Corpora from a sociolinguistic perspective

Corpora sob uma perspectiva sociolinguística

Tyler Kendall*
University of Oregon
Eugene / USA

ABSTRACT: In this paper, I consider the use of corpora in sociolinguistic research and, more broadly, the relationships between corpus linguistics and sociolinguistics. I consider the distinction between "conventional" and "unconventional" corpora (Beal et al. 2007a, b) and assess why conventional corpora have not had more traction in sociolinguistics. I then discuss the potential utility of corpora for sociolinguistic study in terms of the recent trajectory of sociolinguistic research interests (Eckert under review), acknowledging that, while many sociolinguists are increasingly using more advanced corpus-based techniques, many are, at the same time, moving away from corpus-like studies. I suggest two primary areas where corpus developers, both sociolinguistic and non-, could focus to develop more useful corpora: Corpora containing a wider range of non-standard (spoken) varieties and more flexible annotation and treatment of spoken language data.

KEYWORDS: Sociolinguistics; conventional and unconventional corpora; spoken language corpora; data management; annotation methods.

RESUMO: Neste artigo considero o uso de corpora na pesquisa sociolingüística e, de modo mais geral, a relação entre a linguística de corpus e a sociolinguística. Reflito sobre a distinção entre corpora "convencionais" e "não-convencionais" (BEAL ET AL. 2007 a, b) e avalio o porquê de corpora convencionais não terem atraído mais atenção no campo da sociolinguística. Na sequência, discuto a utilidade potencial de corpora para os estudos sociolingüísticos em termos da trajetória recente que tem sido adotada pela pesquisa nesta área (ECKHERT, em avaliação), reconhecendo que, se por um lado, muitos sociolinguistas têm ampliado o seu uso de técnicas avançadas da linguística de corpus, por outro, muitos estão, ao mesmo tempo, se afastando de estudos relaciados a corpora. Sugiro duas áreas principais nas quais compiladores de corpora, independentemente de serem sociolingüísticos ou não, poderiam enfocar para desenvolverem corpora mais úteis: corpora contendo uma amplitude maior de variedades (faladas)não-padrão e um esquema mais flexível de anotação e tratamento de dados orais.

PALAVRAS-CHAVE: Sociolinguística; corpora convencionais e não-convencionais; corpora orais; gerenciamento de dados; métodos de anotação.

* tsk@uoregon.edu

1. Introduction

Much work in sociolinguistics is firmly empirical and based on the analysis, whether quantitative or qualitative, of data of actual language use. As such, sociolinguistics is a field that has many natural connections to corpus linguistics, and these connections have not gone unnoticed. Several recent collections of papers (KRETZSCHMAR; ANDERSON; BEAL; CORRIGAN; OPAS-HÄNNINEN; PLICHTA, 2006; BEAL; CORRIGAN; MOISL, 2007a, 2007b; KENDALL; Van HERK, 2011), articles (BAUER, 2004; ANDERSON, 2008; ROMAINE, 2008), and a book (BAKER, 2010) have explicitly explored some of the relationships between sociolinguistics and corpus linguistics.[1] Despite these connections, however, there is often little direct interaction between scholars in these two fields. For instance, research undertaken on corpora like the British National Corpus (BNC) that might be described as "corpus sociolinguistic" (BAKER, 2010) does not appear to have caught on within mainstream sociolinguists to any large extent.[2]

Why is this the case? That is, why has corpus linguistics not had a larger influence on sociolinguistics? Beal *et al.* (2007a, b) made the useful clarification that much sociolinguistic work involves what they termed "unconventional" corpora, corpora that do not fit the standard mold of resources like the BNC. In fact, their volumes sought

> to establish whether or not annotation standards and guidelines of the kind already employed in the creation of more conventional corpora on standard spoken and written Englishes … should be extended to less conventional corpora so that they too may be 'tamed' in similar ways (BEAL *et al.*, 2007a: 1).

[1] Also, see Kretzschmar's (2009) *The Linguistics of Speech* for an interesting and helpful discussion of some historiographical connections between sociolinguistics and corpus linguistics.

[2] Before proceeding, it is worth commenting that "sociolinguistics" is a term that covers a diverse set of approaches to linguistics across several disciplines and encompasses many different traditions of research, ranging from, e.g., areas of linguistic anthropology, sociology, discourse studies, variationist linguistics, and so on. (Of course, "corpus linguistics" can also be considered a cover term for a number of different approaches to linguistics.) In this paper, I will often refer to "sociolinguistics" and "sociolinguists" in monolithic terms, but I realize that I am perhaps over-generalizing. It may help to explain that my own background is from the "variationist approach," the field-based, quantitative study of language variation and change pioneered by scholars like William Labov and Walt Wolfram. Some readers may find my point of view overly influenced by that flavor of sociolinguistics.

While the main question in this passage, about the applicability of standard corpora annotation to other kinds of corpus-like data, is a difficult question (and ultimately I will steer away from it, offering an alternative possibility in §5.2 of this paper), this notion of "conventional" and "unconventional" corpora is a useful one. Sociolinguistic research often focuses on non-standard varieties of language, and spoken language in particular, and the large "conventional" (i.e. standard language and often primarily written) corpora have simply not been of great use for pursuing sociolinguistic research on non- or less-standard varieties.[3]

In this paper, I consider the role of corpora and corpus linguistic methodologies in sociolinguistics and the division between conventional and unconventional corpora further. I begin, in section 2, by considering the role of corpora (broadly defined) in sociolinguistics. In section 3, I look at the ways that "traditional" corpora have been used to ask sociolinguistic questions, and consider why more work one might describe as "corpus sociolinguistic" (BAKER, 2010) has not been undertaken. I follow this up in section 4, by reviewing recent trajectories of research interest in sociolinguistics and considering how this impacts the relationship(s) between corpus linguistic and sociolinguistic work. In section 5, I consider what future directions corpus linguistics, and corpus development in particular, could take in order to facilitate corpus-based research on sociolinguistic questions. Finally, in section 6, I close with some summary comments.

[3] As a side note, the implementation of representativeness (McENERY; WILSON, 2001; McENERY; XIAO; TONO, 2006) in the construction of the major conventional corpora may be limiting from a sociolinguistic perspective. Corpora like the BNC, the Corpus of Contemporary American English (COCA), and those in the Brown family, in attempting to represent national varieties, are by necessity somewhat normative and exclusive – they downplay and/or normalize over the true diversity of language at a national scale. The notion of representativeness – what larger population of language use or users a corpus trustworthily samples – is crucial in all corpus-based work (see also GRIES, 2006). Yet, the sampled variability included in conventional corpora is often ordered along the dimension of register or genre, not the dimension(s) of social variation. This cannot be helped – as impressively large as, e.g., COCA is (at over 410 million words), it would be impossible for it to fully represent, say, the ethnic diversity of English in the U.S. For that matter, how does one even assess the full extent of ethnic diversity of English in the U.S.?

2. The place of corpora in sociolinguistics

Many approaches to sociolinguistics involve the analysis of bodies of naturally occurring talk. The size of these "bodies" of data range from quite small, such as a single conversation or story, to massive, e.g., hundreds of hours of recorded interviews collected over years of fieldwork. Increasingly, sociolinguists are calling these datasets "corpora".[4] Whether these corpora meet the "proper" definition maintained by corpus linguists (balanced, representative, machine-readable; cf. McENERY; WILSON, 2001; McENERY; XIAO; TONO, 2006) or not (they most often do not), they can still be usefully examined via corpus-based methodologies. Techniques such as examining concordances and collocational patterns, conducting keyword analyses, and the use of corpus analysis software tools themselves can shed useful light into even quite small datasets (cf. BAKER, 2010).

While it sometimes seems to be the case that sociolinguists' borrowings from corpus linguistics are shallow (and, as mentioned in footnote 4, possibly at times only name-deep), corpora certainly have a growing place in sociolinguistic research and some connections have already existed for decades. For instance, in an important paper, Poplack (1989) detailed the creation of her at-the-time "mega-corpus" of spoken Ottawa-Hull French, which contains 3.5 million words from 270 hours of recorded speech.[5] Other early

[4] Considering terminology further, "corpus" and "sociolinguistics" are both terms that are used variously by different groups of scholars. Within corpus linguistics, it often seems to be the case that "sociolinguistics" is used as a cover term for all kinds of corpus-based research that involves extra-linguistic factors. But among sociolinguists, much of this research fails to have traction; it is not always seen as sociolinguistic, or at least as "sociolinguistic enough". Meanwhile, there is also an increasing tendency for sociolinguistic researchers to consider and discuss their data as "corpora" in ways that over-generalize that term. More and more sociolinguistic field-based projects appear to outcome in collections of data that are named as corpora (e.g., hypothetically, "Smalltown USA Corpus"), when for corpus linguists they often have none of the characteristics of "corpus proper". It is hard to see where one draws the line between the "unconventional" corpora described in Beal *et al.* (2007a, 2007b) and data collections that simply are not appropriately considered "corpora". I do not want to dwell on terminological issues too much in this paper, but it is worth considering whether some of the most obvious current connections between sociolinguistics and corpus linguistics may only be name deep.

[5] While 3.5 million words may not seem like a "mega-corpus" to present day corpus researchers, it should be remembered that this word count reflects only natural, conversational spoken language, and, especially for a resource created over twenty years ago is an extremely impressive accomplishment. Within sociolinguistics, it remains one of the largest corpus-like corpora.

sociolinguistic work also focused extensively on describing aspects of their projects that in today's terms would likely be described as "corpus creation" (e.g., SHUY; WOLFRAM; RILEY, 1968; SANKOFF, D.; SANKOFF, G., 1973; cf. KENDALL, 2008). Tagliamonte's (2006) sociolinguistics textbook, *Analysing Sociolinguistic Variation*, outlines an approach to (variationist) sociolinguistics that many corpus linguistics would likely be comfortable with (and would, I think, find quite useful).

It seems clear that in coming years sociolinguistics will make increasing use of corpora and will increasingly interact with corpus-based approaches to linguistics from other areas. Endeavors like the Origins of New Zealand English (ONZE) project (GORDON; MACLAGAN; HAY, 2007), the Newcastle Electronic Corpus of Tyneside English (NECTE; ALLEN; BEAL; CORRIGAN; MAGUIRE; MOISL, 2007), and the Danish LANCHART project (LANguage CHAnge in Real Time; GREGERSEN, 2009), all of which involve the creation of impressive "unconventional" corpora, point to the fact that sociolinguists are paying more serious attention to corpus-based methodologies and the benefits of explicit corpus creation work.

There are also growing connections between sociolinguistics and corpus linguistics in terms of specific research. For instance, Torgersen, Gabrielatos, Hoffmann, and Fox (2011) provide an excellent example of how corpora and corpus linguistic methodologies can be used to pursue core sociolinguistic questions, such as the actuation of language change (WEINREICH; LABOV; HERZOG, 1968). In this work, they examine pragmatic markers (such as "innit" and "if you know what I mean") in two corpora of London speech, the Corpus of London Teenage Language (COLT) and the Linguistic Innovators Corpus (LIC). Through an analysis using the corpus-based heuristics of frequency and spread (i.e. dispersion among speakers) of pragmatic markers, Torgersen et al. shed light into the locus of language change in the highly complex and multi-ethnic urban center of London.

In sum, corpora and corpus-based methods have an important and still growing place in sociolinguistic research. Yet, the similarities are often approximate, and the connections often still indirect. Again, as Beal *et al.* (2007a, 2007b) pointed out, sociolinguists' corpora are typically "unconventional", not conforming to the models used in crafting major corpora like the BNC and the Corpus of Contemporary American English (COCA). In the next section, I change focus from corpora in sociolinguistic research, to sociolinguistics in corpus-based research to discuss the relationship between these two approaches to language more closely.

3. Corpus-based sociolinguistics

There have been several calls for scholars to mine traditional (i.e. "conventional") corpora for sociolinguistic research applications (BENDER, 2002; BAUER, 2004; ANDERSON, 2008; ROMAINE, 2008; BAKER, 2010). Baker's (2010) recent book, *Sociolinguistics and Corpus Linguistics*, points out that the previous literature in these two fields indicates "that some form of 'corpus sociolinguistics' is possible, although it might appear that corpus linguistics has made only a relatively small impact on sociolinguistics" (BAKER, 2010, p. 1). Baker provides many examples of research in support of a "corpus sociolinguistics", such as work on the BNC on sex-related language differences (SCHMID, 2003) and broader social differences (RAYSON; LEECH; HODGES, 1997). Research on sex-based language differences indeed seems an area of sociolinguistics well suited to corpus-based research.[6] One example, not reviewed by Baker, is recent work by Säily (2011; SÄILY; SUOMELA, 2009), who examined the relationship between speaker/writer sex and morphological productivity in the Corpus of Early English Correspondence (CEEC) and the BNC and demonstrated ways in which the productivity of the *–ity* and *–ness* suffixes were similar or differed for males and females in the historical and present-day corpus data – findings of both sociolinguistic and morphological theoretical interest. A second, and perhaps better-known example (and one which is discussed at length in BAKER, 2010) is Biber's extensive research on genre- and register-based variation (e.g., BIBER; FINEGAN, 1989), which has clearly shown the value of corpora for understanding this important dimension of language variation.

Yet, for the most part (and as Baker points out), these kinds of standard corpora have had limited appeal and limited use by sociolinguists. I believe this is not entirely unexplainable. Many sociolinguists are interested first and foremost in spoken language, which is less available in conventional corpora than written language (cf. NEWMAN, 2008). Further, sociolinguistics is often about *fully* or at least *richly* situating language use in its social and interpersonal contexts and standard corpora, even those that are comprised of spoken language

[6] Arguably, the sex of an author or speaker is the easiest social/identity-related factor to annotate in a corpus, or to reconstruct post factum from corpus data. Note, however, that "gender," a socially constructed identity-related variable, is different from "sex," the biologically-based variable, and more difficult to evaluate without ethnographic inquiry (cf. CHESHIRE, 2004; BUCHOLTZ; HALL, 2006).

recordings, are often too divorced from social contextual information to be of use for in depth sociolinguistic study. Most spoken language corpora which currently do exist – such as many of those available from the Linguistics Data Consortium (LDC; http://ldc.upenn.edu/) – have been designed for speech technology and natural language processing research and simply do not capture the kinds of information that would be necessary for even basic sociolinguistic research (such as, the socio-economic class or ethnicity of the talkers).

It is clear from, for instance, work on the BNC (again, RAYSON *et al.*, 1997; SCHMID, 2003; SÄILY, 2011) that certain questions of a sociolinguistic nature, such as the differences in language use by sex, are fairly pursuable through standard and written corpora, but it is much less clear how one would look at more nuanced sociolinguistic patterns through these kind of corpora. For instance, it is difficult to see how researchers could examine social structures like "communities of practice" (WENGER, 2000; e.g., MALLINSON; CHILDS, 2007), groups built around a coordinated set of interests and activities, through a pre-existing corpus, and it is these sorts of questions which have become of most interest to a large number of sociolinguists in recent years. (I will return to this point in the next section.)

Further, much corpus linguistic work that does see itself as sociolinguistic focuses on lexis (i.e. examines socio-cultural, or, e.g., sex-based, differences through lexical patterns in corpora). Many of the studies reviewed by Baker (2010) are of this kind. For instance, Baker (2010, p. 70-73) gives examples of research on personal titles (such as "Mr." and "Mrs.") and gendered nouns ("boy(s)" and "girl(s)") and their changing use over time in the Brown family of corpora. Rayson *et al.* (1997) examine social differences in lexical frequencies in the BNC. These sorts of research endeavors are clearly sociolinguistic, in the sense that they inform us of changes in social structure and/or changes in the discourse on social structure, and have some similarities to work that is conducted in squarely sociolinguistic circles (e.g., D'ARCY; TAGLIAMONTE, 2010, who examine the complex sociolinguistic factors influencing the realization of relative pronouns in spoken English in Toronto). Nonetheless, many sociolinguists in recent years have focused more on morphosyntactic patterns, which are substantially harder to mine through corpus-based methods, and sociophonetic patterns, which seem even less analyzable through corpus methodologies. (Of course, SÄILY, 2011 examined morphological patterns, illustrating that "corpus sociolinguistic" work does not need to be limited to lexis.)

To some degree the lesser interest in lexis may have to do with sociolinguists' focus on spoken over written language. The primacy of words in corpus linguistics – e.g., the fact that we count a corpus' size in terms of word-count – does not perfectly fit research on conversational spoken language, which is characterized by a high occurrence of disfluencies, grammatical "errors," mispronunciations and the like.[7] It is hard to imagine how the sampling frame for the Brown family of corpora, for instance, 500 samples of about 2,000 words each (extracted from the start of a sentence to the completion of the first full sentence 2,000 words later; FRANCIS; KUČERA, 1979; see also BAKER, 2010, p. 59-68), could be appropriately applied to conversational spoken language.

Finally, it must be observed that many sociolinguists are interested in *specific* language varieties or language situations – such as the language practices of a particular (often small) community or social group – and, as a result, pre-existing corpora containing normative or standard varieties are simply not of great interest. Of course, corpora like the BNC and COCA are always valuable as benchmarks for assessing features in other varieties (and the Brown corpus is still a major source for word frequency information for a diverse range of research) but their use as the actual primary data of analysis is much less common among many sociolinguists.

In closing this section, I would agree that Baker's (2010) "corpus sociolinguistics" indeed appears possible and, I would argue, is being realized by some researchers, although I would also note that the uptake for this kind of work appears to be greater among researchers coming from corpus linguistic perspectives than among those coming from sociolinguistic backgrounds. Sociolinguists have been slower to adopt conventional corpora for research, for the reasons I outlined above, one of which, I revisit more fully now.

4. Convergence and divergence

The previous sections of this paper have illustrated, I think, a complex relationship between sociolinguistics and corpus linguistics. On the one hand, these fields have clear similarities. On the other, they also have clear differences.

[7] For instance, in Kendall, Bresnan, and Van Herk (forthcoming), a study of the variable pattern between *give* theme-NP recipient-PP and *give* recipient-NP theme-NP in African American English (discussed again below), we only approximate the word-count of our transcribed spoken data, finding it unrealistic to give the dataset an exact figure.

We observe that in some ways sociolinguistics and corpus linguistics have always been converging at the same time that we observe they have always been diverging in other ways. A consideration of Eckert's (2005, under review) paper, "Three waves of variation study", provides further light on this situation.

In this historiographical assessment of quantitative sociolinguistic work, Eckert classifies the study of sociolinguistic variation into three major categories, or "waves." The first wave is characterized by the study of broad correlational patterns between social features of talkers (and writers) and their use of variable language features. The second wave of study involves ethnography and studying smaller groups of speakers (and writers) to greater depth, focusing on more local patterns of language use. The third wave of study is about practice and agency, rather than social structures. Instead of searching for categories which correlate with language use, research in the third wave focuses more closely on understanding styles and the construction and negotiation of identity(/ies) rather than broad patterns of individual variable features. Eckert points out that these three waves are not necessarily chronologically ordered. Labov's (1963) ground-breaking first study – on Martha's Vineyard, with its deep ethnographic analysis of a small community – is seen as in the second wave, while his second (1966) foundational study – a large-scale survey of English in New York City – is seen as squarely first wave. Yet, despite there not being a direct chronology that corresponds to the three waves, current interest in sociolinguistics is moving increasingly towards third wave-like approaches (see also COUPLAND, 2007). First wave, and to a lesser extent second wave, sociolinguistic research would appear to fit comfortably within a corpus linguistics mold. It is in these "waves" of research that we can draw strong connections between sociolinguistic and corpus linguistic methods and practice where the quantitative large-scale analysis of corpora is most helpful. However, the focus and methodologies of third wave research appear to share less with corpus linguistics (and perhaps have more similarities with Conversation Analysis, cf. LIDDICOAT, 2007, than they do with large-scale corpus-based research).

Eckert provides a nice summary of a major way that third wave work differs from the earlier waves, especially the first wave – the kind of sociolinguistics most implementable through corpus methods.

> The survey method's primary virtues are coverage and replicability, both of which depend on the use of pre-determined social categories and fairly fleeting social contact with the speakers chosen to represent

> those categories. As a result, the social significance of variation can be surmised only on the basis of a general understanding of the categories that serve to select and classify speakers. There is no question that the broad demographic patterns of variation are important. But just as a map of New York City does not tell you what the streets are like, or what it's like to walk on them, the macro-sociological patterns of variation do not reveal what speakers at different places in the socioeconomic hierarchy are doing socially with those variables (Eckert under review: 6).

Further,

> [the] move from the study of structure to the study of practice, giving agency its place in the analysis, has defined the recent history of the social sciences and recent intellectual history more generally […]. It does not negate the importance of structure, but emphasizes the role of structure in constraining practice and, in turn, the role of practice in producing and reproducing structure. In the study of variation, a focus on practice brings meaning into the foreground, as we try to get at what speakers are doing on the ground. At the same time, it moves us closer to the goal of studying the actual process of change (Eckert under review: 14).

These passages help explain the tension in actualizing a "corpus sociolinguistics." "Coverage and replicability" are two major tenets (and advantages) behind corpus-based work. Yet, it appears to be an impossible task to make replicable and generalizable, especially through corpus-based methods, the ethnographic and instance-specific knowledge a researcher must gain in order to understand the actual creation and negotiation of social meaning "on the ground."

From this, it seems that some sociolinguists will continue to be uninterested in corpora and corpus methods. Nonetheless, there are concrete steps that corpus developers could take to enhance the possibilities of a "corpus sociolinguistics" and to increase the utility of corpora for pursuing sociolinguistic research, and I turn to these now.

5. The future of corpora in sociolinguistic research

Technological advancements have been paramount in the development of sociolinguistics. The same is true of course for corpus linguistics. The current research in both approaches would be impossible without modern recording equipment and the ability to store, process, and analyze large amounts of text and audio data through computerized means. While it may be the case that sociolinguistics and corpus linguistics diverge in coming years

in their research orientations and methodologies in some ways (as is indicated by Eckert's third wave), it seems likely that continued technological advancements in the development, annotation, and analysis of corpora will lead to increased opportunities for sociolinguistic engagement with corpora. This is especially true for research investigating aspects of language and social structure (i.e. work in the first and second waves), though I believe it is still the case for work that is less interested in large-scale quantitative study. All researchers working with recorded data can benefit from advancements in the treatment of these data.

Some of these advances will occur, I believe, without the need for an explicit "call to arms." Nonetheless, I here explicate two areas where I suggest corpus work could immediately benefit sociolinguistic research, and, conversely, insight from sociolinguistics could enrich broader corpus-based research: The creation of (spoken language) corpora for more diverse language varieties, and the implementation of annotation schemes that are more flexibly connected to data. I consider these in turn.

5.1. The need for large, publicly available corpora of more diverse (spoken) language varieties and increased sharing of existing data

One might argue that a primary benefit of corpus-based approaches to linguistic analysis is that the development, publication, and sharing of public corpora allows for the best possible advancement of empirical knowledge about language. By allowing (and, further, promoting) the repeated and repeatable analysis of the same publicly available datasets, corpus linguistics fosters an environment that more fully fits the "scientific method" mode of research than many other areas in linguistics. Scholars can question and refine previous findings by (re-)analyzing the original data; they can extend or modify the annotation schemes and data coding used in previous research; they can compare previously analyzed datasets directly to newly developed datasets; and so on. By working from a shared pool of data, researchers are best able to collectively develop agreed upon knowledge about language. This, I believe, is a major benefit of corpus-based work (which in my opinion has been under-boasted about by corpus linguists).

The vast bulk of sociolinguistic research, even that based on thoroughly balanced and representative linguistic databases, has been conducted on proprietary datasets that are not available for peer review or outside

consideration. The common practice in sociolinguistics is for individual (groups of) researchers to develop highly specialized, but closed, databases, which are not made widely available to outsiders. This tendency is not ill intentioned, but rather is the outcome of historical processes in the field. A huge amount of effort, time, and money goes into the collection of sociolinguistic data (and the compilation of any spoken language dataset; NEWMAN, 2008) and within sociolinguistics (as, unfortunately, with many disciplines), academic "credit" has come from the analysis of the data and not its collection or compilation. Researchers traditionally have not wanted to get "scooped" (cf. CHILDS; Van HERK; THORBURN, 2011) on findings after doing the extensive and expensive work of data collection.[8] A second, and perhaps bigger, reason has related to rights management and informant privacy issues, since sociolinguistic fieldwork and interviewing often captures sensitive information that the informants may not want to make public or which fall under contracts with human subjects boards and have restricted access. These issues of anonymity and privacy are complex and difficult to answer when deciding to share fieldwork data (CHILDS *et al*. 2011). Finally, since sociolinguistic datasets have typically been developed in order to research a specific question or set of questions, it has often been assumed that once the original questions have been studied in depth there is not further interest in the datasets themselves. This trend of closed data appears to be changing and it is now the case that more groups of sociolinguistic researchers are making their data available to colleagues and to the public (cf. KENDALL, 2008; CHILDS *et al*. 2011), but it remains the case that sociolinguistics has so far not been able to benefit from the kind of peer review only possible when datasets are widely available for review and re-analysis.[9] This has also, of course, limited the ability of other (i.e. less sociolinguistically oriented) corpus linguists to draw from the vast amount of data collected in recent decades by

[8] As a reviewer aptly pointed out: One hoped for part of a solution would be greater recognition for corpus development work in career advancement, like promotion and tenure, so that corpus developers (sociolinguistic or not) were less incentivized to limit access to their data.

[9] The recent founding of journals like the *Journal of Experimental Linguistics*, <http://elanguage.net/journals/index.php/jel>, which focus on "reproducible research" and the publication of full datasets along with research articles, represents an exciting turn for areas of language research outside of corpus linguistics, most of which, like sociolinguistics, have heretofore, not made a general practice of working from shared data.

sociolinguists. One could imagine there being much richer corpora available, especially "conventional" corpora, if the developers of those corpora could draw on the spoken language data collections of sociolinguists.

To give a specific example, African American English (AAE) has been studied at exceptional length in North American sociolinguistics and has been the subject of a vast body of empirical and quantitative investigations (cf. SCHNEIDER, 1996, p. 3). This research has been driven by numerous exciting questions, from those involving diachrony – such as, how did AAE form in the first place? Is present day AAE the outcome of pidgin/creole forms or of a working class, slave-master variety of British English? Is present day AAE converging with, or diverging from, white varieties or regional varieties of American English? – to more applied sorts of questions about topics like education and social justice – such as, what are the educational positions and responsibilities of school systems towards AAE-speaking children? – and so forth. Many scholars have researched these questions (to list just a few: McDAVID, R.; McDAVID, V., 1951; WOLFRAM, 1969; LABOV, 1972a; DILLARD, 1972; FASOLD, 1972; SMITHERMAN, 1977; 1981; HEATH, 1983; RICKFORD, 1999; POPLACK; TAGLIAMONTE, 2001; WOLFRAM; THOMAS, 2002; CRAIG; WASHINGTON, 2006) and, while consensuses have emerged among sociolinguists in some areas, many questions are still quite actively pursued. However, one could argue that additional progress could be made if scholars had access to a large, shared pool of data against which they could test competing theories or could cite broadly available evidence in order to support or refute particular positions.

While some groups of sociolinguistic researchers have invested in developing thorough transcription and annotation schemes for their data (e.g., POPLACK, 1989; cf. TAGLIAMONTE, 2006), many other sociolinguists do not work with transcribed data, but rather code just the features of interest directly from the audio recordings (e.g., MILROY; GORDON, 2003, though one infers this point through the lack of discussion rather than an explicit statement about transcription). Thus, there are massive amounts of sociolinguistic recordings, which are simply not available in forms that avail themselves to corpus linguistic approaches. The costs of developing complete "corpus-like" data collections can unfortunately be too high, especially when the research questions at hand (often involving particular sociolinguistic variables, cf. WOLFRAM, 1993; MILROY; GORDON, 2003) are more quickly pursued by extracting just the tokens of interest from the audio recording rather than transcribing and annotating everything available.

In recent work investigating the dative alternation in African American English, Kendall, Bresnan, and Van Herk (forthcoming) attempted to take stock of the amount of transcribed sociolinguistic AAE data that was available if one pooled data from across several research groups. All told, we obtained only about a quarter million words of transcribed AAE speech, even though many scholars were extremely generous in making data available to us for analysis. This is not to say that a quarter million words is all that exists, but rather that these data (i.e. accurate transcripts of AAE speech) are scattered throughout the field and not available in any aggregateable form for corpus-based research. It seems clear that doing corpus-based analysis on AAE will require further corpus compilation and creation work.

In sum, countless researchers would be greatly aided by the availability of a large, publicly available corpus of African American English. And this is just one example of a non-standard variety of English. We can readily imagine how many language researchers would benefit from corpora developed for other varieties and varieties of other languages. We need more large-scale publically available corpora of non-standard language varieties.

5.2. Connecting "data" to data and the question of "taming"

A second area from which sociolinguistic research could benefit would be a greater focus on the kinds of annotation available in corpora. Corpus linguists – and also documentary linguists, natural language processing researchers, and others for that matter (e.g., BIRD; LIBERMAN, 2001; SIMONS; BIRD; SPANNE, 2008) – have developed extensive annotation frameworks, but often these annotation frameworks have not focused on capturing some of the information that sociolinguists are most interested in, such as a fuller range of social and demographic information about the speakers/writers and audiences in corpora, as well as the full interactional context and setting of the data.[10] In his "ethnography of speaking" approach

[10] In some cases, it would be more accurate to say that it is the entry of the annotation for particular corpora that fail to capture enough information to be widely useful for sociolinguistic queries rather than failings in the annotation *schemes* themselves. The annotation framework for the 4.2 million word demographically sample spoken portion of the BNC, for example, was designed to capture quite a range of demographic features for the speakers – including speaker sex, age group, education level, occupation, social class, and dialect background. For the recruited participants (those who agreed to carry the recording device) the information for many of these

to language, Hymes (e.g., 1974), for instance, proposed the S-P-E-A-K-I-N-G model, which in present day terms could be understood as an annotation framework. Many of the S-P-E-A-K-I-N-G model's components are included in standard corpora (such as information about the "Genre", and often "Participants"), but many are not (such as the "Act sequence" or "Key"). This is not to propose that Hymes' model in particular be adopted by corpus developers, but more simply to highlight some of the kinds of annotation that would further sociolinguistic research possibilities through corpora and, more generally, might lead to richer annotation frameworks than are most often currently used.

Of course, there are huge difficulties in implementing these kinds of ethnographically informed annotation systems in a general way. They are often not readily applicable on a wide-scale, or individual annotation schemes are too bound up with a specific project, or a specific researcher's agenda, to be of use beyond a specific corpus or a specific research project. Even social measures that may seem straightforward at first glance, like socio-economic class or education level, must often be contextualized for the particulars of the group under study (cf. SANKOFF; LABERGE, 1978). For instance, when studying the language use of non-mainstream populations, such as rural African Americans in the U.S. South, mainstream conceptions of socio-economic class or social status simply do not seem to be relevant and local understandings of social structure are necessary for in depth sociolinguistic research (KENDALL; WOLFRAM, 2009). How to best achieve the kind of annotation necessary to make cross-group comparisons in these sorts of situations, or whether such annotation is possible in the first place, is a difficult question to answer.

This question, however, returns us to the quote on the first page of this paper (BEAL *et al*. 2007a, p. 1). In their two edited volumes about "unconventional" corpora, Beal *et al*. (2007a, 2007b) discuss the difficulties of "taming" these unconventional corpora. Poplack, in her foreword to the volumes, explains,

---

fields is available (since the recruits were solicited based on their region, sex, age group, and social class), but for the talkers with whom the recruits interacted much less information is known. There are also additional problems that must be considered (such as, inaccurate or even false information) when relying on the self-disclosure of social information from recruits such as those in COLT and the BNC (cf. STENSTRÖM; ANDERSEN; HASUND, 2002, p. 18-19).

> Taming, as understood here, is largely a question of representation: How to represent forms for which there is no standard orthography, what to represent, how much to annotate, how much analysis to impose on the materials, how to represent ambiguities and indeterminacies, how to represent the finished product to the end-user (POPLACK, 2007, p. ix-x).

While I find Beal *et al*.'s discussion helpful (and Poplack's foreward particularly insightful), and while the papers in their volumes provide an excellent overview of current work on unconventional corpus development, I am not sure that I like the term "taming" for what needs to happen for less standard language datasets to be usefully developed into corpora. It seems to me that one reason traditional corpora have not been used as extensively for sociolinguistic research is precisely because they have been extensively "tamed," and this "taming" has rendered them less sociolinguistically "real" or useful.[11] A more preferable model might be one which embraces the multi-dimensionality of spoken language data and attempts to maintain the full richness of those dimensions through the corpus development process. (I resist the temptation to label this something like "data left in the wild".)

In Kendall (2008), I proposed a model for considering data within sociolinguistics that attempts to maintain close connections between layers of annotation or metadata. Crucially, this involves being explicit about layers of abstraction (steps away from the original source data) in our annotation and metadata creation processes. Figure 1, from that paper, contrasts what I consider to be a traditional approach to sociolinguistic analysis and data management with an approach that I believe has greater benefits. The basic premise is that sociolinguists are interested in understanding patterns of language in their social contexts, but that all quantitative work (or in fact any work based on records of speech, including audio-only and even video recordings, since recordings never capture the entirety of a real-world event)

[11] For example, the BNC's demographically sampled spoken component was built following social survey research practices (BURNARD, 2007; see also RAYSON *et al*., 1997) and, at face value, appears to be quite similar to the sort of large-scale dialect survey sociolinguists might undertake. Rayson *et al.* (1997), in their examination of social factors in differences in lexical frequency in the spoken component, note however that work on the social differentiation of language in this part of the corpus is limited by the simplified transcription system. One might argue that it is primarily the extent of its "taming" that makes this part of the BNC less sociolinguistically useful than it otherwise would be.

involves abstractions away from the true, contextualized language data, the actual real-world speech event. In the "traditional" model, layers upon layers of annotation are developed, many of which increase the distance between the "data" (in quotes, indicating some level of abstraction from the actual or ideal data) and the real-world speech events that are ultimately the objects of interest, the true data (no quotes).

For example, if I am interested in studying variable realizations of the English past tense (like unmarking or non-standard past tense marking), I might audio record a speech event and from that recording develop a transcript, which, for sake of the example, we will assume accurately captures the variable realizations of the English past tense morpheme. I then extract the frequencies of the various realizations of the past tense morpheme along with other contextual information and then compile this as a spreadsheet, which I add to compiled data from other speakers and other speech events. In the end, I have a data file ready for quantitative analysis, but I have also moved several steps away from the original speech event. My language data has become a spreadsheet of frequencies or data tokens with very little available matrix talk, perhaps a concordance-like "keyword in context" amount of surrounding context. It is no longer quite "language," having been separated from its full communicative context. This likely does not matter as far as the success of my quantitative analysis goes, but the closer examination of individual tokens has become difficult, as has my ability to question the original coding of the morphemes.

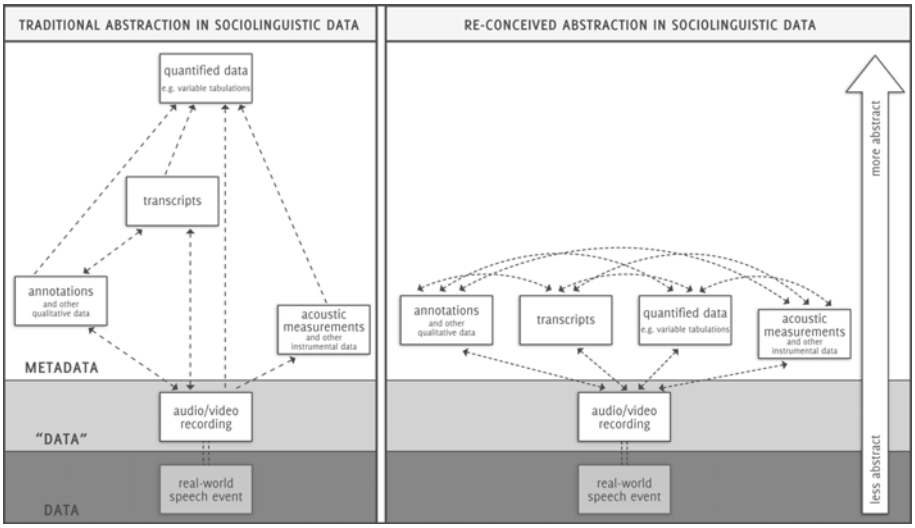

FIGURE 1 – Layers of abstraction in sociolinguistic data (from KENDALL, 2008, p. 346, Figure 5)

The "re-conceived" model of Figure 1 focuses primarily on maintaining linkages between levels and types of annotation. As in the hypothetical example discussed for the traditional model, I may wish to transcribe the recording and then to extract quantitative data from that. However, here the emphasis would be on maintaining links between each of these layers of data with the other layers. This is achieved through a focus on accurate time-stamping and the development and use of software built for time-aligned linguistic (or at least audio) annotation. Returning to Hymes – who wrote "the most common, the most serious, defect in most reports of speaking probably is that the message form, and, hence, the rules governing it, cannot be recaptured" (1974, p. 54) – we can observe that, while an accurate transcript may capture the lexical and syntactic form of an utterance, no transcript or text-based annotation can be expected to accurately encapsulate its full form, such as its prosody, the nuanced particulars of the speaker's voice, and so on.

The Sociolinguistic Archive and Analysis Project (SLAAP; <http://ncslaap.lib.ncsu.edu/>; cf. KENDALL, 2007, 2008) and the Online Speech/Corpora Archive and Analysis Resource (OSCAAR; <http://oscaar.ling.northwestern.edu>; cf. KENDALL, 2010) are two examples of ways that one might approach implementing this sort of model. Both of these projects feature a time-aligned transcription model which is dynamically linked to the underlying audio recordings and to any additional researcher notes or quantitative data. For example, Figure 2 displays one view of SLAAP's transcript feature for a stored recording. In addition to the transcript text, the user has direct access to the recording audio, as well as to fine-grained information about where silences occur and their lengths. Users can also get "close up" views of individual transcript lines, as in Figure 3, which displays the text of a line along with the audio itself, as well as a spectrogram and pitch track for the utterance (created dynamically from the audio). Users can extract phonetic information directly from this view (only pitch data is illustrated in Figure 3).

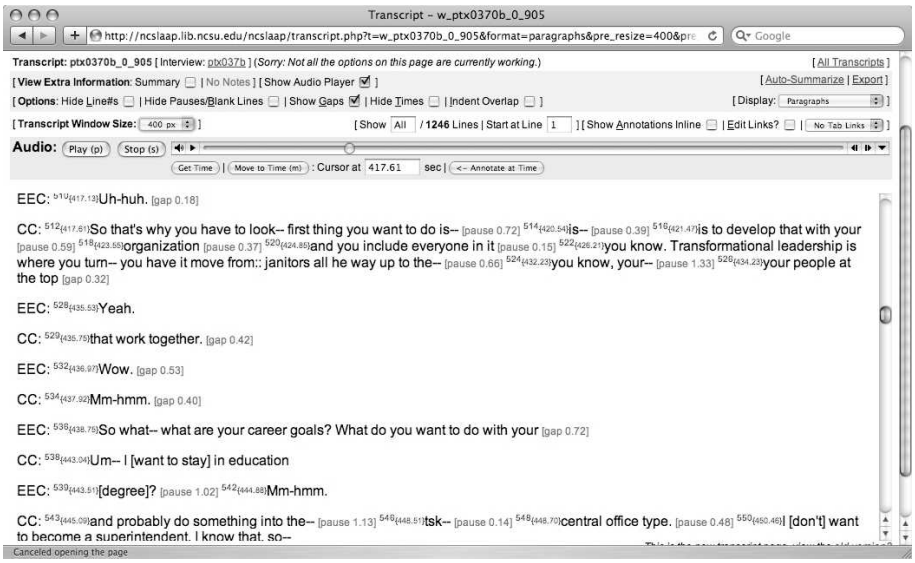

FIGURE 2 – A transcript view in SLAAP (a sociolinguistic interview with "CC", a Mexican American female in Southern Texas; "EEC" is the interviewer)

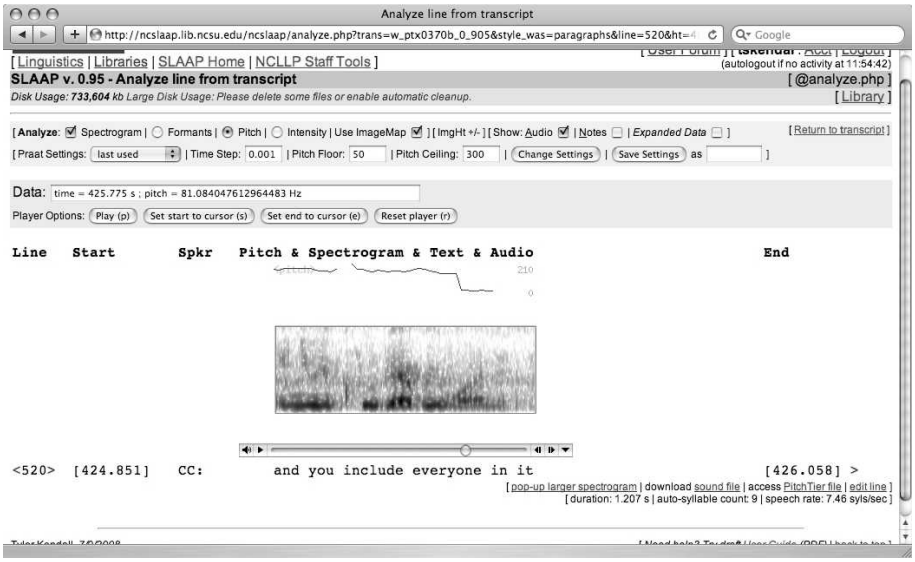

FIGURE 3 – SLAAP's "close up" of the line "and you include everyone in it" from Figure 2

Transcripts in SLAAP are dynamic entities and can be reformatted in numerous ways, from textual representations, like the columnar format suggested by Ochs (1979), to various graphical formats (screenshots of these

other transcript views and a fuller discussion of SLAAP's transcript model are available in KENDALL, 2007, 2008). Traditional corpus analysis features are available, such as in Figure 4, which displays the highest frequency bigrams (on the left) and a sample concordance (on the right; for the phrase "high school") from the same transcript shown in Figures 2 and 3. Since all of the utterances are time-stamped, SLAAP is able to show a graphical timeline (at the top-right) indicating where each of the concordance lines occurs in the recordings (the single line that extends the length of the timeline image represents the temporal duration of the full recording; the filled bar that extends roughly across the left-half of the line represents the transcribed portion of the recording; the dots below the lines show when the concordance lines occur in time). While SLAAP's maintenance of the linkage between audio and text is a centerpiece of the archive, transcripts can also be exported as plain text (or as Praat TextGrids; BOERSMA; WEENINK, 2010) and then manipulated via standard corpus or text analysis tools.

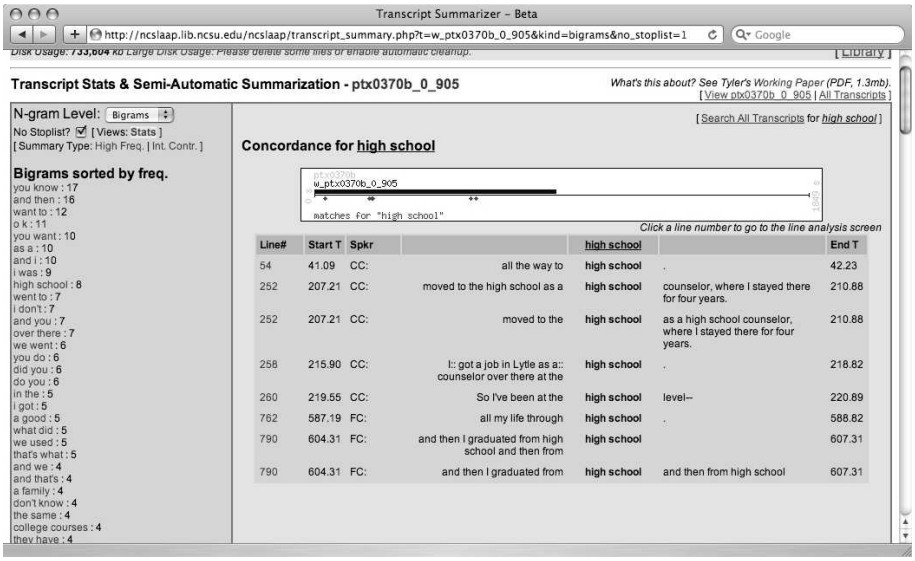

FIGURE 4 – A concordance view from SLAAP for "high school" in the transcript shown in Figures 2-3

The connection between the audio recording and the transcript (and other annotation layers) is not the only step available towards spoken language corpora that fit the "re-conceived" model of Figure 1, but it is, I believe, a large step towards improved spoken language data. Further, SLAAP's transcript

implementation has been shown here only as *one* demonstration of a way that this can be accomplished[12] and SLAAP, itself, is meant only as one possible example. The TalkBank website (<http://www.talkbank.org/>; MacWHINNEY, 2007) provides another excellent example of the advantages of time-aligned annotation linked to audio (and video) via specialized software, as does the Origins of New Zealand English (ONZE) project (<http://www.lacl.canterbury.ac.nz/onze/>; GORDON; MACLAGAN; HAY, 2007) and the growing list of projects using the ONZE Miner software (recently renamed as LaBB-CAT; <http://onzeminer.sourceforge.net/>; FROMONT; HAY, 2008). Finally, the Annotation Graph Toolkit (<http://agtk.sourceforge.net/>; BIRD; LIBERMAN, 2001) provides a formal framework for the development of these kinds of interfaces to data. Such systems, by basing the annotation on the temporal record of the recording, allow for multiple versions of annotation (and multiple versions even of transcription) and give the end-users, the analysts, the ability to customize their interfaces with the data.

The "re-conceived" model of abstraction for (socio)linguistic data in Figure 1 is perhaps less a proposal for the future than it is a way to think about and steer the changes that are occurring in the ways that audio-based spoken language recordings are manageable and increasingly managed. By focusing on building flexible annotation systems that maintain links through various levels of annotation and, most importantly, to the source recording, we can build corpora, which, instead of needing to be "tamed", can be utilized in a richer variety of ways than currently possible. I believe these sorts of models present the best opportunities for fruitful future work at the interface of corpus linguistics and sociolinguistics. They also would yield more flexible spoken language corpora for a range of applications beyond sociolinguistics.

[12] Other features in SLAAP also seek to minimize the separation of annotation from the source recording. For example, in addition to its transcript features, SLAAP has tools developed specifically for variationist sociolinguistic analysis that also follow a similar time-stamped and linked model. Analysts can extract and code variables (cf. WOLFRAM, 1993; TAGLIAMONTE, 2006) directly from the audio player or from the transcript views. These variable codes are stored along with their time-stamps and users can later return directly to the moment in the audio associated to each extracted variable at the click of the mouse. See Kendall (2007, 2008) for more on SLAAP's variable analysis features.

6. Conclusion

In this paper, I have outlined some areas where corpus linguistics and sociolinguistics have strong existing connections and some areas where these connections are less strong.[13] I have also discussed some wished for items for the future – namely, a broader range of "unconventional" corpora, which document a diverse range of language varieties, and an orientation to corpus-based data that maintains its connection to its context (and audio or video recording) and minimizes the amount of abstraction away from the actual source speech (or writing). These are advancements that I believe would greatly aid sociolinguistic research, as well as non-sociolinguistically oriented corpus-based research, and would build stronger bridges between sociolinguists and corpus linguists. The bulk of this paper has approached the relationship between sociolinguistics and corpus linguistics primarily from the perspective of sociolinguistics and, as such, has largely framed its discussion in terms of

[13] Further, I have focused in this paper on corpora, i.e. data, rather than other areas of intersection among corpus linguistic and sociolinguistic methods and practice. However, much could also be said about these other areas of overlap. For instance, both approaches involve extensive use of quantitative methods, although these exact methods differ in significant (but sometimes subtle) ways. Traditional corpus linguistic quantitative methods, in their focus on (normalized) frequencies of occurrence, can fail to account for what is *not* in a corpus. As D'Arcy (2005, in preparation) indicates through an analysis of discourse particle "like", corpus linguists might benefit from greater attention to variationist sociolinguistic quantitative methods (e.g., variable analysis and its principal of accountability; cf. LABOV, 1972b; TAGLIAMONTE, 2006), which attend not only to how many times the form of interest was realized by language users, but also to what *else* was realized in the places where that form was a relevant option. Meanwhile, much can also be said about Variable Rule Analysis (Varbrul), which for over three decades has been the dominant statistical technique in the sociolinguistic literature. Varbrul, a specialized form of logistic regression developed specifically for sociolinguistic variable analysis (CEDERGREN; SANKOFF, 1974) was a huge advancement over other available techniques for multivariate analysis when it was first developed, but in recent years, an array of powerful statistical techniques have been developed in corpus linguistics and other areas of language research (cf. BAAYEN, 2008; JOHNSON, K., 2008; GRIES, 2009; JOHNSON, D. E., 2009) that are, oftentimes, relevant and more appropriate for sociolinguistic analysis than Varbrul. This has been a point of contention among some language researchers in recent years, but, I believe, sociolinguists are rapidly incorporating these available techniques (see in particular JOHNSON, K., 2008, p. 174-180 and JOHNSON, D. E., 2009) and that this is becoming an area of fruitful, cooperative methodological advancement.

what corpus linguistics "can do" for sociolinguistic research. Yet, these suggestions have important ramifications on corpus linguistics more generally and I hope these ramifications are clear to readers: The development of more spoken language corpora, from a range of varieties and with more flexible annotation, will benefit corpus linguistic research widely.

As my discussion of Eckert's "three waves" account of the development of (variationist) sociolinguistic research indicates, it will likely be the case that much important sociolinguistic work remains heavily engaged in and devoted to a kind of analysis that is likely impossible through the use of corpora. Although, at the same time, as Baker (2010) points out, tools from corpus linguistics can still be used for examining transcribed data, regardless of the overall direction the research or data takes (provided it is transcribed, of course). Software-based archives, like that demonstrated by SLAAP above, can help bring corpus-based methods and a more explicit focus on data to sociolinguistic research, even that which is not interested in large-scale analysis.

I would like to end by posing the question: What can corpus linguists do *now* to best advance sociolinguistic research and to best promote the use of corpora and corpus methodologies in sociolinguistics? There are clearly several answers to this question and while others may respond differently, my own wish would be that corpus linguists (especially those who have extensive experience in corpus development) work directly with sociolinguists (especially those who focus on field-based research and ethnography) to develop sociolinguistically rich, "unconventional" corpora, to make those corpora publically available to researchers, and to work towards developing best-practices for the corpus-like treatment of *sociolinguistic* (spoken language) data. As I have argued elsewhere (KENDALL, 2008), sociolinguistic data and data management practices could greatly benefit from the knowledge and expertise of corpus linguists and language documentarians. Luckily, with the growth of projects like ONZE and LANCHART, and corpora like COLT and the LIC, I believe that we are on our way towards achieving this needed collaboration.

Acknowledgments

I thank the editors and the anonymous reviewers for excellent and helpful comments on an earlier draft of this paper. I also thank Gerard Van Herk, for many conversations relating to these topics, and Charlotte Vaughn, who helped with the original conception and production of Figure 1. Any errors, of course, remain my own.

References

ALLEN, W.; BEAL, J.; CORRIGAN, K.; MAGUIRE, W.; MOISL, H. A linguistic "time-capsule": The Newcastle Electronic Corpus of Tyneside English. In: BEAL, J.; CORRIGAN, K.; MOISL, H. (Ed.). *Creating and Digitizing Language Corpora*. New York / Basingstoke, Hampshire: Palgrave-Macmillan, 2007b. V. 2: Diachronic Databases.

ANDERSON, W. Corpus linguistics in the UK: Resources for sociolinguistic research. *Language and Linguistics Compass*, v. 2, n. 2, p. 352-371, 2008.

BAAYEN, R. H. *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press, 2008.

BAKER, P. *Sociolinguistics and Corpus Linguistics*. Edinburgh: Edinburgh University Press, 2010.

BAUER, L. Inferring variation and change from public corpora. In: CHAMBERS, J. K.; TRUDGILL, P.; SCHILLING-ESTES, N. (Ed.). *The Handbook of Language Variation and Change*. Malden, MA / Oxford: Blackwell, 2004.

BEAL, J.; CORRIGAN, K.; MOISL, H. (Ed.). *Creating and Digitizing Language Corpora*. New York / Basingstoke, Hampshire: Palgrave-Macmillan, 2007a. V. 1: Synchronic Databases.

BEAL, J.; CORRIGAN, K.; MOISL, H. (Ed.). *Creating and Digitizing Language Corpora*. New York / Basingstoke, Hampshire: Palgrave-Macmillan, 2007b. V. 2: Diachronic Databases.

BENDER, E. Corpus methods for sociolinguistics. Workshop at New Ways of Analyzing Variation (NWAV) 31. Palo Alto, CA: Stanford University, 2002. Available at: <http://faculty.washington.edu/ebender/corpora_sociolx.html>. Retrieved: April 1, 2011.

BIBER, D.; FINEGAN, E. Drift and the evolution of English style: A history of three genres. *Language*, v. 65, n. 3, p. 487-517, 1989.

BIRD, S.; LIBERMAN, M. A formal framework for linguistic annotation. *Speech Communication*, v. 33, n. 1-2, p. 23-60, 2001.

BOERSMA, P.; WEENINK, D. Praat: Doing phonetics by computer. 2010. Software. Available at: <http://www.fon.hum.uva.nl/praat/>. Retrieved: September 18, 2010.

BUCHOLTZ, M.; HALL, K. Gender, sexuality and language. In: BROWN, K. (Ed.). *Encyclopedia of Language and Linguistics*. 2. ed. Oxford: Elsevier, 2006. V. 4.

BURNARD, L. (Ed.). *Reference guide for the British National Corpus (XML edition)*. Published for the British National Corpus Consortium by Oxford University Computing Services, 2007. Available at: <http://www.natcorp.ox.ac.uk/docs/URG/>. Retrieved: April 1, 2011.

CEDERGREN, H.; SANKOFF, D. Variable rules: Performance as a statistical reflection of competence. *Language,* v. 50, n. 2, p. 333-355, 1974.

CHESHIRE, J. Sex and gender in variationist research. In: CHAMBERS, J. K.; TRUDGILL, P.; SCHILLING-ESTES, N. (Ed.). *The Handbook of Language Variation and Change*. Malden, MA / Oxford: Blackwell, 2004.

CHILDS, B.; VAN HERK, G.; THORBURN, J. Safe harbour: Ethics and accessibility in sociolinguistic corpus building. *Corpus Linguistics and Linguistic Theory*, v. 7, n. 1, p. 163-180, 2011.

COUPLAND, N. *Style*: Language variation and Identity. Cambridge: Cambridge University Press, 2007.

CRAIG, H.; WASHINGTON, J. *Malik Goes to School:* Examining the Language Skills of African American Students from Preschool to 5th Grade. Mahwah, NJ: Lawrence Erlbaum Associates, 2006.

D'ARCY, A. Tracking the development of discourse 'like' in contemporary (Canadian) English. Doctoral Thesis Proposal. University of Toronto, Toronto, Canada, March 16, 2005.

D'ARCY, A. Counting matters: Normalization and accountability. In preparation.

D'ARCY, A.; TAGLIAMONTE, S. Prestige, accommodation, and the legacy of relative *who*. *Language in Society*, v. 39, n. 3, p. 383-410, 2010.

DILLARD, J. L. *Black English:* Its History and Usage in the United States. New York: Random House, 1972.

ECKERT, P. Variation, convention, and social meaning. Paper presented at the 2005 Annual Meeting of the Linguistic Society of America, Oakland, CA, 2005.

ECKERT, P. Three waves of variation study: The emergence of meaning in the study variation. Under review. Available at: <http://www.stanford.edu/~eckert/PDF/ThreeWavesofVariation.pdf>. Retrieved: September 7, 2010.

FASOLD, R. *Tense Marking in Black English*: A Linguistic and Social Analysis. Washington, DC: Center for Applied Linguistics, 1972.

FRANCIS, W. N.; KUČERA, H. Brown Corpus manual. Revised and amplified, 1979. Available at: <http://icame.uib.no/brown/bcm.html>. Retrieved: September 7, 2010.

FROMONT, R.; HAY, J. ONZE Miner: The development of a browser-based research tool. *Corpora,* v. 3, p. 173-193, 2008.

GORDON, E.; MACLAGAN, M.; HAY, J. The ONZE Corpus. In: BEAL, J.; CORRIGAN, K.; MOISL, H. (Ed.). *Creating and Digitizing Language Corpora*. New York / Basingstoke, Hampshire: Palgrave-Macmillan, 2007b.

GREGERSEN, F. The data and design of the LANCHART study. *Acta Linguistica Hafniensia*, v. 41, p. 3-29, 2009.

GRIES, St. Th. Exploring variability within and between corpora: Some methodological considerations. *Corpora,* v. 1, n. 2, p. 109-151, 2006.

GRIES, St. Th. *Statistics for Linguistics with R:* A Practical Introduction. Berlin: De Gruyter Mouton, 2009.

HEATH, S. B. *Ways with Words*: Language, Life, and Work in Communities and Classrooms. New York: Cambridge University Press, 1983.

HYMES, D. *Foundations in Sociolinguistics*: An Ethnographic Approach. Philadelphia: University of Pennsylvania Press, 1974.

JOHNSON, D. E. Getting off the Goldvarb standard: Introducing Rbrul for mixed-effects variable rule analysis. *Language and Linguistics Compass,* v. 3, n. 1, p. 359-383, 2009.

JOHNSON, K. *Quantitative Methods in Linguistics*. Malden, MA / Oxford: Blackwell, 2008.

KENDALL, T. Enhancing sociolinguistic data collections: The North Carolina Sociolinguistic Archive and Analysis Project. *University of Pennsylvania Working Papers in Linguistics*, v. 13, n. 2, p. 15-26, 2007.

KENDALL, T. On the history and future of sociolinguistic data. *Language and Linguistics Compass*, v. 2, n. 2, p. 332-351, 2008.

KENDALL, T. Developing web interfaces to spoken language data collections. *Proceedings of the Chicago Colloquium on Digital Humanities and Computer Science*, v. 1, n. 2, Chicago: University of Chicago, 2010.

KENDALL, T.; BRESNAN, J.; VAN HERK, G. The dative alternation in African American English: Researching syntactic variation and change across sociolinguistic datasets. *Corpus Linguistics and Linguistic Theory*, forthcoming.

KENDALL, T.; VAN HERK, G. (Ed.). Corpus linguistics and sociolinguistic inquiry. *Corpus Linguistics and Linguistic Theory*, Special issue, v. 7, n. 1, 2011.

KENDALL, T.; WOLFRAM, W. Local and external standards in African American English. *Journal of English Linguistics*, v. 37, n. 4, p. 305-330, 2009.

KRETZSCHMAR, W. Jr. *The Linguistics of Speech*. Cambridge: Cambridge University Press, 2009.

KRETZSCHMAR, W. JR.; ANDERSON, J.; BEAL, J.; CORRIGAN, K.; OPAS-HÄNNINEN, L. L.; PLICHTA, B. Collaboration on Corpora for Regional and Social Analysis. *Journal of English Linguistics*, v. 34, n. 3, p. 172-205, 2006.

LABOV, W. The social motivation of a sound change. *Word*, v. 19, p. 273-309, 1963.

LABOV, W. *The Social Stratification of English in New York City*. Washington, D.C.: Center for Applied Linguistics, 1966.

LABOV, W. *Language in the Inner City:* Studies in the Black English Vernacular. Philadelphia: University of Pennsylvania Press, 1972a.

LABOV, W. *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press, 1972b.

LIDDICOAT, A. *An Introduction to Conversation Analysis*. London / New York: Continuum, 2007.

MACWHINNEY, B. The TalkBank Project. In: BEAL, J.; CORRIGAN, K.; MOISL, H. (Ed.). *Creating and Digitizing Language Corpora*. New York / Basingstoke, Hampshire: Palgrave-Macmillan, 2007a.

MALLINSON, C.; CHILDS, B. Communities of practice in sociolinguistic description: Analyzing language and identity practices among Black women in Appalachia. *Gender and Language*, v. 1, n. 2, p. 173-206, 2007.

MCDAVID, R.; MCDAVID, V. The relationship of the speech of American Negroes to the speech of whites. *American Speech*, v. 26, n. 1, p. 3-17, 1951.

MCENERY, T.; WILSON, A. *Corpus Linguistics*. 2. ed. Edinburgh: Edinburgh University Press, 2001.

MCENERY, T.; XIAO, R.; TONO, Y. *Corpus-based Language Studies*: An Advanced Resource Book. New York / London: Routledge, 2006.

MILROY, L.; GORDON, M. *Sociolinguistics*: Methods and Interpretation. Malden, MA / Oxford: Blackwell, 2003.

NEWMAN, J. Spoken corpora: Rationale and application. *Taiwan Journal of Linguistics*, v. 6, n. 2, p. 27-58, 2008.

OCHS, E. Transcription as theory. In: OCHS, E.; SCHIEFFELIN, B. (Ed.). *Developmental Pragmatics*. New York: Academic Press, 1979.

POPLACK, S. The care and handling of a mega-corpus: The Ottawa-Hull French Project. In: FASOLD, R.; SCHIFFRIN, D. (Ed.). *Language Change and Variation*. Amsterdam / Philadelphia: John Benjamins, 1989.

POPLACK, S. Foreward. In: BEAL, J.; CORRIGAN, K.; MOISL, H. (Ed.). *Creating and Digitizing Language Corpora*. New York / Basingstoke, Hampshire: Palgrave-Macmillan, 2007a.

POPLACK, S.; TAGLIAMONTE, S. *African American English in the Diaspora*. Malden, MA / Oxford: Blackwell, 2001.

RAYSON, P.; LEECH, G.; HODGES, M. Social differentiation in the use of English vocabulary: Some analyses of the conversational component of the British National Corpus. *International Journal of Corpus Linguistics*, v. 2, n. 1, p. 133-152, 1997.

RICKFORD, J. R. *African American Vernacular English*: Features, Evolution, Educational Implications. Malden, MA / Oxford: Blackwell, 1999.

ROMAINE, S. Corpus linguistics and sociolinguistics. In: LÜDELING, A.; KYTÖ, M. (Ed.). *Corpus Linguistics*: An International Handbook. Berlin / New York: Mouton de Gruyter, 2008.

SÄILY, T. Variation in morphological productivity in the BNC: Sociolinguistic and methodological considerations. *Corpus Linguistics and Linguistic Theory*, v. 7, n. 1, p. 119-141, 2011.

SÄILY, T.; SUOMELA, J. Comparing type counts: The case of women, men and *-ity* in early English letters. In: RENOUF, A.; KEHOE, A. (Ed.). *Corpus linguistics*: Refinements and reassessments. Amsterdam: Rodopi, 2009.

SANKOFF, D.; LABERGE, S. The linguistic market and the statistical explanation of variability. In: SANKOFF, D. (Ed.). *Linguistic Variation*: Models and Methods. New York: Academic Press, 1978.

SANKOFF, D.; SANKOFF, G. Sample survey methods and computer-assisted analysis in the study of grammatical variation. In: DARNELL, R. (Ed.). *Canadian Languages in their Social Context*. Edmonton, Alberta: Linguistic Research, 1973.

SCHMID, H-J. Do men and women really live in different cultures? Evidence from the BNC. In: WILSON, A.; RAYSON, P.; MCENERY, T. (Ed.). *Corpus Linguistics by the Lune*: A Festschrift for Geoffrey Leech. Frankfurt: Peter Lang, 2003.

SCHNEIDER, E. *Focus on the USA*. Amsterdam / Philadelphia: John Benjamins, 1996.

SHUY, R.; WOLFRAM, W.; RILEY, W. *Field Techniques in an Urban Language Study*. Washington, D.C.: Center for Applied Linguistics, 1968.

SIMONS, G.; BIRD, S.; SPANNE, J. (Ed.). Best practice recommendations for language resource description. Open Language Archives Community document, 2008. Available at: <http://www.language-archives.org/REC/bpr-20080711.html>. Retrieved: September 18, 2010.

SMITHERMAN, G. *Talkin and Testifying*: The Language of Black America. Boston: Houghton Mifflin, 1977.

SMITHERMAN, G. (Ed.). *Black English and the Education of Black Children and Youth*: Proceedings of the National Symposium on the King Decision. Detroit, MI: Center for Black Studies, Wayne State University, 1981.

STENSTRÖM, A-B.; ANDERSEN, G.; HASUND, I. K. *Trends in Teenage Talk*: Corpus Compilation, Analysis, and Findings. Amsterdam / Philadelphia: John Benjamins, 2002.

TAGLIAMONTE, S. *Analysing Sociolinguistic Variation*. Cambridge: Cambridge University Press, 2006.

TORGERSEN, E. N.; GABRIELATOS, C.; HOFFMANN, S.; FOX, S. A corpus-based study of pragmatic markers in London English. *Corpus Linguistics and Linguistic Theory*, v. 7, n. 1, p. 93-118, 2011.

WEINREICH, U.; LABOV, W. HERZOG, M. Empirical foundations for a theory of language change. In: LEHMANN, W. P.; MALKIEL, Y. (Ed.). *Directions for Historical Linguistics*. Austin, TX: University of Texas Press, 1968.

WENGER, E. *Communities of Practice*: Learning, Meaning, and Identity. Cambridge: Cambridge University Press, 1998.

WOLFRAM, W. *A Sociolinguistic Description of Detroit Negro Speech*. Washington, DC: Center for Applied Linguistics, 1969.

WOLFRAM, W. Identifying and interpreting variables. In: PRESTON, D. (Ed.). *American Dialect Research*. Amsterdam / Philadelphia: John Benjamins, 1993.

WOLFRAM, W.; THOMAS, E. R. *The Development of African American English*. Malden, MA / Oxford: Blackwell, 2002.

Recebido em 09/09/2010. Aprovado em 08/05/2011.