of second language (L2) and related disciplines, for example, including English as a Second Language, English Language Teaching and its subfield English for Academic Purposes, and foreign language instruction, have been using corpora and corpus methods to inform language study and language instruction. And they've been applying this work to enhance L2 writing pedagogy in a range of areas, including vocabulary knowledge (e.g. Nation, 2001), genre knowledge (e.g. Henry & Roseberry, 2001;

Tribble, 2001; Tribble, 2002), grammatical knowledge (e.g. Diniz & Moran, 2005; Clear, 2000), and citation practices (e.g. Thompson & Tribble, 2001).

Writing researchers as well have begun extending corpus methods to both native language (L1) and second language writing research. Research teams in the U.K. and the U.S., for example, have designed large reference corpora of student writing. The developers of the British Academic Written English (BAWE) Corpus, "which contains 2761 pieces of proficient assessed student writing" ("BAWE and BAWE Plus Collections") and which is utilized by two of the papers published in this special issue (Sharpling; Henderson and Barr), suggest that the corpus "has the potential to chart growth patterns such as whether students' arguments became more complex as their education advanced, whether students learned to integrate material from different sources in formulating conclusions, and whether students' vocabulary became more specialized and precise" (Nesi et al., 2004, p. 446). The Michigan Corpus of Upper-Level Student Papers, which is discussed in Römer and Wulff's contribution, provides a corpus of *circa* 2.6 million words and offers researchers the opportunity to quantitatively and qualitatively examine student writing in areas as diverse as writing development, genre variation, and disciplinary differences ("MICUSP").

The trend toward corpus-informed approaches to writing research also continues on a smaller scale. Given the ease with which individual teachers and researchers can create and then mine corpora using text analysis software, the development of small corpora by writing teachers who adopt the role of compiler-analyst (Flowerdew, 2005) is providing another avenue for corpus application.

Yet, because corpus approaches introduce new methods and new theoretical models within the field of writing research, L1 writing research in particular, to date, relatively few writing studies have integrated corpora and corpus methods, and there have been very few comprehensive discussions of the work undertaken this area. As Römer and Wulff point out, "In spite of the growing recognition of the usefulness of corpus linguistics for professional communication research in general and writing research in particular, it is hard to find a basic introduction to corpus linguistic methods tailored to the needs of writing researchers" (p. 101).

Although the seven articles in this special issue represent contributions from a diverse, international group of authors specializing in a broad spectrum of writing-related disciplines and addressing a wide array of research interests, they are unified by a single, common theme: the use of corpora in writing research. The papers examine writing of both native language and L2 writers and demonstrate the use of corpora to investigate the writing of students in primary through university levels. They explore topics such as writing development, writing assessment, and writing instruction and employ methodologies ranging from genre-based to comparative approaches. Taken together, the articles in this collection have been selected to provide readers with a comprehensive and informative introduction to the field and to exemplify how researchers are developing and exploiting corpora and corpus methods to improve writing research and writing instruction.

3. Introduction to the articles in this Special Issue

The seven articles brought together in this special issue represent the range of application for corpus-informed writing research. Although a number of the papers may overlap in topic and/or approach, for convenience, the issue and this introduction have been organized into four main themes: designing and analyzing large reference corpora; large corpora and writing assessment; genre-based approaches; and small corpora and L2 college student writing.

3.1 Designing and analyzing large reference corpora

It is fitting to begin with the article authored by Römer and Wulff, *Applying corpus methods to written academic texts: Explorations of MICUSP*. While this article does offer a brief, introductory overview of the recently released Michigan Corpus of Upper-level Student Papers (MICUSP), the authors' chief aim is "to acquaint readers who may not be familiar with corpus work with the core techniques in corpus analysis" and "to demonstrate the potential of corpus-analytic techniques for the field of writing research at large, be it as a primary method of investigation, or a supplementary method to test, complete, and qualify given assumptions" (p. 101).

The article will be especially valuable to readers who are new to corpus-analytic techniques, as section three employs MICUSP in an extended tutorial detailing the "Central steps in corpus analysis," and section four provides an illustrative case study, "Attended and unattended *this* in student writing," which is designed to offer readers an exemplar which not only explicates corpus-analytic techniques but which also illustrates how researchers can exploit a corpus such as MICUSP to explore a writing research question. The tutorial guides readers, step-by-step, through fundamental corpus-analytic methods using the freeware program AntConc. Instruction begins with how to open the program and call up a file and continues through an explanation of how to generate collocates and clusters. The case study builds on the tutorial and offers critical discussion to assist readers in understanding how and why to refine search and analysis methods.

In *A dual purpose data base for research and diagnostic assessment of student writing*, Parr introduces another dimension of corpus-informed writing research, the development of a *mediated* corpus consisting not of writing but of students' scores on a writing test. The study involves a large, national sample of writing produced by New Zealand students in grade levels 4-12. Students were asked to respond to one of seven purpose-based writing prompts; these essays were scored by trained evaluators who used one of seven rubrics designed to match the seven different writing prompts. The assessment results were entered into a corpus, and, as explained by Parr, these data were "interrogated" to analyze "patterns of development with age;" "performance […] in relation to curriculum expectations;" "relative performance in each of the different purposes for writing through the course of schooling;" and "differential performance by gender or by ethnicity" (p. 135-136). Parr's contribution presents readers with a novel approach as well as an in-depth discussion of the analyses it has enabled.

3.2 Large corpora and writing assessment

Deane and Quinlan's paper, *What automated analyses of corpora can tell us about students' writing skills*, represents a shift in focus from the use of human-based assessment scores (as discussed by Parr) to the current and potential use of automated essay scoring (AES) systems, such as Educational Testing Service's e-rater. E-rater was "developed by analyzing large corpora of student essays, first to identify useful features and then to build scoring models in which human ratings of essay quality are used as an external criterion" (p. 152). The AES system depends on natural language processing techniques to "capture machine-detectable features germane to writing quality" (p. 152) and can be used in "Predicting Human Judgments of Essay Quality" (p. 154), in "Predicting the Developmental Level of Student Writing" (p. 159), and in "Identifying Dimensions of Linguistic Variation in Student Essays" (p. 160).

For example, citing Spandel and Stiggins (1990), Deane and Quinlan argue that although educators generally agree on the features that define quality writing, measuring student achievement of these quality standards can be difficult in part because of because of inter-rater reliability problems. While Deane and Quinlan acknowledge that e-rater cannot replicate a human reader's ability to identify and evaluate writing features such as "voice," "ideas," or "quality of argumentation," it can be trained to measure the presence or absence of linguistic features that correlate positively with quality writing, including, for example, "fluency, word choice, adherence to conventions, and use of appropriate discourse structures" (p. 155). Measurement of these machine-detectable features, to the extent that they do correlate with the human-detectable features, can yield consistent and precise large-scale assessment.

In their discussion of the potential use of AES, the authors describe new work designed to investigate the writing of students engaged in different stages of the writing process, including prewriting, drafting and revising, and suggest the possibility of teaming assessment data with writing process data to provide greater insight into student writing.

Sharpling's contribution, *When BAWE meets WELT: The use of a corpus of student writing to develop items for a proficiency test in grammar and English usage*, explicates the use of the British Academic Written English corpus in the design and development of a new version of the Warwick English Language Test (WELT), "an English language proficiency test for candidates across the world applying to be accepted for further study by Higher Education (HE) institutions within the United Kingdom" (p. 180). As Sharpling suggests, the key contribution of this paper is in demonstrating how student writing can be used in the development of an assessment tool that reflects actual – rather than artificial – written usage.

3.3 Genre-based approaches

The potential of purpose-built corpora in the analysis of student academic writing in English by Hüttner points out that although students in non-English speaking countries

are increasingly expected to write academic papers in English, the scholarly community's understanding of these students' writing proficiency and writing practices is insufficient. Hüttner elucidates the need to increase English as a Foreign Language (EFL) resources, specifically corpora of non-native student writing, and she argues that English for Academic Purposes (EAP) research in particular demands new, corpus-driven methods. Through the expansion and analysis of EFL writing corpora, Hüttner proposes to cultivate "theory-informed" teaching practices and advocates an "extended genre analysis" model which "take[s] into account the special status of student genres" and "systematically integrate[s] corpus linguistic tools into the analysis" (p. 199).

The study described in her paper applies extended genre analysis "to a corpus of 55 student paper conclusions produced by non-native speakers in the initial phase of their studies" (p. 199) and compares it with an expert corpus of 55 articles in order to illustrate how EAP pedagogy can benefit from such an approach. Notably, this paper discusses the study's "Implications for teaching practice" (p. 215) and describes how instructors can use corpora to generate models of writing and how students can excavate corpora for "discovery-learning" purposes.

3.4 Small corpora and L2 college student writing

In *A concordance-based study of the use of reporting verbs as rhetorical devices in academic papers*, Bloch takes the position that "in order to become successful academic writers" college student writers must "understand how [their] grammatical choices...can affect their credibility as researchers" and "enhance the rhetorical impact of [a] claim" (p. 220). According to Bloch, non-native English writers in particular can have difficulty using reporting verbs fluently. This paper describes the creation of two small corpora from the journal *Science* which are used together with a third small corpus of English as a Second Language (ESL) college student papers to compare students' use of reporting verbs with that of published authors and to develop a database of authentic usage for use in an ESL college composition environment. Bloch details how the corpora facilitate study of academic writers' use of reporting verbs to create a repository of example sentences for use in developing authentic teaching materials and how they facilitate data-driven learning opportunities in the composition classroom.

The pilot study shared in Henderson and Barr's paper, *Comparing indicators of authorial stance in psychology students' writing and published research articles*, is similar to Bloch's study in that it compares a learner corpus of (French psychology) student papers written in English with a corpus of native English students' psychology papers and a corpus of published psychology articles. Although Bloch focuses on reporting verbs and research stance and Henderson and Barr on pronouns, adjectives, and adverbs and authorial stance, both call attention to the fact that L2 students' fluency in their target language, in this instance English, is determined in part by their ability to use certain lexical bundles with a degree of proficiency that matches or nearly matches that of L1 writers. If L2 writers are to become members of the target discourse

community, they need greater awareness of the language patterns in their target language.

Using the introductory sections of papers, which Henderson and Barr suggest fulfill the discursive function of establishing authorial stance, Henderson and Barr evaluate French ESL students' use of pronouns, adjectives, and adverbs beside that of L1 students and published authors. Notably, Henderson and Barr discern patterns in the writing of ESL students that were not anticipated yet do not convey "absence of an idea or function" (p. 261). This finding, they observe, has implications for learner corpora analysis methodology and demands further exploration.

4. Closing Remarks

Finally, while this collection does aim to offer an introduction to corpus-informed writing research, it does not endeavor to provide an all-encompassing perspective. All of the papers in this collection, because they employ corpora, advance corpus-informed approaches to writing research. Most notably, all employ usage-based theoretical models. A more comprehensive overview would, conceivably, include discussion of the challenges and objections to corpus methods as well as usage-based approaches. It is acknowledged as well that while this special issue provides an introduction, there is, indeed, considerably more to cover. These shortcomings notwithstanding, it is hoped that the strength of the issue will lie in its breadth and scope and in its potential to illuminate the connections between corpus methods and writing research.

References

BAWE and BAWE Plus Collections. Warwick Centre for Applied Linguistics. 21 April 2010. <http://www2.warwick.ac.uk/fac/soc/al/research/collect/bawe/>.

Biber, D., Conrad, S., & Reppen, R. (2004). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Clear, J. (2000). Do you believe in grammar? In L. Burnard & T. McEnery (Eds.), *Rethinking Language Pedagogy from a Corpus Perspective* (pp. 19-30). Frankfurt: Peter Lang.

Diniz, L. & Moran, K. (2005). Corpus-Based Tools for Efficient Writing Instruction. *Essential Teacher*, *2*(3), 36-39.

Flowerdew, L. (2005). An integration of corpus-based and genre-based approaches to text analysis in EAP/ESP: countering criticisms against corpus-based methodologies. *English for Specific Purposes*, *24*, 321-332. doi: 10.1016/j.esp.2004.09.002

Henry, A. & Roseberry, R. L. (2001). Using a small corpus to obtain data for teaching a genre. In M. Ghadessy, A. Henry, & R. L. Roseberry (Eds.), *Small Corpus Studies and ELT: Theory and Practice* (pp. 93-134). Philadelphia: John Benjamins. doi: 10.1075/scl.5.10hen

MICUSP. University of Michigan English Language Institute. 21 Jan. 2010. <http://micusp.elicorpora.info/>.

McEnery, T., & Wilson, A. (2004). *Corpus Linguistics: An Introduction*. 2nd edition. Edinburgh: Edinburgh University Press.

Nation, P. (2001). Using Two Small Corpora to Investigate Learner Needs: Two Vocabulary Researcher Tools. In M. Ghadessy, A. Henry, & R. L. Roseberry (Eds.), *Small Corpus Studies*

and ELT: Theory and Practice (pp. 31-46). Philadelphia: John Benjamins. doi: 10.1075/scl.5.06nat

Nesi, H., Sharpling, G., & Ganobcsik-Williams, L. (2004). Student papers across the curriculum: Designing and developing a corpus of British student writing. *Computers and Composition*, *21*, 439-450. doi: 10.1016/j.compcom.2004.08.003

Sinclair, J. (2005). Corpus and Text - Basic Principles. In M. Wynne (Ed.), *Developing Linguistic Corpora: a Guide to Good Practice* (pp. 1-16). Oxford: Oxbow Books. Available online from http://ahds.ac.uk/linguistic-corpora/ [Accessed 2009-11-13].

Spandel, V., & Stiggins, R. J. (1990). *Creating writers: Linking assessment and writing instruction* (2nd ed.). London: Longman.

Thompson, P. & Tribble, C. (2001). Looking at Citations: Using Corpora in English for Academic Purposes. *Language Learning & Technology*, *5*(3), 91-105.

Tribble, C. (2001). Small Corpora and Teaching Writing: Towards a Corpus-Informed Pedagogy of Writing. In M. Ghadessy, A. Henry, & R. L. Roseberry (Eds.), *Small Corpus Studies and ELT: Theory and Practice* (pp. 381-408). Philadelphia: John Benjamins. doi: 10.1075/scl.5.22tri

Tribble, C. (2002). Corpora and corpus analysis: New windows on academic writing. In J. Flowerdew (Ed.), *Academic Discourse* (pp. 131-149). London: Longman.