Prolegomenon for an analysis of dialect coding conventions for data sharing[1]

Malcah Yaeger-Dror and Christopher Cieri

Linguistic Data Consortium, University of Pennsylvania, University of Arizona

Introduction

For the past two decades, the Linguistic Data Consortium[2] at the University of Pennsylvania has fostered the intersection of linguistics and language technologies, building and sharing datasets for use in language related sub-disciplines. Over that time, we have observed a gradual convergence of desiderata and, to a lesser extent, methodologies. Speech technology developers who have been sharing data and relying upon publicly available data (Garofalo 1988; Godfrey, Holliman, & McDaniel 1992) have tended to focus more on variation in sensors and communication channels (Doddington 1985; Campbell 1997). Sociolinguistic inquiry has tended to focus on individual speech communities with only infrequent attempts to trace variation across regional or national space and therefore had less motivation for data sharing.[3] However, recent efforts in human language technologies (e.g. speaker and language recognition) have incorporated dialect information to improve system performance (Chen, Shen, Campbell, & Schwartz 2009), while linguists have begun to exploit publicly available corpora (Mackenzie 2013) and to share their own (Fox 2001; Kendall 2011).

Sharing across the boundaries of these disciplines has led to the recognition of potential advancements for all participants, such as the need to improve methodology in such surprising areas as sociolinguists' own coding of demo-

1 This work was conducted with the help of NSF Grant BCS #1144480 and supplemental funding from LDC. Particular thanks go to Christine Massey of the University of Pennsylvania Institute for Research in Cognitive Science, whose patience and clear-sightedness permitted this work to come to fruition. Thanks also to the many participants in the NSF sponsored workshop on this theme, http://projects.ldc.upenn.edu/NSF_Coding-_Workshop_LSA/index.html, and to Brittany McLaughlin and Laurel MacKenzie for their unflagging assistance. We would also like to thank the editors of this volume who provided feedback that hopefully permitted us to clarify our presentation.

2 http://www.ldc.upenn.edu

3 At present, the only ways to learn about many corpora are to ask the authors directly (which is neither a scalable procedure, nor one guaranteed to produce consistent results), or to read the published papers mentioning those corpora (which rarely have enough detail due to space constraints and the need to focus on research results).

graphic, attitudinal, and situational metadata.[4] This paper describes research that highlights the necessity for a deeper understanding and improved coding of the categories along which language varies. We suggest that more complete and appropriately granular metadata will facilitate more accurate analyses, comparison across studies, and better understanding of trends that span speech communities. Based on a review of recent publications and available corpora, we identify several discussion foci to facilitate the development of sharable databases, identify metadata which recent studies have demonstrated to be related to speech variation, and support the development of metadata protocols that lead to more thorough and comparable elicitation and analysis.

Changes in current metadata practice

Data sharing initiatives have developed proposals for capturing, normalizing, and sharing metadata that have necessarily evolved and yet still fall short of capturing the metadata needed in contemporary sociolinguistic studies. In the infancy of the British National Corpus[5], its creators maintained that one of their goals was to provide metadata that would permit archival comparisons and data sharing. Those metadata sufficed for that study's scope, but do not support all current needs for studying linguistic variation in the UK. The Open Language Archives Community[6] provides a Dublin Core[7] compliant standard for recording, hosting, and searching metadata, but it does not attempt to characterize individual speakers or sessions. The General Ontology for Language Description[8] (GOLD; Farrar & Langendoen 2003) is a central repository of very careful and thoughtful definitions of terminology for particular linguistic phenomena, but it does not consider social factors.

Sociolinguists assume that their own corpora have followed a shared protocol for data collection with relevant coding to permit judicious comparisons (Labov 1984; Poplack 1989; Tagliamonte 2012). In actuality, conventions are often not shared and not explicitly stated, with the result that corpora which could and should be comparable are not. There are exceptions, particularly

4 We use the terms *coding* and *metadata* nearly interchangeably because they are roughly so in the current scope. Human Language Technology developers tend to use the term *annotation* for judgments on the spoken, written, signed, or gestured data, and *metadata* for information about speakers, attitudes, and situations. Sociolinguists *code* all of these.

5 http://www.natcorp.ox.ac.uk/corpus

6 http://www.language-archives.org/

7 http://dublincore.org

8 http://linguistics-ontology.org

among the research groups in Toronto, Ottawa, and Montreal, where corpora follow the explicit conventions first discussed in Sankoff and Sankoff (1973) and further elaborated when the Montreal French corpus expansion required a set of rules to permit comparability with the earlier collection (Thibault & Vincent 1990). Subsequent studies in Ottawa (Poplack 1989), Denmark (Gregersen 2009), and Toronto (Tagliamonte 2012) have followed comparable protocols. Overall, however, corpora are not reliably comparable and are rarely shared.

The US National Science Foundation (NSF) has enacted new rules which require that proposals describe how data will be maintained and shared across fairly broad research communities.[9] A more recent memo from the White House Office on Science and Technology Policy indicates that researchers will experience increasing pressure to share their research products effectively, including academic papers and corpora.[10] The members of the Variationist List have also been discussing how best to gather and code corpora to permit subsequent sharing of data from different regions and from different communities[11].

Of course, it is inevitable that later papers would refine the metadata categories suggested in the seminal works on language variation and change. Moreover, when many of these works were carried out, it was inconceivable that we would be sharing data the way we can today. In consequence, the research sponsorship of the time lacked the enthusiastic and growing support for data sharing we see today. It is natural that the advent of easy data sharing brings with it the need to revise methodology.

Coding demographics and attitude

The focus of this paper is a treatment of the pre-requisite issue of coding specific demographic information shown to influence speech and speakers' attitudes toward their own and other groups. As such, this paper can be considered a prolegomenon to a study of metadata requirements and will provide suggestions for finer-grained demographic coding for more accurate analyses of speech variation directly and via the attitudes those demographic factors engender.

Community affiliation

Each speaker belongs to at least one local or regional community (Labov 2001), and most studies control for community affiliation. However, many studies em-

9 http://www.nsf.gov/publications/pub_summ.jsp?ods_key=grantsgovguide

10 http://www.whitehouse.gov/sites/default/files/microsites/ostp

11 http://j.mp/VAR-L

ploy an under-differentiated coding scheme for both speaker's dialect and regional heritage within a given language community.

Regional identity versus political boundaries

Many studies code regional identity as the nation or state in which speakers were raised, though ample evidence shows that dialect boundaries do not always conform to political ones. In the US, New York (e.g. Dinkin 2009), Ohio (e.g. Thomas 2011), and Illinois (e.g. Habick 1991; Bigham 2010), to name only three states, exhibit extreme dialect variation.

In coding for language or dialect heritage, correlations to political or regional boundaries become even more tenuous. Numerous studies have shown differences among first, second, and later generation immigrants; other studies have drawn connections among the multiple sub-communities with which speakers identify and the linguistic variables associated with those groups (e.g. Hall-Lew & Starr 2010; Llamas & Watt 2010; Bayley, Cárdenas, Treviño Schouten, & Martin Vélez 2012; Hall-Lew & Yaeger-Dror 2012). This work underscores the necessity of distinguishing factors that are most critical to a given community's identity from those that are more friable into disparate identities, each of which evokes attitudes about speech and sub-communities of speakers. The importance of each of these factors has been extensively discussed in the work of Labov (most recently, 2011), Preston (e.g. Preston & Long 2002) and Giles (Giles 1973; Bourhis & Giles 1977), and their students. Beyond such regional identities, other community identities have also been shown to influence language variation and linguistic attitudes.

Ethnicity: linguistic, regional, religious or 'racial' heritage

Ethnicity has been frequently used as a cover term for linguistic heritage (e.g. Italian, Polish, Cantonese, Latino)[12], regional heritage (e.g. Northern, Southern), 'racial' heritage (e.g. African American, Asian, Native American) [13], and religious heritage (e.g. Jewish, Christian, Muslim). The question facing the researcher who would compare such studies is not merely with which 'ethnic' groups a given speaker associates, but whether all potential personal identities are coded consistently to permit comparison regarding the possible strength of competing

12 Many studies inaccurately treat linguistic and national heritage as identical; see discussion in Verkuyten and Thijs (2010).

13 Any discussion of the validity of the concept of 'race' is beyond the scope of this paper. We use the term merely to refer to its common use as a very broad categorization.

influences on group members' speech. Even if we consider only these four categories, it is clear that speakers whose parents come from identical backgrounds may still differ in how they identify with different groups. The comparative strength of those linguistic identities cannot be measured in research that acknowledges only partially overlapping subsets. The problem arises most clearly in discussions of 'ethnicity' if non-congruent criteria are chosen by different studies which one then hopes to compare or archive together. In addition, speakers' sense of affiliation with any of these groups is often correlated with the interlocutors and the relationships among them, the topic discussed, and how well the interaction itself is progressing (e.g. Giles 1973; Bourhis & Giles 1977; Eckert & Rickford 2001; Labov 2001; Rickford & Price 2013).

Communities: self-defined versus externally-defined

Even with language choice fixed, sub-groups (e.g. Northern vs. Southern Italian, Indian, French, Mexican, Norwegian) hold specific attitudes toward those from other areas within the same heritage language, even if the speakers appear to outsiders to share the same cultural heritage. The situation becomes more complicated as linguistic, 'racial', or religious variables come into play. While many American communities have large Polish populations, the Catholic and Jewish Polish heritage speakers may not consider themselves as sharing an ethnic or community affiliation. Similarly, Iranians, Syrians, Iraqis, or Lebanese speakers from Copt, Maronite, Circassian, or other Christian heritage may have a very limited sense of shared community heritage with their Sunni or Shia neighbors or with each other (see, for example, Blanc 1964). Conversely, consider those who share a religion but differ in their regional heritage. The degree to which they wish to emphasize regional linguistic features or specific religious shibboleths may vary by individual or situation. Attitudes toward the in- and out-groups can all influence speech, making coding each such feature prudent.

Bailey (2000), Toribio (2003), Guardado (2008), and others have discussed the fact that 'Latino' is not one uniform ethnicity. It is 'a heterogeneous US ethnic group' (Bayley et al. 2012) which varies radically with both racial and regional heritage communities, the speakers' socioeconomic background, and the regions from which and to which speakers' families migrate (which in turn influences the degree to which they are integrated into a larger local Hispanic or Anglo community, or are isolated from other communities).[14] Alim, Ibrahim,

14 Looking to larger sociological studies may not help either. Both the US Census and Pew Research Center specify that 'Latino' is a term used for people of any Spanish speaking

and Pennycook (2008), Cutler (2010), Roth-Gordon (2012), and others have demonstrated the extent to which we oversimplify speakers' ethnic identity if we code 'Black' speakers of English as only 'African American'. Wong and Hall-Lew have pointed out the extent to which 'Asian' is neither racially nor culturally a valuable cover term. They have also shown that even limiting the discussion of 'Asian Heritage' speakers to Cantonese speakers, their affiliation with this heritage and their choice of sociophonetic features to flag that identity may differ in New York and San Francisco, and may be influenced as well by other demographic variables (e.g. Hall-Lew & Star 2010; Wong & Hall-Lew 2012).[15]

When religion is coded, it is sometimes relegated to 'ethnicity' (Tannen 1981; Meechan 1999) for lack of a generally accepted set of conventions for coding religious affiliation. Since religion is coded in demographic studies, we have been able to determine the degree to which it should be regarded as a factor group quite distinct from 'race' or regional heritage. On the other hand, if religion is coded under 'ethnicity', other aspects of 'ethnicity' discussed above are ignored and cannot be retrieved later. Just as speakers from the same heritage language but of different religions may feel they share no community of interest, members of the same religion from different linguistic, regional, or racial backgrounds may share few religious community ties (e.g. Miller 2007; Walters 2007; Ben Rafael & Sharot 2008; Wagner 2013).

These findings argue that each of these 'ethnic' designation types should be treated individually rather than lumped together as ethnicity. We can ignore neither religious distinctions nor other in-group membership distinctions with which they interact. The full extent to which group memberships influence speech will only be determined when studies are coded consistently for each of these features. We further propose that while all these factors are not critical in every study, consistently coding a larger inventory of factors increases opportunities to uncover correlations and interactions. In this way, we may determine that some factors are more important than initially assumed. We may also be able to demonstrate that in a given community a factor generally assumed to be critical to linguistic variation is not (e.g. Hinton & Pollock 2000).

region, whatever their racial or religious background, yet within each Hispanic region, narrower regional, religious, and racial identities are often quite salient.

15 Similarly, Sharma (2011) shows the extent to which British 'Asian' speakers' linguistic choices have been highly correlated with the ethnolinguistic vitality of the community that they are members of, and Bayley and Bonnici (2009) suggest that linguistic changes over the last generation are linked to the ethnolinguistic vitality of the US Latino communities to which speakers belong; see discussion below.

Multiple heritage

Many sociolinguists now agree that the interview protocol should permit speakers to acknowledge multiple heritage group membership. We must code variables conflated under the 'ethnicity' umbrella, allowing for the possibility that a speaker may have more than one identity for each, and acknowledging that sense of identity will be dependent on other factors, such as the identities of their interlocutors and other situational factors (Eckert & Rickford 2001).

Ideological commitment and degree of affiliation

Degree of religious observance has also been shown to influence adaptation to a local community dialect (e.g. Mallinson & Childs 2007, inter alia; Al-Khatib, Alzoubi, & Abdulaziz 2009; Baker & Bowie 2009). Though open-ended questioning can reveal interviewees' positions, we see evidence only infrequently of a uniform plan to elicit and encode such information. Without such a plan the researcher hoping to compare the impact of the nature and strength of religious affiliation must either abandon corpora that lack this information or attempt to code it with recourse only to the static recordings.

Mallinson and Childs (2007) showed that even in a single religious community, religious and nonreligious friendship groups differ consistently in their sociophonetics. Comparison of degree of religious observance of Alberta (Meechan 1999) and Utah (Baker & Bowie 2009) Mormons, or different groups of Jews (Ben Rafael & Sharot 2008) reveals that degree of religious observance can be as significant an influence on indexicality as that of racial or regional heritage. As sociolinguists we may assume that the differences are caused primarily by social networks (Milroy 1980; Mallinson & Childs 2007), but social attitudes toward other groups often differ, or are intensified, by speakers' degree of religious (Benor 2011; Alam & Stuart-Smith 2013) or political (Abu-Elhij'a 2011; Bourhis et al 2009; Hall-Lew et al 2012) commitment. Some of the above-cited studies present evidence to support their claim that it is group ideology rather than community of practice, or social network, which is the strongest influence on the attitudes toward appropriate speech behavior.

Language, attitudes, and evaluation

Labov (1972, 2001) and others have strengthened our understanding of the importance of education and class, even in the 'classless' New World. However, often when data are archived only the most skeletal remains of educational or

other socioeconomic coding are retained. Rough assessment of speakers' educational achievements are generally found somewhere in the transcript even if they have not been coded, but while earlier studies were prone to follow a rather rigid *socioeconomic scale* (SES) or more nuanced scales like the *Linguistic Marketplace* (ML) rating developed in Montreal (Sankoff & Laberge 1978), recent studies may not even code for how all speakers make a living, much less their SES or ML. Such factors influence attitudes toward interlocutors (e.g. Fabricius & Mortensen 2013).

As a rule, we do not think of speakers' politics as creating a community social network in the same way as their religious beliefs may. On the other hand, publications by social psychologists, descriptive linguists, and sociolinguists have found that political opinions influence attitudes towards linguistic and other cultural assimilation, and consequently speech—particularly for members of heritage language communities (Dubois & Horvath 2000; Bourhis, Barrette, El-Geledi, & Schmidt 2009). This is the case even within a uniform dominant community and particularly when the politics form part of a speaker's public persona (e.g. Hall-Lew, Starr, & Coppock 2012; Hernández Campoy & Cutillas Espinosa 2012). This should come as no surprise, when attitudes toward 'urbanity' (Labov 2001; Habick 1991; Hall-Lew et al. 2012; Miller 2007) are understood to be correlated with group members' politics, and can also influence sociophonetic variation.

External factors also influence speakers' attitudes. Giles and Johnson (1987) specify a range of external demographic factors that influence a community's ethnolinguistic vitality. It is understood that ethnolin-guistic vitality influences speaker's attitudes toward when to speak the dialect [or language] of their in-group. Speakers of the same ethnic background(s), but who are embedded in different community environments, actually ought to be compared with each other very carefully because the influence of external factors within their respective communities impinge radically on the speakers' attitudes toward themselves, their peers, and the out-groups with whom they share a larger community. Speakers' ability and willingness to return to their heritage area and ability to access their heritage language locally may strongly influence their willingness to acknowledge this heritage or speak the language.

Dubois and Horvath (2000) was perhaps the first sociolinguistic article to demonstrate the extent to which political attitudes of speakers in different times can favor radically different linguistic outcomes, based on a renaissance of positive affect toward cultural artifacts and access to them in the public media. Labov's (2001) discussion of changing social dynamics on Martha's Vineyard also demonstrates the extent to which changing external factors influence a community's set of attitudes toward themselves and others, radically changing the dia-

lect from one generation and the next. Bayley and Bonnici (2009) point out that one factor that distinguishes Latinos of the youngest contemporary generation from those of previous generations is that they have greater access to online and media outlets in Spanish, as well as greater opportunities to return periodically to their heritage-land(s). Such factors influence attitudes toward the maintenance of language and culture (Giles 1973). Ito (2010) has come to similar conclusions in her studies of US immigrants. Sharma (2011) has pointed out that the experience of the older generation of Bangladeshi immigrants to the UK, who were a small minority in their London neighborhoods and who could not get along in either their L1 or their dialect of English, should not be compared to the experience of those who arrive to neighborhoods where they are the demographic majority. Similarly, Wong and Hall-Lew (2012) have pointed out the incomparability of the 19th and 20th century 'New Chinese Immigrants' (who arrived as 'coolies' for the railroads with no education or educational opportunities, no access to Chinese media, books, or opportunity to visit China) with recent high tech arrivals. To label both 'first generation', without taking into consideration the group's linguistic vitality, as well as individuals' educational and socioeconomic situations, is disingenuous.

Since these demographic factors all influence listener attitudes, which in turn influence linguistic variation (Labov 1964, 1972b, 2001; Fabricius & Mortensen 2013; Wagner 2013), future research protocols should take advantage of recent analyses which have demonstrated that a finer-grained coding protocol is called for. Each of these factors is an important element in speakers' image of themselves. Often the information has already been recorded as part of the interview itself, and coding it into a metadata file consistently will save hours of repetitious hunting. In short, it is clear that revising data gathering and coding protocols to permit us to capture information which is easily included will improve the research uses of resulting corpora, both for our own studies and as archived data to be available for future studies.

Conclusions

While many studies take the path of least resistance and incorporate metadata from the US Census and the like, sociolinguists have concluded that such metadata are too coarse-grained for appropriate coding for our needs. Many (if not all) of the factors discussed above have been shown to be sociolinguistically sensitive; evidence has demonstrated that each is correlated with linguistic choices in some communities.

Consistent elicitation protocols are particularly important when we are considering subsequent archival storage, since it has been demonstrated that different ways of posing a given question will elicit different responses (Bowie 2012). The demographic information should be elicited in a comparable way; the features should be coded and stored consistently to permit archiving and future study.

What is yet to be achieved are shared resources for elicitation and coding of speakers' social attitudes toward both their own group, and out-groups within the local community. It is also critical to develop resources for the analysis and coding of speaker attitudes engendered within a given interaction, to permit quantification of the relationship between a given set of attitudes and speakers' use of a given linguistic variable within a given interactive setting. Speakers' multiple identities should also be recognized for their (sometimes inconsistent or divergent) influences on attitudes, which influence speech both directly and independently. Our metadata should permit the analysis of possibly relevant factors for all interlocutors in a given interaction. Granted, an attempt to consider every possible factor simultaneously may initially convince researchers that the task is unwieldy. However, given our improved ability to consider features that have proven relevant in previous studies, and given that what has not been coded is much more difficult to incorporate later, we propose that taking a more comprehensive and circumspect view of metadata is not only feasible, but actually more effective in the long run.

References

Abu-Elhij'a, Dua'a 2011. *Variation in representation of Arabic consonants and grammatical variables in Facebook*. Unpublished MA thesis, University of Haifa.

Al-Khatib, Mahmoud, A. Alzoubi, and A. Abdulaziz. 2009. The impact of sect-affiliation on dialect and cultural maintenance among the Druze of Jordan: An exploratory study. *Glossa* 4: 186-219.

Al-Wer, Enam and Rudolf de Jong (eds.). 2009. *Arabic dialectology: In honour of Clive Holes on the occasion of his sixtieth birthday*. Leiden: Brill.

Alam, Farhana and Jane Stuart-Smith. 2013. Identity, ethnicity, and fine phonetic detail: An acoustic phonetic analysis of syllable-initial /t/ in Glaswegian girls of Pakistani heritage. In M. Hundt & D. Sharma (eds.), *English in the Indian diaspora*. Amsterdam: Benjamins. PAGES?

Alim, Samy, A. Ibrahim, and A. Pennycook (eds.). 2008. *Global linguistic flows: Hip Hop cultures, youth identities and the politics of lg*. NY: Routledge.

Bailey, Benjamin. 2000. Language and negotiation of ethnic/racial identity among Dominican Americans. *Language in Society* 29: 555-582.

Baker, W. and David Bowie. 2009. Religious affiliation as a correlate of linguistic behavior. *PWPL* 15 (Article 2) URL: <repository.upenn.edu/pwpl>.

Bayley, Robert and Lisa Bonnici. 2009. Recent research on Latinos in the USA and Canada, Part 1: English varieties. *Language and Linguistic Compass* 3: 1300-1313.

Bayley, R., N. Cárdenas, B. Treviño Schouten, C. Martin Vélez 2012. Spanish dialect contact in San Antonio, Texas. In K. Geeslin & M. Días-Campos (eds.), *Selected proceedings of the 14th Hispanic Linguistic Symposium*. Somerville, MA: Cascadilla Proceedings Project. 48-60. URL: http://-www.lingref.com/cpp/hls/14/index.html.

Ben Rafael, Eliezer and S. Sharot. 2008. *Ethnicity, religion and class in Israeli society*. Cambridge: CUP.

Benor, Sarah (ed.). 2011. Special issue of *Language and Communications* 31/2.

Bigham, Doug. 2010. Mechanisms of accommodation among emerging adults in a university setting. *Journal of English Linguistics* 38: 193-210.

Blanc, Haim. 1964. *Communal dialects in Baghdad*. Cambridge, MA: Harvard University Press.

Bourhis, R. Y. and Howard Giles. 1977. The language of intergroup distinctiveness. In H. Giles (ed.), *Language, ethnicity and intergroup relations*. London: Academic Press. 119-135.

Bourhis, Richard, G. Barrette, S. El-Geledi, and R. Schmidt. 2009. Acculturation orientations and social relations between immigrants and host community members in California. *Journal of Cross-Cultural Psychology* 40: 443-467.

Bowie, David. 2012. Age as a sociolinguistic variable. Presented at the NSF Workshop on Linguistic Archives – URL: http://projects.ldc.upenn.edu/-NSF_Coding_Workshop_LSA/index.html.

Campbell, J. 1997. Speaker recognition: A tutorial. *Proceedings of the IEEE* 85/ 9: 1437-1462.

Chen, Nancy, Wade Shen, Joseph Campbell, and Reva Schwartz. 2009. Large-scale analysis of formant frequency estimation variability in conversational telephone speech. *Proceedings of Interspeech 2009*, Brighton, UK. 2203-2206.

Cheshire, Jenny, Sue Fox, Paul Kerswill, and Eivind Torgersen. 2008. Ethnicity, friendship network and social practices as the motor of dialect change: Linguistic innovation in London. In U. Ammon and K. Mattheier (eds.), *Sociolinguistica: International Yearbook of European Sociolinguistics*. Berlin: Max Niemeyer Verlag. 22, 1-23.

Cutler, Cecelia. 2010. Hip-hop, white immigrant youth, and African American Vernacular English: Accommodation as an identity choice. *Journal of English Linguistics* 38: 248-269.

Dinkin, A. 2009. Dialect boundaries and phonological change in upstate New York. Doctoral dissertation, University of Pennsylvania.

Doddington, G.R. 1985. Speaker recognition—Identifying people by their voices. *Proceedings of the IEEE 73/11*. 1651-1664.

Dubois, Sylvie and Barbara Horvath. 2000. When the music changes, you change too. *Language Variation and Change* 11: 287-313.

Eckert, Penelope 2000. *Linguistic variation as social practice: The linguistic construction of identity at Belten High*. Oxford: Blackwell Publishers.

Eckert, Penelope 2008. Variation and the indexical field. *Journal of Sociolinguistics* 12: 453-476.

Eckert, Penelope and John Rickford (eds.). 2001. *Style and sociolinguistic variation*. Cambridge: Cambridge University Press.

Fabricius, A. and J. Mortensen. 2013. Language ideology and the 'construct resource': A case study of modern RP. In T. Kristiansen & S. Grondelaars, (eds.), *Language (de)standardisation in Late Modern Europe*. Oslo: Novus Forlag.

Farrar, S. and D.T. Langendoen. 2003. A linguistic ontology for the semantic web. *GLOT International* 7: 97-100.

Fox, Michelle A. 2001. Syllable-Final /s/ Lenition. Linguistic Data Consortium, Philadelphia.

Garofolo, John S. 1988. *Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database*. National Institute of Standards and Technology (NIST), Gaithersburgh, MD 107.

Giles, Howard. 1973. Accent mobility: A model and some data. *Anthropological Linguistics* 15: 87-105.

Giles, H. and P. Johnson. 1987. Ethnolinguistic identity theory. *International Journal of the Sociology of Language* 68: 69-99.

Godfrey, J., E. Holliman, and J. McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development, *ICASSP-92, IEEE International Conference on Acoustics, Speech, and Signal Processing* 1: 517-520.

Guardado, José Martin. 2008. Language socialization in Canadian Hispanic communities: Ideologies and practices. Doctoral dissertation, University of British Columbia.

Habick, Timothy 1991. Burnouts versus Rednecks: Effects of group membership on the phonemic system. In P. Eckert (ed.), *New ways of analyzing sound change*. San Diego: Academic Press. 185-212.

Hall-Lew, Lauren and Rebecca Starr. 2010. Beyond the second generation: English use among Chinese Americans in the San Francisco Bay Area. *English Today* 26/3: 12-19.

Hall-Lew, Lauren, Rebecca Starr, and Elizabeth Coppock. 2012. Style-shifting in the U.S. Congress: The vowels of 'Iraq(i)'. In J.M. Hernandez Campoy & J.A. Cutillas Espinosa (eds.), *Style-shifting in public: New perspectives on stylistic variation*. Amsterdam: John Benjamins. 45-63.

Hall-Lew, Lauren and Malcah Yaeger-Dror (eds.). 2012. Evolving perspectives on the concept of ethnolect, LSA, Portland. [To appear as double issue of *Language and Communication* 2014.]

Hay, Jennifer, Paul Warren, and Katie Drager. 2006. Factors influencing speech perception in the context of a merger-in-progress. *Journal of Phonetics* 34: 458-484.

Hernandez Campoy, Juan Manuel, and J.A. Cutillas Espinosa (eds.). 2012. *Style-shifting in public: New perspectives on stylistic variation*. Amsterdam: John Benjamins.

Hinton, Linette and Karen Pollock. 2000. Regional variations in the phonological characteristics of African American Vernacular English. *World Englishes* 19: 59-71.

Ito, Rika 2010. Accommodation to the local majority norm by Hmong Americans in the twin cities. *American Speech* 85: 141-162.

Kendall, Tyler. 2011. Corpora from a sociolinguistic perspective. *Brazilian Journal of Applied Linguistics* 11: 361-389. URL: http://ncslaap.lib.ncsu.-edu/pdfs-/Kendall2011_BJAL_CorpSocioling.pdf.

William A. Kretzschmar, et al. 2012. Digital Archive of Southern Speech. Linguistic Data Consortium, Philadelphia.

Labov, William. 1984. Field methods for the project on language change and variation. In J. Baugh & J.Sherzer (eds.), *Language in use*. Englewood: Prentice Hall. 28-53.

Labov, William. 2001. *Principles of linguistic change: Social factors*. Oxford: Blackwell Publishing.

Llamas, Carmen and Dom Watt (eds.). 2010. *Language and identities*. Edinburgh: Edinburgh University Press.

MacKenzie, Laurel. 2013. Variation in English auxiliary realization: A new take on contraction. *Language Variation and Change* 25: 17-41.

Mallinson, Christine and Becky Childs. 2007. Communities of practice in sociolinguistic description: Analyzing language and identity practices among Black women in Appalachia. *Gender and Language* 1: 173-206.

Meechan, Marjory 1999. The Mormon drawl: Religious ethnicity and phonological variation in southern Alberta. Doctoral dissertation, University of Ottawa.

Miller, Catherine 2007. *Arabic in the city: Issues in dialect contact and language variation*. NY: Routledge.

Milroy, Lesley. 1980. *Language and social networks*. Oxford: Blackwell.

Poplack, Shana 1989. The care and handling of a megacorpus. In R. Fasold & D. Schiffrin (eds.), *Language change and variation*. Amsterdam: John Benjamins. 411-451.

Preston, Dennis and Daniel Long (eds.). 2002. *Handbook of perceptual dialectology*. Amsterdam: Benjamins.

Rickford, John and Mary Price. 2013. Girlz II Women: Age-grading, language change and stylistic variation. *Journal of Sociolinguistics* 17: 143-179.

Roth-Gordon, Jennifer 2012. Linguistic techniques of the self. *Language and Communication* 32: 36-47.

Sankoff, D. and S. Laberge. 1978. The linguistic market and the statistical explanation of variability. In D. Sankoff (ed.), *Linguistic variation: Models and methods*. New York: Academic Press. 239-250.

Sharma, Devyani. 2011. Style repertoire and social change in British Asian English. *Journal of Sociolinguistics* 15: 464-492.

Tagliamonte, Sali A. 2012. *Variationist sociolinguistics: Change, observation, interpretation*. New York: Wiley-Blackwell Publishers.

Tannen, Deborah. 1981. New York Jewish conversational style. *International Journal of the Sociology of Language* 30: 133-149.

Thibault, Pierrette and Diane Vincent. 1990. Un corpus de français parlé. Montreal: Recherches Sociolinguistiques, 1.

Thomas, Erik. 2010. A longitudinal analysis of the durability of the Northern/Midland dialect boundary in Ohio. *American Speech* 85: 375-430.

Toribio, Almeida Jacqueline. 2003. The social significance of language loyalty among Black and White Dominicans in New York. *The Bilingual Review/La Revista Bilingüe* 27: 3-11.

US Census. 2013. http://factfinder2.census.gov/faces/nav/jsf/pages/search-results.xhtml?refresh=t.

Wagner, Suzanne Evans. 2013. We act like girls and we don't act like men: Ethnicity and local language change in a Philadelphia high school. *Language in Society* 42: PAGES.

Walters, Keith. 2007. Language attitudes. In K. Versteegh et al. (eds.), *Encyclopedia of Arabic language and linguistics*. Vol. II. Leiden: Brill. 650-664.

Wang, Wendy. 2012. The rise of intermarriage: Rates, characteristics vary by race and gender. Washington: Pew Social and Demographic Trends. http://www.pewsocialtrends.org/files/2012/02/SDT-IntermarriageII.pdf

Wong, A,y and Lauren Hall-Lew. 2012. Coding for Asian (American) ethnic identities. Paper delivered at the NSF sponsored workshop at LSA in Portland. *Creating and digitizing language corpora*.