SPi	Journal Code				Article ID				Dispatch: 06.11.14	CE: Diane E. Petersen
	L	N	C	3	1	2	1	2	0	No. of Pages: 7	ME:

Language and Linguistics Compass (2014): 1–7, 10.1111/lnc3.12120

Introduction to the Special Issue on Archiving Sociolinguistic Data

Malcah Yaeger-Dror* and Christopher Cieri Q1
Department of LinguisticsUniversity of Arizona

Abstract
▪ Q2

1. *Background*

The growth in computing power, storage, and networking has encouraged and enabled an increasing interest in big data, which one might define as volumes beyond what any individual researcher can collect and analyze directly. In linguistics, this naturally leads to increased emphasis on data sharing, which is one way to amass data of sufficient size and variety to permit new kinds of insight. At the same time, there is a growing consensus among sociolinguists that researchers must refine their views on demographic, situational, and attitudinal factors that correlate with linguistic differences in order to permit not only more effective data sharing and comparison but also more accurate conclusions.

2. *History & Purpose*

The roots of the current volume go back to the editors' first correspondences in 1998 concerning corpora that might be useful for sociolinguistic research, their formats, and the metadata and annotations they contain. Over the intervening years, we have grappled with the problem of how corpora initially created to support human language technology (HLT) development might be re-used or augmented for linguistic analysis, conversely how data collected by linguists might help in building and improving HLTs and (equally important) how the integration of HLTs with our own corpora might improve linguistic analysis by reducing manual labor and allowing analysts to address new questions using larger volumes of data with additional layers of annotation. One of the first challenges we encountered dealt with mismatches in metadata. Among corpora focusing on the correlation of linguistic and social phenomena, one finds great variation even among demographic categories represented in the literature for decades. Social situation is even less likely to be coded systematically, and there is almost no consensus on how or when to code for language attitudes (Llamas/Watt, Noels, Q3 Rickford, this issue). Yet, there is a clear consensus that each should be considered independently of other measurable features. Although many of the contributors to this special issue have previously discussed data sharing via a series of workshops, classes, and papers, we concluded that a new workshop focused specifically on the question of metadata interoperability would permit researchers to discuss concerns that have arisen in their own studies and propose appropriate methods and metadata for comparable data sets. The plan was for each invited speaker to suggest a straightforward set of distinctions, which are critical for analysis of the variables they discussed. It is indicative of the complexity of the problems the field faces that

many of these 'simple' variables need further study before a definitive list can be proposed. We also uncovered other factors in corpus development, which should be foci for specific papers.

The goal of this special issue then is to publish papers from the workshop on Sociolinguistic Archive Preparation presented at the Linguistic Society of America in January of 2012, with funding from NSF (BCS #1144480). The workshop's goal was to discuss and hopefully agree upon techniques for developing corpora that can be shared effectively among research groups and related disciplines. Presenters identified numerous instances in which they gained new insights by eliciting metadata not commonly considered in the field or by refining existing metadata categories. Although every speech community, social network, community of practice or other domain of focus may display unique features, the workshop confirmed the need for linguists to carefully consider the demographic, situational, and attitudinal dimensions of the interactions they observe as well as the coverage and granularity of the metadata they elicit. Individual papers provided key examples that inspire the reader to take a fresh and critical look at metadata. In doing so, the workshop laid the foundation for the systematic development of future corpora and archives that will permit analyses not possible today. As will be obvious from the table of contents, we have selected specific metadata areas in which our contributors have expertise and avoided others for which there is insufficient consensus while hoping that they too will become foci of future discussions.

3. *Organization*

3.1. The papers in this special issue relate to multiple themes

The first theme relates to the problem of identifying appropriate metadata for a given study in the context of data sharing. The point of holding a workshop before the publication of these papers was to permit authors to reach consensus on how to identify and analyze specific demographic, situational, and attitudinal features in order to capture and preserve critical information, which is only determinable at the time of interviews, and which researchers have found to influence speech they have analyzed. Each of the contributors has carried out extensive data collection and archiving projects: Gary Simons at SIL International and the Open Language Archives Community where his work has advanced the state of language documentation, Christopher Cieri at the Linguistic Data Consortium where he has overseen the development of hundreds of datasets and more than 100,000 distributions and Tyler Kendall at NCSU and Oregon where he has developed sociolinguistic archives and advised numerous others. Each reports on insights they have gleaned from their own efforts and suggests how these can be applied to future corpora.

At least within the US, studies involving human subjects must submit a protocol for approval by an *Institutional Review Board* (IRB), itself subject to Federal regulation, before work can begin. Q4 Ironically, the administrative requirements of the IRB force researchers to consider, at the outset, the very questions that the science demands but that could otherwise be postponed: What is the goal of the study? What data, metadata, and annotation will be required to achieve that goal? How exactly will the study acquire them? At the same time, gaining approval for a protocol can be challenging as researchers, expert in their own domains, seek to explain methods, risks, and benefits to board members whose training is necessarily more broad and less detailed. Less successful applications result in multiple iterations, long delays, and excessively restrictive terms of use. The contributors to this segment have very different perspectives: Di Persio has a great deal of experience formulating successful protocols that result in the megacorpora for which LDC is known. Warner carries out research in phonetics and language documentation at a large regional university and ~~sits on~~ that university's board of ethics, as well

as the LSA's. Their individual experiences in presenting research to review boards are complementary and critical in combination.

Sociolinguists who recognize the importance of data sharing have sometimes assumed (Yaeger-Dror/Cieri 2013) that the field already has an accurate and complete evaluation of specific *demographic factors*. Demographic designations, while part of the sociolinguistic 'collective consciousness' for decades, appear to require finer distinctions than many have acknowledged – or analyzed – in the past. Recent studies have revealed the inadequacy of specific common demographic distinctions such as *Age* (Bowie, this issue, Wagner 2012), *Sex* (Eckert 2008, this issue), and *Ethnicity* (Bayley, Blake, Wong/Hall-Lew this issue, Hall-Lew/Yaeger-Dror 2014). An additional topic of interest for future sharing of LDC and other corpora is that demographic coding varies across corpora. As Cieri (this issue) shows, inconsistencies in coding (though his examples are linguistic, not demographic) may preclude accurate comparison of apparently comparable corpora. Future collaborative research will benefit from the papers on this theme, which demonstrate the insights gained by adopting more refined categories, and encouraging readers to adopt a similar approach. These works blaze the trail for efforts that systematically coordinate data collection across community boundaries and advance the field by addressing critical scientific questions that require such data.

While ethnic coding has sometimes followed the choices made available by the census bureau or the Pew Trust, recent work, even in the popular press (Funderburg/ Schoeller 2013), has demonstrated how underdetermined the choices have been. In fact, the editors have proposed distinguishing component features that have been conflated into *ethnicity*. Yaeger-Dror & Cieri (2013) emphasize that even the term 'ethnicity' underdetermines variation in critical ways and propose that specific sources of variation previously labeled 'ethnic' – for example regional-heritage, linguistic-heritage, and religion – should be considered separately. Blake discusses various regional-heritage designations that ingroup 'Black' speakers find relevant to their identity as members of a 'Black' community. Hall-Lew and Wong discuss the extent to which, even within the Cantonese-~~speaking~~ community, other demographic factors influence speakers' own perception of their Asian and/or Cantonese identity. ~~Bailey~~ considers analyses of Latin-American speakers and clarifies the extent to which recent research has made such a designation not only underspecified, but inappropriate in some instances. He reminds us that even within the US, Latin Americans from different regions differ radically, both because their regional and racial affiliations differ and because their history in the US has differed. Given the space limitations, he does not even broach the extent to which speakers who have Latin~~a~~ and African heritage actually consider themselves in a disjunct group from other speakers who share ethnic attributes with them (as shown by the work of e.g., Vasquez et al 1997, Toribio 2003, and others). Contributions to this theme also point out that, just as US speakers often have multiple identities, their ethnic allegiances may include more than one identity. Blake, in this issue, reminds us of Coupland's (2003) warning that we should distinguish social network (*community-as-association*) from actual demographic identity (*community-as-demography*), and both from speakers' chosen self-identifier(s) at a given moment (*community-as-value*). We should be aware of individual speakers' demographic history, but also their social network, how they see themselves and how they would wish to be seen – within the specific interaction, as well as in their broader self-identification.

While the question of religious choice has been limited in both the US Census and sociolinguistic protocols, Yaeger-Dror (this issue) reviews literature demonstrating the importance of *religion* ~~in~~ studies of linguistic variation and presenting evidence that future studies should permit more fine-grained religious metadata, given the evidence that different subcommunities within a larger community may visualize themselves (*community-as-value*), and form social groups based on those affiliations (*community-as-association/* community of practice)

and present themselves linguistically as members of a specific more narrowly defined community. The paper suggests specific religious designations, which have already been shown to influence speakers' linguistic self-presentation and cites studies showing that even within a specific religion there are narrower communities-as-value, where speakers use both linguistic and sartorial choices to display more narrowly defined religious identities. The paper discusses accumulating evidence that religion cannot be used as a stand-in for 'ethnicity', but should be distinguishable from other sources of ethnic self-identification. The evidence also leads to the conclusion that *community-as-value* concepts entail that speakers' political beliefs also are relevant to their linguistic choices (e.g., Hall-Lew et al 2012, Bourhis et al 2009), and might properly be regarded as relevant metadata for linguistic study.

Sociolinguists are particularly attuned to the Observer's Paradox (Labov 1972) and to the various factors that all appear to be included in what we loosely refer to as 'style' (Eckert and Rickford 2001), but *social situation* (Eckert/Rickford 2001, Rickford, and Tagliamonte, this issue, Becker 2014) is less likely to be coded systematically than demographics. Two sociolinguists who focus on very different corpora, and who have given the problem much thought, weigh in with their review of situational factors that correlate with linguistic variation but are often ignored in research protocols and therefore in resulting publications as well, since they are generally impossible to specify after the fact. Rickford reminds us to pay attention to 'serendipitous situational switching' while Tagliamonte reminds us of a vast trove of situational factors, which we all have been aware of, but which few studies code for systematically, and which, therefore, often remain unacknowledged creating later confusion. The work of both of these authors and much of the work from New Zealand (Gibson and Bell 2012; Drager et al 2010) and Hawaii (Drager et al 2013) demonstrate that evaluation of one's own and one's interlocutor's identity are influenced by the immediate social situation, and vice versa, so that interaction among the demographic and situational features can become quite complex. It is now generally accepted that a speaker's attitude toward both his or her own 'ingroup' and toward other social groups will influence the linguistic choices that she/he makes in a given social situation. Speakers tend to adapt their degree of linguistic affiliation with a specific group to the social situation within which the conversation takes place. Conversations among speakers from the same group are more likely to demonstrate their in-group bona fides than the same speakers interacting with ~~and~~ out-group speakers or interviewer. They are also more likely to demonstrate in-group bona fides where the demographic factor being discussed is relevant (Schegloff 1972). The interaction among simultaneously relevant ethnic identities provides a rich environment for such manipulation of multiple identity markers (Hall-Lew & Yaeger-Dror 2014, especially Becker 2014). These observations suggests that fieldworkers need to control for and document features, which could be relevant to a speaker's presentation of self. Eckert, Rickford, and Tagliamonte (this volume) all discuss documenting such situational features. There are now several techniques for coding *social attitudes* to permit the analysis of the interaction between attitudes and linguistic choices. One set of choices developed by sociolinguists (Watts & Llamas, this issue) and a discussion of choices that have been made by social psychologists (Noels, this issue) are represented here as well.

4. *Limitations and Themes for Future Studies*

Although the workshop and the resulting special journal issue provide a detailed and balanced discussion of several problems involved in developing a methodology for interoperable metadata, there are other topics, which have been left to future studies.

The demographic features discussed in depth have been considered primarily from an American point of view. The relevant demographics may differ radically for other research

communities. For example, when Americans use the hyphenated terms African- (Blake, this issue) or Asian- (Wong/Hall-Lew, this issue), they describe groups whose demographic makeup differs radically from what one might imagine to be the equivalents in the UK. When Americans say 'Asian' they typically refer to east Asian community identities, often Cantonese speakers, while when British sociolinguists use the term, the default understanding is that the speakers are Bangladeshi, or Pakistani (Sharma 2011,~~;~~ Alam & Stuart-Smith 2014). Similarly, Bayley's review points out that most American sociolinguists use the term 'Latin' to refer to speakers from a broad spectrum of regional, racial, and religious heritages, with different contact situations with the larger US English-speaking community. The papers in this volume should be immediately helpful to the US research community but require adaptation for researchers working elsewhere. Nor have we tried to address some other demographic considerations critical to sociolinguistic analysis: for example, the understanding of and interaction among *Education* (Jahangiri & Hudson 1982), *Socioeconomic Status* (Labov ~~2011~~), and *Linguistic Marketplace* (Sankoff/Laberge 1978) could require a ~~workshop~~ of their own.

While it is generally assumed that a given corpus will be fairly uniform in *social situation*, as Rickford and Tagliamonte have pointed out, even within the limits of the sociolinguistic interview situation is often more complex than generally acknowledged. Most studies only code for the background of one person in an interview, leaving the situational considerations dependent on interaction among two or more speakers inadequately documented. ~~B~~both papers identify other situational features relevant to linguistic choices ~~such as~~ the interactive context and relationship between the speakers.

Perhaps the most important concern that the workshop and special issue have not had time to resolve is, given the limited resources of time, fieldworker-training and good-will among talkers, the optimal use of these conclusions to produce corpora that are appropriate for their initial intended purpose ~~but~~ also adequate for possible reanalysis. If there is an interview, how much time should the interviewer spend on detailed elicitation of the demographic, situational and attitudinal factors discussed? To what extent should we control for or document the identity of the interviewer, setting, and visual cues like those shown to be relevant in Hay & Drager (2010). Given the influence of the form of a question on the answers received, how shall we control for or document the form of questions? To further limit the observer's paradox (Labov 1972), would researchers be better advised to request that potential subjects complete an online survey as LDC sometimes does or should they favor, as Noels (this issue) suggests, presenting the questionnaire as part of the interview protocol, so the interviewer can tease out more detailed answers when necessary? Finally, how do we code subject responses consistently and document the numerous methodological choices that were made so that future researchers can understand how the information was gathered, make use of shared corpora, and replicate findings effectively? Without expecting to ever create the ideal set of metadata categories, elicitation instruments and coding specification, we hope the special issue has raised consciousness about variables which have been under-differentiated in previous studies and which we should attend to in future research.

Acknowledgements

Efforts of this type always rely upon the generous contributions of many scholars and organizers without whose help the entire endeavor becomes impossible. The editors are grateful to NSF, LDC the LSA secretariat, and the authors who contributed to the workshops and special issue. Particular thanks go to Joan Maling and Christine Massey who gave unstintingly of their time to facilitate workshop planning and to David Robinson whose unflappable organizational skills saved the day (actually 3 days)! We also thank Denise DiPersio, Marian Reed, Laurel

MacKenzie, and Brittany McLaughlin who did the heavy lifting as we planned and ran the workshop. We would also like to thank those who contributed to the special issue, the authors themselves, many of whom also refereed other papers in the volume, outside referees: Alex D'Arcy, Cece Cutler, Lauren Hall-Lew, Michael Newman, Benjamin Tucker, Suzanne Wagner, Keith Walters, and the editors of the journal, whose patience surpasses all. Finally, thanks to all the workshop attendees whose presence and enthusiasm and subsequent activities including data sharing via LDC and otherwise demonstrate the importance of topics raised by the workshop.

~~References to articles in the issue (all have been referred to! Tell us if you want us to provide a list, or if you can just transpose an extra copy of the toc here)~~

Short Biographies

Malcah Yaeger-Dror is a research scientist in the Department of Linguistics and Cognitive Sciences at the University of Arizona. She carries out sociophonetic research on English, French, Spanish, and Hebrew. Her research considers interdisciplinary questions having to do with language, identity, human subjectivity, the social psychology of language choice, and the cognitive underpinnings of situational variation in speech.

Christopher Cieri trained as a sociolinguist at the University of Pennsylvania and is Executive Director of the Linguistic Data Consortium, where he has overseen the development and distribution of hundreds of corpora containing collections of text, broadcast, conversations, prompted speech, meetings, and interviews that are subsequently transcribed, aligned, and annotated in many ways including sociolinguistic coding. His own research interests include variation and language contact especially in phonetics and phonology, and methodological issues in corpus use.

Note

* Correspondence address: Malcah Yaeger-Dror, Department of Linguistics, University of Arizona. E-mail: malcah@gmail.com

Works Cited

Alam, F., and J. Stuart-Smith. 2014, forthcoming. Identity, ethnicity, and fine phonetic detail: an acoustic phonetic analysis of syllable-initial /t/ in Glaswegian girls of Pakistani heritage. *English in the Indian Diaspora (varieties of english around the world)*, ed. By M. Hundt and D. Sharma. Amsterdam: Benjamins.

Bourhis, R. Y., G. Barrette, S. El-Geledi, and R. Schmidt. 2009. Acculturation orientations, social relations between immigrant and host community members in California. *Journal of Cross-Cultural Psychology* 40.443–467.

Drager, K., J. Hay, and A. Walker 2010. Pronounced rivalries: Attitudes and speech production. *Te Reo* 53.27–53.

Drager, K., M. J. Kirtley, J. Grama, S. Simpson. 2013. Language, variation and change in Hawai'i English. PWPL 19:Article 6.

Eckert, P., and J. Rickford, eds. 2001. *Style and sociolinguistic variation*. NY: Cambridge University Press.

Funderburg, L., and M. Schoeller. 2013. The changing face of America. *National Geographic,* 83-119. (URL, downloaded 11/28, http://ngm.nationalgeographic.com/2013/10/changing-faces/funderburg-text#; http://ngm.nationalgeographic.com/2013/10/changing-faces/schoeller-photography)

Gibson, A., and A. Bell. 2012. Popular music singing as referee design. *Style shifting in public*, ed. by J. Hernández-Campoy, and J. Antonio Cutillas-Espinosa, 139–164. Philadelphia: Benjamins.

Giles, H., R. Y. Bourhis, and D. M. Taylor. 1977. Towards a theory of language in ethnic group relations. *Language, ethnicity, and intergroup relations*, ed. by H. Giles, 307–48. London: Academic Press. Q13

Hall-Lew, L., R. Starr, and E. Coppock. 2012. Style-shifting in the U.S. congress. *Style shifting in public*, ed. by J. Hernández-Campoy and J. A. Cutillas-Espinosa, 45–64. Philadelphia: Benjamins.

Hall-Lew, L., and M. Yaeger-Dror (eds) 2014. New perspectives on linguistic variation and ethnic identity in North America. *Language and Communications* 35(1) Q14

Hay, Jennifer, and K. Drager. 2010. Stuffed toys and speech perception. *Linguistics* 48. 865–892. Q15
Labov. 1972. *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.
Nagy, N., J. Chociej, and M. F. Hoffman. 2014. Analyzing ethnic orientation in the quantitative sociolinguistic paradigm Q16
language and communication 35/1.
Schegloff, E. 1972. Notes on a conversational practice: Formulating place. *Studies in social interaction*, ed. by D. Sudnow, 75–119. New York: MacMillan.
Sharma, D.. 2011. Style repertoire and social change in British Asian English. *Journal of Sociolinguistics* 15. 464–492.
——. Forthcoming. Changing identities: Asianness and social class in Britain. Racing language, languaging race, ed by H. Samy Alim, A. Ball, and J. Rickford. Stanford: Stanford University Press.
Toribio, A. J. 2003 The social significance of Spanish language loyalty among Black and White Dominicans in New York. *Bilingual Review* 27. 2—11.
Vasquez, L. A., E. Garcia-Vasquez, S. Bauman, and A. Sierra. 1997. Skin color, acculturation, and community interest among Mexican American students: A research note. *Hispanic Journal of Behavioral Sciences* 19. 377–386.
Yaeger-Dror, M., and C. Cieri. 2013. Prolegomenon for an analysis of dialect coding conventions for data sharing. *Methods in dialectology*, ed by A. Barysevich, A. D'Arcy and D. Heap, 189–204. Bamberg Series: Lang.

The added references include

Becker K. 2014. Linguistic repertoire and ethnic identity in New York City. Language and Communication 35. 43-54.
Coupland, N. 2003. Sociolinguistic authenticities. Journal of Sociolinguistics 7. 417-431.
Eckert, P. 2008. Variation and the indexical field. Journal of Sociolinguistics 12. 453-476.
Jahangiri, Nader & R. Hudson. 1982. Patterns of variation in Tehrani Persian. Sociolinguistic Variation in Speech Communities, ed by S. Romaine, 49-64. London: Edward Arnold.
Labov, W. 2001. Principles of Linguistic Change. II Social Factors. Malden: Blackwell.
Wagner 2012. Age grading in sociolinguistic theory. Language and Linguistic Compass 6.371-382.