Toward the Harmonization of Metadata Practice for Spoken Languages Resources

Christopher Cieri[*], Malcah Yaeger-Dror[*●]
[*]Linguistic Data Consortium, University of Pennsylvania, [●]University of Arizona
[*]3600 Market Street, Suite 810, Philadelphia, PA 19104, USA
E-mail: ccieri@ldc.upenn.edu, malcah@email.arizona.edu

Abstract

This paper addresses issues related to the elicitation and encoding of demographic, situational and attitudinal metadata for sociolinguistic research with an eye toward standardization to facilitate data sharing. The discussion results from a series of workshops that have recently taken place at the NWAV and LSA conferences. These discussions have focused principally on the granularity of the metadata and the subset of categories that could be considered required for sociolinguistic fieldwork generally. Although a great deal of research on quantitative sociolinguists has taken place in the Unites Stated, the workshops participants actually represent research conducted in North and South America, Europe, Asian, the Middle East, Africa and Oceania. Although the paper does not attempt to consider the metadata necessary to characterize every possible speaker population, we present evidence that the methodological issues and findings apply generally to speech collections concerned with the demographics and attitudes or the speaker pools and the situations under which speech is elicited.

Keywords: metadata, sociolinguistics, standards

1. Introduction

The brief history of building digital, shareable language resources (LRs) to support language related education research and technology development is marked by numerous attempts to create and enforce standards. The motivations behind the standards are numerous. For example, standards offer the possibility of making explicit the process by which LRs are created, establishing minimum quality levels and facilitating sharing. Nevertheless, there have been instances in which the pre-mature or inappropriate promulgation or adoption of standards has lead to its own set of problems (Osborn 2010, p. 74ff, Mah, et. al. 1997) as researchers struggle to apply to their use cases standard that were not truly representative and perhaps not intended to be. To reduce the potential effort expended in developing, promoting and using proposed standards that may subsequently be found difficult to sustain, we propose that standardization is a late step in a multipart process that begins with understanding, progresses to documentation that may itself encourage consistency in practice within small groups at which point the question of standardization begins to ripen.

2. Background

The present workshop seeks to survey current initiatives in speech corpus creation with an eye toward standardization across sub-disciplines. Such standardization could permit resource sharing among researchers working in conversation and discourse analysis, sociolinguistics and dialectology among others and between those fields and others who depend upon similar kinds of data including language engineers (Popescu-Belis, Zufferey 2007). Coincidentally, the authors have been involved in a number of workshops on related themes including a series taking place at the annual NWAV (New Ways of Analyzing Variation) meetings on speech data collection, annotation and distribution including documentation and metadata description. More recently they lead a workshop funded by the U.S. National Science Foundation at the 2012 winter meeting of the Linguistics Society of America[1]. The principal topics of the latter were metadata description and related legal issues in the creation of spoken language corpora for sociolinguistics. This paper constitutes a summary of efforts within that community to begin understanding metadata encoding practice as a first step toward consistency, sharing and standardization.

3. Towards Standardization

Before metadata practice can be standardized, individual researchers must first understand their practices, the variations among them, the causes for variation, the tradeoffs of different approaches and their potential uses. In particular, researchers need to know if they can apply their metadata categories consistently, a question that is not frequently asked but must be if the goal is to adopt a standard that will be used by many independent groups with the intent of sharing corpora. Once the practice is understood it must be documented so that potential users can evaluate it and competing practices can be harmonized to permit appropriate comparisons. With adequate documentation independent researchers can decide if they want to adopt consistent practices.

4. Metadata

Within sociolinguistics, some researchers' position is that each study requires its own set of demographics. However, the ultimate consensus at the workshops was that cross community comparative corpus-based studies are only possible if there is a shared set of specific coding choices. Some of the demographic information is generally accepted within the larger sociolinguistic community: sex, birth year, years of education, and some designation of job description are fairly common

[1]http://projects.ldc.upenn.edu/NSF_Coding_Workshop_LSA/index.html

demographic fields, as are designations for where the speaker grew up, where the speaker lives at the time of the interaction, along with what years a given speaker has spent in specific regions.

5. Ethnicity

Within the American linguistics community 'ethnicity' frequently conflates three quite distinct demographic features: race, region and religion. Each of these will be discussed in turn.

5.1 Race[2]

While recent US based studies generally distinguish between black[3] or African American, Hispanic and other ethnic categories, sometimes referred to as "dominant dialect"; this is now understood to be insufficient: "black" speakers may be of Haitian, Jamaican, Dominican, or African provenance, and may not consider their primary identity as African American [henceforth AA]. Within the US, African Americans whose parents grew up in the North, can generally be distinguished from those whose parents grew up in the South. So if ethnicity choices are limited to the above three, there may be no confusion in a community where all "black" speakers are in fact African American, but in large cities much confusion could result from the failure of the coarse term to capture the three-way distinction [Blake and Shousterman 2010]. Speakers of mixed race [e.g., Purnell 2009 2010] have also been shown to differ consistently from both "white" and "black" linguistic groups within their communities.

While both the Pew Trust[4], and the Mumford Center[5] have treated Asian as a viable group, it is clear that speakers whose parents emigrated from India and Pakistan have very little in common [ethnically, regionally or religiously] with those whose parents hailed from Japan or Korea or China. It has been shown that even different Chinese groups can be distinguished from each other [Hall-Lew/Wong 2012]. It has also been shown that coding subjects for when their forebears left their country of origin reveals correlation with linguistic choice, a connection that in retrospect should not be surprising since the settlement patterns and trajectory of integration into the larger community differed for speakers arriving at different times [Sharma 2011; Wong/Hall-Lew 2012].

5.2 Regional/Linguistic Heritage

Given that Hispanic ancestry speakers are racially quite diverse within the Americas, the discussion of Hispanic heritage speakers of various racial and regional provenance is even more complicated. While the English syntax of Hispanic ancestry speakers seems to be convergent [Bonnici/Bayley 2010], the English phonology differs for even the most similar regional groups, for example Cuban, Puerto Rican and Colombian-Costeños [Bayley/Bonnici 2009] or Mexican, Texan, Californian and New Mexican Chicanos. As with the 'Asian' speakers discussed earlier, the interaction of settlement conditions with date of arrival has a strong influence on speaker variation. [Bayley/Bonnici 2009].

5.3 Religion

There have now been many studies which demonstrate that specific racial and regional heritage groups should also be divided by religion: For example, it has long been known that even in Ireland, [Milroy, 1980], Wales [Bourhis/Giles 1978] Belgium [BOURHIS, et al 1979) and the Middle East [Miller 2007, Walters 2011] different religious groups, which share the same racial and regional heritage, speak quite differently from each other, even to the extent of using different languages. For example Sunni, Shia, Copt, Maronite all speak quite differently, despite the fact that they are ethnically 'Egyptian' [Miller 2005]. Conversely, the 'New York Jews' referred to in Tannen's early work [Tannen 1981] – not to mention 'Muslims' [Miller 2007] or, for that matter 'Christians' [Wagner to appear] can belong to quite different racial and regional heritage groups, and are often linguistically quite distinct. As a result, conflating 'racial heritage' 'regional heritage' and 'religion' threatens to obscure distinctions that have been shown to be significant in numerous community studies.

Within individual studies, it is necessary that field sociolinguists determine which racial, regional and religious heritage speakers are likely to be included in their sample and prepare to control effectively for these distinctions. Unfortunately, such information is generally not coded for easy access. In fact, among the corpora currently available, even in the few cases that include protocols for eliciting speaker metadata, the protocols generally do not suggest asking these questions of speakers. Even sociolinguist interviewers, who are 'primed' by their protocol to elicit appropriate demographic information fail to probe in order to distinguish among relevant subgroups. Moreover, researchers often assume that if subjects have answered demographic questions, these answers are somehow available, despite the fact that the information may be buried in the often untranscribed interview audio. Furthermore, Lieberson (1992) shows that interviewees are not always honest or accurate in their representation of the regional, racial and religious background they belonged to during their formative years.

5.4 The Melting Pot and Multiple Identities

While in some societies, there may be little mixing among demographic or religious groups, in the US large numbers of those born since the 1970's actually belong to, and identify with, multiple demographic groups (Blake and Shousterman 2010). Coding practice needs to

[2] Any discussion of the validity of the concept or label 'race' is well beyond the scope of this paper. When we use the term here, we are merely referring to the traditional use of the term as a very broad categorization.
[3] We use the term occasionally to highlight the lack of further analysis.
[4] http://www.pewtrusts.org
[5] http://mumford1.dyndns.org/cen2000

permit the association of multiple values even for a single speaker and a single variable. A researcher may decide to give priority to the first-named 'identity', but the schema should allow for multiple listings. Mature metadata schema should also acknowledge the possibility of changing affiliation over time.

6 Encoding Demographics

Sociolinguists, historically, have assumed that the best way to do so is to incorporate relevant questions about 'ethnicity' and attitudes toward 'ethnicity' into a questionnaire executed during an interview. However, unless the interviewer has been sensitized to the fact that finer distinctions are needed, they may feel no obligation to spend time on the relevant questions. Furthermore there are no generally accepted instructions for encoding subjects' free form answers into regularized form so that future researchers can access it without having to listen to the interview in its entirety. In short a protocol for eliciting information about demographic and attitudes must be accompanies by a protocol for encoding this information into a form searchable by future scholars even if future scholars is ultimately only the same researcher returning to the data after some hiatus. Recent work has made clear that an accurate assessment of dialect change requires returning to a community 20 or more years later [Wagner 2012, in press], by which point even the original research team may no longer recall the details of an original interview. Even someone returning to a group of speakers previously studied will be under-served by a coding protocol that assumes that demographic information is adequately encapsulated in the interview itself and need not be formally coded.

7 Socioeconomic Information

Although many corpora include metadata for 'years of education', years spent in a technical school are not distinguished from those spent in what is commonly referred to as 'higher education', a fact that some research communities has already noted (Graff, pc). Moreover, multiple studies have demonstrated the usefulness of community specific scales for the importance of the 'dominant dialect' among speakers with different job descriptions, the so-called *linguistic marketplace* [Sankoff, Laberge 1978]. Even where a scale has not been devised in a given community, each speaker's occupation could be listed as well, which will permit subsequent scaling of socioeconomic and linguistic marketplace variation within a given community.

8 Politics

While it is not always possible to ask speakers about their political opinions, there have been recent articles showing that since speakers' politics strongly influence their attitudes toward their own and other groups, and their attitude toward the 'dominant dialect' of their region [Abrams et al 2011, Bourhis et al 2009, Hall-Lew et al 2010]. Some awareness of speakers' politics should be coded if possible.

9 Social Situation

Labov's early work clearly demonstrated the importance of the social situation. (See Labov 2001 for an overview.) However, the presumptions on the part of sociolinguists that every speaker is equally aware of the current social situation, that those speakers present an accurate view of the situation to interviewers and that the knowledge the community researcher has come internalize is equally obvious to outside readers are all likely to mislead. A transparent means for encoding and preserving descriptions of social situations would improve the usefulness of data sets and the ability to compare one to the other.

9.1 Interlocutor Dynamics

It has been shown that even in a straightforward interaction, the actual interlocutor is not necessarily the principal 'audience' [Bell, 1984]. At the same time, even in an interview situation, the interlocutor [interviewer] effect is pervasive [Hay/Drager 2010; Llamas et al 2009]. That said, very few corpora provide adequate descriptions of the interlocutors, including interviewers, despite the fact that this is significant in the analysis of the subject's speech.

9.2 Social Attitudes

The recent workshop at LSA as well as 4 decade's evidence from social psychological studies documented the importance of speakers' attitudes toward their own and other groups for the analysis of their speech [Giles 1973, Giles et al 1977]. In fact, the earliest studies in the social psychology of language demonstrated the variability of social attitudes even within one interaction [Giles 1973]. These factors could also be coded for, particularly if a post interaction questionnaire could be provided. While social psychologists have proposed elaborate and extensive questionnaires (Abrams et al 2009, Bourhis et al 2009, Noels 2012.) Recent work by Labov et al (2011) and by Llamas and her coworkers (Llamas 2012) have shown that the critical information can be determined with fewer questions, and with those questions presented online.

10 Broader Methodological Issues

Although our focus has thus far centered on studies conducted by sociolinguists, frequently within the United States, a number of tensions have emerged for which we have no solutions yet but which must figure into any discussion of metadata for speech corpora. We have seen that conflating 'racial heritage' 'regional heritage' and 'religion' may obscure distinctions we wish to preserve. Taken to its logical extreme, the desire for completeness and fine-granularity in elicited speaker metadata must necessarily be constrained by the limited time available for any single speaker given the other requirements of a representative speaker sample. We also see tensions between the communities with which a speaker may identify and those with which an outsider may associate the speaker. A third tension exists among the actual

methods for eliciting metadata. Checklists and multiple choice questionnaires offer the promise, perhaps misleading, of clean distinctions between metadata categories and values while ethnographic style interviews tend to recognize the inherent ambiguity of categories but exact a cost later in the analytic process of rendering textual descriptions into categories of comparison.

10 Conclusion

To reduce the effort expended in developing, promoting and using proposed standards that may subsequently be found difficult to sustain, standardization should be a late step in a process that begins with understanding, progresses to documentation that hopefully leads to consistent practice and the ultimately to standardization. The research community focusing on quantitative analysis of language variation has begun to examine its own processes and identifies a number of challenges even in the assignment of metadata for speakers and interview sessions. Among them we have noted too the use of metadata categories that are too coarse to reveal correlation already shown to exist in the literature, the conflation of multiple dimensions into a single super-category that, again, fails to capture distinctions expected to be significant. In addition we have noted a generally absence of explicit descriptions of the complete elicitation and encoding practices and, presumably as a result, a tendency to avoid entire metadata categories that other scholars have found to be revealing. By carefully enumerating the opportunities for improving metadata elicitation and providing infrastructure to support new efforts, such as template questions and coding schemata, it is the authors' hope that the community will begin to move toward consistent practice that facilitates greater data sharing and the benefits that naturally result from it.

11 Acknowledgements

We are grateful for the funding supplied by NSF BCS Grant #1144480, which made much of this work possible.

12 References

Abrams, Jessica, Valerie Berker & Howard Giles. 2009. An examination of the validity of the Subjective Vitality Questionnaire Journal of Multilingual and Multicultural Development. 30:59-72.

Bayley, R. & Lisa Bonnici. 2009. Recent research on Latinos in the United States and Canada, part 1: language maintenance and shift and English varieties. Language and Linguistics Compass 3:1300 –1313.

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

Bell, A. 1984. Language Style as audience design. Language in Society 13: 145-204.

Benor, Sarah (ed) 2011. Special issue of Language & Communication 31

Blake and Shousterman 2010. Diachrony and AAE: St. Louis, Hip-Hop, and Sound Change outside of the Mainstream Journal of English Linguistics 38: 230-247

Bonnici, Lisa & R. Bayley 2010 Recent research on Latinos in the USA. Part 2: Spanish Varieties. Language and Linguistic Compass 4: 121-134.

Bourhis, Richard, G. Barrette, S.El-Geledi, R. Schmidt 2009. Acculturation Orientations and Social Relations between Immigrants and Host Community Members in California. Journal of Cross-Cultural Psychology 40: 443-467.

BOURHIS, R. Y. & GILES, H. 1977. The Language of Intergroup Distinctiveness. In H. Giles (Ed.), Language, Ethnicity and Intergroup Relations. London: Academic Press. Pp 119-135.

BOURHIS, R.Y., GILES, H., LEYENS, J.P. & TAJFEL, H. 1979. Psycholinguistic distinctiveness: Language divergence in Belgium. In H. Giles & R. St-Clair (Eds.), Language and Social Psychology, Oxford: Blackwell. Pp. 158-185.

Bowie, David 2012. Religion: elicitation and metadata. Presented at the LSA in Portland, to appear. (http://projects.ldc.upenn.edu/NSF_Coding_Workshop_LSA/index.html)

Giles, H. 1973. Accent mobility: A model and some data. Anthropological Linguistics, 15, 87–105.

Hall-Lew, Lauren 2010. Ethnicity and sociolinguistic variation in San Francisco Language and Linguistic Compass 4(7):458-72.

Hall-Lew, Lauren, Elizabeth Coppock and Rebecca L. Starr. 2010. Indexing Political Persuasion: Variation in the Iraq Vowels. American Speech, 85(1):91-102.

Hay, Jennifer and Katie Drager 2010. Stuffed toys and speech perception. Linguistics 48(4):865-892.

Hay, Jennifer, Paul Warren and Katie Drager 2010. Short-term exposure to one dialect affects processing of another. Language and Speech 53(4):447-471.

Hay, Jennifer, Katie Drager and Paul Warren 2009. Careful who you talk to: An effect of experimenter identity on the production of the NEAR/SQUARE merger in New Zealand English. Australian Journal of Linguistics 29(2):269-285.

Labov, William 2001. Principles of Linguistic Change Vol. II: Social Factors. Blackwell: Oxford.

Labov, William, Sharon Ash, Maya Ravindranath, Tracey Weldon, Maciej Baranowski and Naomi Nagy 2011. Properties of the sociolinguistic monitor. Journal of Sociolinguistics 15: 431–463

Lieberson, Stanley 1992. The enumeration of ethnic and racial groups in the census: Some devilish principles. In: J. Charest & R.Brown (eds) Challenges of measuring an ethnic world. US Govt Printing Office: Washington.

Llamas, Carmen 2012, to appear. Paper presented at the LSA Workshop. (http://projects.ldc.upenn.edu/NSF_Coding_Workshop_LSA/index.html)

Llamas, C., D. Watt, & Daniel Ezra Johnson. 2009. Linguistic accommodation and the salience of national identity markers in a border town. Journal of Language and Social Psychology 28(4). 381-407.

Mah, Carole, Julia Flander, John Lavagnino. 1997, Some Problems of TEI Markup and Early Printed Books, Computers and the Humanities 31:31–46.

CAROLE MAH, JULIA FLANDERS and JOHN LAVAGNINO

Miller, Catherine 2005. Between accommodation and resistance: Upper Egyptian migrants in Cairo. Linguistics 43(5): 903–956.
Miller, Catherine 2007. Arabic in the city: issues in dialect contact and language variation. Routledge.
Milroy, L. 1980. Language and Social Networks. Oxford: Blackwell Publishers.
Noels, Kim. 2012, to appear. Paper presented at the LSA Workshop. (http://projects.ldc.upenn.edu/NSF_Coding_Workshop_LSA/index.html)
Osborn, Don, 2010, African Languages in a Digital Age: Challenges and Opportunities for Indigenous Language Computing, Ottawa, International Development Research Center.
Popescu-Belis, Andrea, Sandrine Zufferer, 2007, Contrasting the Automatic Identification of Two Discourse Markers in Multiparty Dialogues, in Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue, pages 10–17, Antwerp, Association for Computational Linguistics.
Purnell T. 2009. Convergence and contact in Milwaukee: Evidence from select African American and white vowel space features. Journal of Language and Social Psychology 28(4): 408-427
Purnell, Thomas C. 2010. The Vowel Phonology of Urban Southeastern Wisconsin. In Yaeger-Dror and Thomas, eds, AAE speakers and their participation in local sound changes: A comparative study. Publications of American Dialect Society #94. Raleigh: Duke University Press. 191-217.
Sankoff, D. and Suzanne Laberge 1978. The linguistic market and the statistical explanation of variability. In D. Sankoff (ed.), Linguistic Variation: Models & Methods. NY: Academic Press. 239-50.
Sharma, Devyani 2011. Style repertoire and social change in British Asian English. Journal of Sociolinguistics 15: 464-492.
Suarez, Eva Maria 2010. Dominican identity and lg choice in the Puerto Rican diaspora. Nwav39.
Tannen, D. 1981. New York Jewish conversational style. IJSL 30:133-149.
Wagner, Suzanne Evans. 2012, in press. Age grading in sociolinguistic theory. Language and Linguistics Compass.
Wagner, Suzanne (to appear) Linguistic correlates of Irish-American and Italian-American ethnicity in high school and beyond. In Yaeger-Dror & Guy. PADS #97. Duke University Press: Raleigh.
Walters, Keith 2011. Gendering French in Tunisia: language ideologies and nationalism
IJSL 211: 83-111.