Corpus linguistics and sociolinguistic inquiry: Introduction to special issue

TYLER KENDALL and GERARD VAN HERK

1. Introduction

This issue is unlike most issues of *Corpus Linguistics and Linguistic Theory* in that the papers it contains approach corpus linguistics from the explicit perspective of sociolinguistic inquiry. The seven papers presented here were selected from submissions to an open call for papers, which sought work that might speak to current overlaps between these two areas of linguistics. What unites the papers is not a single research question or pursuit, but a general interest in considering the role of corpus linguistics and corpus-oriented methodologies for sociolinguistic research and in thinking deeply about the data used in sociolinguistic research.

There are many methodological overlaps between corpus-based approaches to linguistics and the central pursuits of sociolinguistic inquiry, especially in the variationist paradigm of sociolinguistics pioneered by William Labov (e.g., 1963, 1966), with its interest in the systematic, quantitative analysis of real-world language use. Despite these overlaps (e.g., representative samples of language use, a focus on frequencies of use and probabilities in distribution), which we might expect to lead to a large body of related work in sociolinguistics and corpus linguistics, there is often less interaction between the fields than might be expected (cf. Baker 2010; Kendall under review).

Sociolinguists have rarely considered their work as falling squarely within corpus linguistics; work in many areas of sociolinguistics is less interested in the large datasets, the large-scale analyses, and the quantitative rigor maintained by corpus linguists. At the same time, sociolinguists often describe their data collections as "corpora", but with a few exceptions (e.g., Poplack 1989; cf. Kendall 2008, under review) these sociolinguistic corpora tend to fall short of the definition maintained by corpus linguists (true balancedness, representativity, machine-readability; cf. McEnery and Wilson 2001; McEnery, Xiao, and Tono 2006). Further, sociolinguistic data have rarely been made available beyond the initial research group and therefore lack the "public" orientation that characterizes many standard corpora. From the perspective of standard

Corpus Linguistics and Linguistic Theory 7–1 (2011), 1–6 1613-7027/11/0007–0001
DOI 10.1515/CLLT.2011.001

corpora, much research undertaken in a sociolinguistic vein by corpus linguists has tended not to be embedded enough in social practice to have traction among many sociolinguists. While there have been some attempts to draw sociolinguists to existing corpus linguistic resources (e.g., Bauer 2002; Anderson 2008), few corpora have been developed by corpus linguists that appear to fully meet the needs of sociolinguists, or to be robustly adequate for examining a wide range of sociolinguistic questions (cf. Kretzschmar et al. 2006; Beal, Corrigan, and Moisl 2007a, b; Kendall under review).

In some ways, this issue could be considered as a follow up to the two-volume *Creating and Digitizing Language Corpora* series edited by Beal, Corrigan, and Moisl (2007a, b), which highlights a number of projects developing "unconventional" corpora, or, more generally, as part of a growing interest within sociolinguistics to be more explicit and more publicly minded about data (cf. Kendall 2008). The contents of the issue span several broad topics from questions like those in Beal et al. on how to develop sociolinguistically useful corpora, to how to use those "unconventional" corpora for analyses and how to examine standard, conventional corpora, like the BNC (as in Säily, this issue), for sociolinguistic patterns.

2. Where do corpus linguistics and sociolinguistics intersect?

We, the guest editors, approached this issue and its call for papers with an open mind to where these two broad approaches to linguistics overlap. From the responses we received, we suggest three (albeit somewhat related) themes which emerge in answer to the question "where do corpus linguistics and sociolinguistics intersect?":

Studying language change in real-time and *comparing across datasets*. As Säily (this issue) writes, "the role of corpus linguistics [. . .] is crucial in the study of the past" and, as several other papers in this issue (Buchstaller, Gregersen and Barner-Rasmussen, King et al.) imply, there is little doubt that this is the case. For much of the history of sociolinguistics, research projects have typically been oriented inwards, on their own data, often collected in the field for a particular purpose or to answer a particular question. Broader findings and claims have been made by comparing results from a particular study with results from earlier studies, through the use of agreed-upon methods and metrics – like the variable heuristic and variable rule analysis, Varbrul (cf. Cedergren and Sankoff 1974; Tagliamonte 2006; Wolfram 1993). However, as sociolinguistic data and corpora become available for a longer time-depth, the possibilities for conducting real-time studies of language change increase dramatically, and these real-time studies are best conducted through primary analyses of the raw data, rather than drawing inferences from across numerous studies' findings.

In drawing comparisons across datasets, we find strong connections between sociolinguistic and corpus linguistic practice and much room for contributions to sociolinguistic thinking from corpus linguistics. As papers in this issue indicate, this will likely be an area of growing connection between the two fields. In fact, sociolinguists' orientation to their data as corpora, in the robust sense used by corpus linguists, is beneficial as sociolinguists seek to draw on data across multiple studies, designed sometimes for different purposes and/or through different research or interview designs, rather than just within a single, singularly designed field project. As several papers in this issue explore in depth, questions surrounding comparability are key in projects that deal with disparate data sources.

Sharing sociolinguistic data. Sociolinguistic data collections have largely remained closed resources, available to the original researchers, and possibly other colleagues, but rarely beyond that. It is still extremely rare for a sociolinguistic data collection to be made openly available to the public, especially without some sort of licensing agreement. In some ways this approach is a function of the nature of the data – sociolinguistic interviews often elicit personal information and stories, and the tension between developing natural spoken language data and preserving an informant's rights is a difficult one (see Childs et al., this issue). However, as several papers in this issue discuss (in particular, Childs et al. and Pope and Davis), many groups of sociolinguists are now adopting models from corpus linguistics and documentary linguistics to find ways to balance privacy concerns with making data collections maximally accessible for research. Sociolinguistic data collections (like all natural spoken language recordings elicited for study) are costly to collect, in terms of time and resources, and they are valuable for gaining linguistic insight, oftentimes beyond the design of the original research. The previous, dominant model of considering sociolinguistic data as too valuable to "part with" or to share appears to be giving way to a model where sociolinguistic data is considered to be too valuable not to share. Corpus linguistics' experience with corpus development and distribution is clearly proving to be helpful to sociolinguists here.

3. Contents of the issue

As a prelude to the issue, we outline the papers herein. As we believe readers will note, there are similarities among many of the papers at the same time that each approaches different facets of the possible connection(s) between sociolinguistic work and corpus linguistic work.

Through data available in the LANCHART project, *Gregersen and Barner-Rasmussen* take up questions common in corpus linguistic work and often downplayed in sociolinguistic research – those involving strict comparability of data collected at different times and by different researchers. They focus in depth on the actual content of the interviews: the questions and prompts used by interviewers, the way those initiatives are interpreted and situated by respondents, and how those interpretations change as the social meaning of The Interview changes over time. They elaborate on the need to consider genre-based variation in the content of different interviews from different settings and assess this notion by examining variability and changes in the realization of /æ/, a socially and stylistically variable vowel in Danish.

Comparability is also central to the discussion by *King et al.* of the MAONZE project. They explore the decisions behind expanding a corpus of spoken language recordings of Māori, the indigenous language of New Zealand, to be able to address a large time depth for sociolinguistic and historical study. Unlike many corpus creation projects (but like Gregersen and Barner-Rasmussen), the MAONZE project focuses on developing a dataset for the study of sound change. As such, King et al. have to address a different range of concerns than corpus developers interested in morphosyntax or lexis.

A methodological concern with comparability also surfaces in the paper by *Buchstaller*, which examines in great depth the changing quotative system in the Diachronic Corpus of Tyneside English (DECTE), a new corpus compiled from three sociolinguistic collections of speech from North-Eastern England. Buchstaller's study is one of the most thorough of the English quotative system to date and, among other foci, assesses the benefits and drawbacks of different methods for an analysis. It also demonstrates the ways that "different strategies" in research impact the outcome of the research.

Continuing with the issue of comparability across datasets, *Torgersen et al.*'s project involves combining data from two sociolinguistic spoken language corpora, the Corpus of London Teenage Language (COLT) and the Linguistic Innovators Corpus (LIC). The authors adopt corpus linguistic methods – in particular, frequency and spread (dispersion) analysis – to address a sociolinguistic question, the means of diffusion of changes in pragmatic marker use. From a sociolinguistic perspective, doing so lets them deal with an expanded definition of the sociolinguistic variable ("multiple ways of doing the same thing") without assuming perfect symmetry in the variable context for the use of such forms.

Saily's paper turns our attention to the study of sociolinguistic variation in more conventional corpora. She addresses sex-based and class-based differences in the productivity of *-ity* and *-ness* suffixes in the British National Corpus (BNC), and, in doing so, provides a thorough assessment of methods for the study of morphological productivity in corpus-based and sociolinguistic research. Her work explores a connection not only between sociolinguistics and corpus linguistics, but also to morphological theory.

Pope and Davis describe a spoken language corpus designed for multiple audiences, including health care professionals. Their corpus, the Carolinas Conversations Collection (CCC), is specifically constructed to address the issue of interlocutor effects, through the use of multiple interviews with each individual, differing in the degree to which interviewer and interviewee share social characteristics and (perceived) experiences. The origins of (part of) the corpus in data collected for non-linguistic purposes also permits the opening of a number of questions relating to data sharing and perceptions of the purpose of recorded interviews. As such, it provides a unique vantage point from which to consider the creation and use of (socio)linguistic corpora.

Finally, *Childs et al.* consider corpus creation from the perspective of sociolinguistic fieldworkers and take the issues (or, more accurately, problems) raised by the desire to share data as their major point of discussion – in particular, the tension between the sociolinguist's need to maintain informant confidentiality and the corpus builder's desire to share as much information as possible. They focus here on what can be done in the way of anonymization to resolve this tension (an issue also raised by Pope and Davis, given the sensitivity of their medically-related data).

4. In closing

Taken together, we believe the papers in the issue permit a number of insights (sometimes related across papers and sometimes unrelated) into the connections between corpus linguistics and sociolinguistics and the role that corpus linguistics can play in advancing sociolinguistic inquiry. We can observe from this, along with other recent considerations (Kretzschmar et al. 2006; Beal, Corrigan, and Moisl 2007a, b; Baker 2010), that corpus linguistics – its methods for analysis, its orientation to data, its research questions, etc. – has a growing place in the sociolinguistic endeavor. (While our focus here is not on the contributions of sociolinguistics to corpus linguistic research, we also believe sociolinguistics has a lot to offer corpus linguistics, but that will remain for future consideration.) Yet there remains ample room for further engagement between these disciplines. We hope that readers take from this special issue an increased desire to consider the connections, and to propose new ones of their own.

References

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

Bauer, Laurie. 2004. Inferring variation and change from public corpora. *The handbook of language variation and change*, eds. J. K. Chambers, Peter Trudgill, & Natalie Schilling-Estes, 97–114. Malden, MA / Oxford: Blackwell.

Baker, Paul. 2010. *Sociolinguistics and corpus linguistics*. Edinburgh: Edinburgh University Press.

Beal, Joan, Karen Corrigan, & Hermann Moisl (eds.). 2007a. *Creating and digitizing language corpora. Volume 1: Synchronic databases*. New York / Basingstoke, Hampshire: Palgrave-Macmillan.

Beal, Joan, Karen Corrigan, & Hermann Moisl (eds.). 2007b. *Creating and digitizing language corpora*. *Volume 2: Diachronic databases*. New York / Basingstoke, Hampshire: Palgrave-Macmillan.

Cedergren, Henrietta & David Sankoff. 1974. Variable rules: Performance as a statistical reflection of competence. *Language* 50(2). 333–355.

Kendall, Tyler. 2008. On the history and future of sociolinguistic data. *Language and Linguistics Compass* 2(2). 332–351.

Kendall, Tyler. under review. Corpora from a sociolinguistic perspective.

Kretzschmar, William Jr., Jean Anderson, Joan Beal, Karen Corrigan, Lisa Lena Opas-Hänninen, & Bartlomiej Plichta. 2006. Collaboration on corpora for regional and social analysis. *Journal of English Linguistics*, 34(3). 172–205.

Labov, William. 1963. The social motivation of a sound change. *Word* 19(3). 273–309.

Labov, William. 1966. *The social stratification of English in New York City*. Washington, D.C.: Center for Applied Linguistics.

McEnery, Tony & Andrew Wilson. 2001. *Corpus linguistics*. Second edition. Edinburgh: Edinburgh University Press.

McEnery, Tony, Richard Xiao, & Yukio Tono. 2006. *Corpus-based language studies: An advanced resource book*. New York / London: Routledge.

Poplack, Shana. 1989. The care and handling of a mega-corpus: The Ottawa-Hull French Project. *Language change and variation*, eds. Ralph Fasold and Deborah Schiffrin, 411–51. Amsterdam / Philadelphia: John Benjamins.

Tagliamonte, Sali. 2006. *Analysing sociolinguistic variation*. Cambridge: Cambridge University Press.

Wolfram, Walt. 1993. Identifying and interpreting variables. *American dialect research*, ed. Dennis Preston, 193–221. Amsterdam / Philadelphia: John Benjamins.