Evolving question formats for data coding conventions.[1]

For the past **several** years the Linguistics Data Consortium at the University of Pennsylvania [LDC url: www.upenn.ldc.edu] has been collecting and archiving linguistic data for future use. At different points over these years I have made use of the LDC corpora, as well as the Montreal French Corpus, and various Talkbank corpora. I have also had the fortune [or misfortune] to be a frequent referee. In the process it has become quite clear that sociolinguists need help from those who have carried out previous archives, because we often assume that our corpora have transparently followed a shared protocol for research, and which permits them to be compared, when in actuality, there are data gathering conventions which are not shared, and are not even expressly stated so that others can determine the degree to which corpora are sharable even within the same community.

There are exceptions, primarily in Canada [Toronto-Ottawa-Montreal], where corpora follow the conventions first discussed in the work of Sankoff and Cedergren (Sankoff & Sankoff 1973), and further elaborated when the 1984 Montreal corpus required a set of rules to permit comparability (Thibault/Vincent 1989/1990). There are also comparable corpora in the UK, where even in the 'infancy' of the BNC, the research community's focus was on corpus gathering that would permit archival comparisons.

In addition, PADS#94 was made up of corpora that were built as parallel corpora of data from different areas of the country so they could be shared, and there has been a noteworthy effort on the part of NYU researchers, and those who are carrying out parallel studies in other cities, to formulate protocols which will permit comparability.

HOWEVER, the bulk of recent corpora are not necessarily comparable, and there is not always a way to determine the degree to which a corpus conforms to the general rules for sociolinguistic corpora or not.

At the same time, the NSF has enacted new rules which require that we all document in a proposal how the data will be made sharable by researchers from other communities. The members of the Variationist List have also been discussing how best to gather data, so that it is later sharable. With all that in mind, Chris Cieri from LDC and organized this workshop to help address some of the difficulties which we know will crop up.

1. **DATA**: The first section will consist of speakers who already have been working with very large corpora, making them available in open archives, and even making their transcription and coding available….Brian MacWhinney (www.talkbank.org), Gary Simons (www.sil.org; www.gold….), Maxine Eskenazi (), and Wade Shen (www.ll.mit.edu) will all provide some perspective.
2. **Ethics Issues**: How do we maximize our ability to be sensitive to the ethical issues involved in setting up a project, while minimizing the amount of time on IRB genuflection/expediting the IRB process, so (student) researchers can carry out their studies? Natasha Warner from the LSA's own ethics committee and Denise DiPersio from LDC will both make suggestions and answer questions.
3. **Metadata**: --Labov Rule #1 is:Code for any variable that might be relevant, because one cannot go back and add more distinctions later without totally disrupting the study. But what information is the minimal amount of information that we should store? How do we ask for it [given that different types of questioning will get different [and probably noncomparable] answers? How can we insure that our question formats as well as the number of choices available are the same? There seem to be several foci for coding that are often not addressed at all, much less addressed using identical question formats. Metadata which can be shared are also in their infancy. NSF has helped us to invite several speakers

[1] Thanks to both the NSF, and the LDC for the funding which has made this **study** possible. Thanks also to the many sociolinguists who have agreed to take part in the NSF sponsored workshop.

who have already coded for specific variables, and can provide perspective on how to ask appropriate questions, so that the answers can be used for coding:

a. **Demographic coding** is constantly upgraded to reflect the information on significant community subgroups, but there are still recurring problems where our coding has turned out to be too 'coarse grained', but most studies include
   i. birthdate,
   ii. speaker sex,
   iii. speaker age at the time of recording,
   iv. racial heritage,
   v. and some form of discussion of other sociodemographic factors.

b. **Demographics** we are less 'on top of' include such issues as:
   i. the regional heritage of the speaker; often studies retain an outmoded under-differentiated regional coding scheme. Since this type of coding is more dependent on the specific region one is working in, it will not be discussed here today.
   ii. While speaker sex has been consistently coded for over the last several years, evidence has been piling up that sexuality can also be a critical determinant of speaker variation, and **Penny Eckert** will present a background paper on the importance of sexuality as a factor in language variation….
   iii. Ethnicity—specifically ethnic designations that we all tend to overgeneralize in our work—will form one section today, with **Renée Blake** discussing various social groups which do not consider themselves 'African American', or which have multiple self-identities some of which may outweigh that ethnic identity, at least some of the time; **Amy Wong & Lauren Hall-Lew** discuss various 'Asian' subgroups for which 'Asian' might be an inappropriate identifier, and **Carmen Fought** discussing the fact that 'Latino' is not one uniform ethnicity, but that different national, regional, racial or ethnic sub-identities are more salient in many –or even most--situations.
   iv. The speakers' educational achievements are generally found somewhere in the transcript even if they haven't been coded for, but while earlier studies were prone to code for a rather rigid 'socioeconomic scale', which was explicated in the text, recent studies may not even code for how a speaker makes a living, much less the SES or ML.
   v. Despite the early work of the Milroys (e.g., 1987), the work of descriptive linguists, and of social psychologists like Giles and Bourhis (1973), which has shown that the religious and political background of the speakers strongly influences social attitudes and opinions (which in turn influence speech), both political and religious persuasion are generally ignored. Even when they are discussed, information on individual religious persuasion (much less commitment to that religion) is very rarely considered in a sociolinguistic study. [When religion is coded, it is relegated to a category within 'ethnicity' for lack of a generally accepted 'religion' category. It is embarrassing to read the work of descriptive linguists, like Catherine Miller (2005) or Clive Holes (e.g, 1986), who have been describing for years the degree to which dialect features which initially appear to be regional are actually traceable to speakers' degree of religious commitment and political attitudes, and those, in turn, influence the dialect variation, so that a more conservative part of a country will differ considerably from the more modern areas. Sociolinguists have not been coding for either of these. **David Bowie** will discuss questions which have been found to successfully provide this

key information without giving offence or leading to less-than-honest responses.

4. **Social Attitudes**: Although Howard Giles and many other social psychologists publish work on how **social attitudes** influence language *choice*, it is only recently that sociolinguists have begun to ask questions that will permit us to code for relevant attitudes in our analyses of dialect variation. **Carmen Llamas** (Llamas, Watt & Johnson 2009; Watt, Llamas & Johnson 2010), **Naomi Nagy** (2011, 2012) and their colleagues, and **Shana Poplack** (e.g., 2007) have all spent years comparing ways of eliciting information about social attitudes without contaminating the data with the 'interviewer effect', and **Kim Noels**, from the social psychology community which studies language variation, will all share some of their insights on how to ask such questions, how to preserve the coded responses to permit later comparison with the linguistic results, and what questions may be most useful.
5. **Situation**: While we generally assume that we have been coding for the **social situation**, and there have been many studies of 'style' or 'register', quite often individual corpora totally ignore situation, do not code for it, even superficially, and it is not always automatically determinable from a transcript, even when there is one. This is all the more embarrassing given not only Labov's work, but that of Hymes (1964), and Giles (as early as Giles & Powesland 1975), as well as the more recent work which is often cited (Eckert&Rickford 2001; Bell 1984, 2001; Coupland 2007….), which all point out that [among other things] one aspect of the social situation is actually dependent on the speakers' attitudes toward their interlocutors' social group memberships. **Sali Tagliamonte** will be presenting some coding conventions which she feels have served her well for the analysis of interview data, and which we can follow, to carry out more appropriate coding of situation, while **John Rickford** will also consider situational features which are needed for data from other social settings.….
6. **Discussion & Conclusions**: The last section will be devoted to a generalized discussion, so those about to go carry out fieldwork can ask questions, and those who have been carrying out fieldwork, and who have perhaps discovered more efficient ways to code and archive can do so. All the speakers will be available, as will others, like **Tyler Kendall**, who have also been involved in sociolinguistic archiving of data both here and in Europe.

References

Abney, S. & S. Bird. 2010. The Human Language Project: Building a universal corpus of the World's languages. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden.

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.

Allard, R. & R. Landry, 1994. Subjective Ethnolinguistic Vitality: A Comparison of Two Measures. – *International Journal of Sociology of Language*, 108: 117-144.

Allen W., Joan Beal, K. Corrigan, W. Maguire, H. Moisl. 2007. A linguistic 'time-capsule': The Newcastle Electronic Corpus of Tyneside English. In Beal, Corrigan, Moisl, eds. *Creating and Digitizing Language Corpora: Diachronic Databases*. Basingstoke:Palgrave. 16-48.

Auer, Peter, F. Hinskens, & P. Kerswill eds, 2005. *Dialect change: Convergence and divergence in European languages*. Cambridge University Press, Cambridge. ISBN 978-0521806879

Beal, J. C. , K. P. Corrigan & H. L. Moisl, eds, 2007. Creating and Digitizing Language Corpora: *Synchronic Databases*, Vol.1. Basingstoke: Palgrave-Macmillan.

Becker, Kara and Elizabeth L. Coggshall. 2009. The sociolinguistics of ethnicity in New York City. *Language and Linguistic Compass* 3.3: 751-66.

Bell, Allan. 1984 Language style as audience design. *LiS* 13: 145-204.

Back in style: re-working Audience Design. In P. Eckert and J. R. Rickford (eds), Style and sociolinguistic variation. New York: Cambridge University Press. 139-69.

Biber, D. and S. Conrad 2009. *Register, Genre and Style*. Cambridge: CUP

Bird, Steven and Gary Simons. 2004. Building an Open Language Archives Community on the DC Foundation. In Diane Hillmann and Elaine Westbrooks (eds.), *Metadata in Practice*, pp. 203–222. Chicago: American Library Association.

Blake, Renée and Cara Shousterman 2010a Diachrony and AAE: St. Louis, Hip-Hop, and Sound Change outside of the Mainstream. *Journal of English Linguistics* 38:230-248.

___&___ 2010b Second generation West Indian Americans and English in NYC. *English Today* 103(26#3) 35–43.

Bourhis, Richard and Geneviève Barrette 2006. Notes on construction of a 'subjective vitality questionnaire' for Barrette. Notes on the immigrant acculturation scale (IAS). Working Paper, *LECRI*, Département de psychologie, Université du Québec à Montréal, Canada. November.

___, ___, 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.

Bowern, Claire. 2010. Fieldwork and the IRB: A snapshot. *Language* 86: 897-905.

Boyd, Sally, M. Hoffman, J. Walker. 2011. Sociolinguistic practice among multilingual youth in Sweden and Canada. ISB8. Oslo.

Brown, R. & A. Gillman 1960. Pronouns of power and solidarity. In: Sebeok, T. (Ed.), *Style*. Cambridge, MA: MIT Press. 253–276.

Cheshire, Jenny & Paul Kerswill 2011. The emergence of Multicultural London English. Delivered at ISB8.

Childs, Becky, G. Van Herk, and J. Thorburn 2011. Safe Harbor: Ethics and accessibility in sociolinguistic corpus building. *Corpus Linguistics and Linguistic Theory* 7:163-180.

Cieri, C. & M. Yaeger-Dror 2011. An evolving perspective on the concept of ethnolect. Paper presented at Methods in Dialectology 14, UWO, August.

Coggshall, Elizabeth L. 2008. The prosodic rhythm of two varieties of native American English. *Penn Working Papers in Linguistics* 14.2:1-9. PDF.

___ and Kara Becker. 2010. A vowel comparison of African American and white New York City residents. Malcah Yaeger-Dror and Erik R. Thomas, eds. Pp. 101-128.

Coupland, Nik 2007. *Style: Language Variation and Identity*. Cambridge: CUP.

Di Paolo, Marianna and Malcah Yaeger-Dror, eds. 2010. *Sociophonetics*. Routledge.

Dobrin, Lise and Claire Bowern. 2009. Making the inevitable valuable: Ethics and ethics review in linguistic fieldwork. Paper presented at the annual meeting of the Linguistic Society of America, San Francisco, CA, January 2009.

Eckert, P. 2000. *Linguistic Variation as Social Practice*. Oxford: Blackwell.

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

Eckert, Penelope. 2008b. Where do ethnolects stop? *International Journal of Bilingualism* 12: 25–42.

___ and J. Rickford 2001. *Style and Sociolinguistic Variation*. Cambridge: CUP.

Eskenazi, Maxine 1995. Hot Topics in Speaking Style Research, in *European Studies in Phonetics and Speech Communication*, Bloothooft, Hazan, Huber, Llisterri, eds., OTS Publications, The Netherlands. P. 58 - 62.

___& A. Black. 2010. SDC: The Spoken Dialog Challenge, invited presentation at SIGDIAL.

Fought, Carmen 2006. *Language and Ethnicity*. Cambridge, UK: Cambridge University Press.

Giles, Howard 1973. Accent mobility. *Anthropological Linguistics* 15: 87-105.

Giles, H. & Richard Bourhis 1973. Dialect perception revisited. *Quarterly Journal of Speech* 59:337-342.

Giles, Coupland and Coupland (eds) 2001.*Contexts of Accommodation*. Cambridge: Cambridge University Press.

Giles, Howard & Peter Powesland (eds.) 1975 *Speech style and social evaluation*. NYC: Academic Press:

Gregersen, Frans & M. Barner-Rasmussen 2011. The Logic of comparability: On genres and phonetic variation in a project on language change in real time. *Corpus Linguistics and Linguistic Theory* 7:7-36.

Haeri, Niloofar 2003. *Sacred Language, Ordinary People.* Palgrave.

Hall-Lew,L. 2005. One Shift, Two Groups: When fronting alone is not enough *PWPL* 10.2:105-116.

___ 2009. 'Ethnicity and phonetic variation in a San Francisco neighborhood.' Unpublished PhD dissertation. Palo Alto: Stanford University.

___ 2010. Ethnicity and sociolinguistic variation in San Francisco *Language and Linguistic Compass* 4(7):458-72.

___ & A. Wong 2011. Chinese-Americans in NYC and SF. For upcoming LSA.

Hernández Campoy & Cutillas Espinosa (eds) in press *Style-Shifting in Public: New Perspectives on Stylistic Variation.* Philadelphia: Benjamins.

Holes, Clive. 1986 The social motivation for convergence in three Arabic dialects. *IJSL* 61: 33-51

Hoffman, Michol 2011. The importance of being ethnolect: an international perspective. Symposium at International Symposium on Bilingualism (ISB)8, Oslo.

Hymes, Dell 1964. *Language in Culture and Society*. Harper & Row.

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

Kendall, Tyler 2011 Beyond Research Alone. Paper presented at Methods in Dialectology 14, UWO, August.

___ & G. Van Herk 2011. *Corpus Linguistics and Linguistic Theory* 7/1.

Koops, Christian 2010. /u/-Fronting is not Monolithic: Two Types of Fronted /u/ in Houston Anglos. *PWPL* 16. 2: Article 14

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

___, Ash, Boberg 2006. *Atlas of North American English (ANAE)*. Berlin: De Gruyter.

Llamas, Carmen

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

MacWhinney, B. 2006. The multidisciplinary analysis of talk. *Hungarian Studies* 20:143-162.

___ 2007a. Opening up video databases to collaborative commentary. In R. Goldman, R. Pea, B. Barron & S. Derry (Eds.), *Video research in the learning sciences*, pp. 537-546. Mahwah, NJ: Lawrence Erlbaum Associates.

___ 2007b. The TalkBank Project. In J. C. Beal, K. P. Corrigan & H. L. Moisl (Eds.), Creating and Digitizing Language Corpora: Synchronic Databases, Vol.1. . Houndmills, Basingstoke, Hampshire: Palgrave-Macmillan.

___ 2008. Enriching CHILDES for morphosyntactic analysis. In H. Behrens (Ed.), *Trends in corpus research: Finding structure in data*, pp. 165-198. Amsterdam: John Benjamins.

Miller, Catherine (2005) Between accommodation and resistance: Upper Egyptian migrants in Cairo. *Linguistics* 43(5): 903–956.

Miller, Catherine, Enam Al-Wer, Dominique Caubet & Janet C.E. Watson (eds.) 2007 *Arabic in the City: Issues in Dialect Contact and Variation.* London: Routledge.

Milroy, Lesley 1987. Observing and analyzing natural language. Oxford: Blackwell.

Myhill, John. 2006. L*anguage, Religion and National Identity in Europe and the Middle East: A historical study* (Discourse Approaches to Politics, Society and Culture) Amsterdam: Benjamins.

Nagy, Naomi 2011. Heritage language variation and change: Corpus construction and use. Paper to be presented at Methods in Dialectology 14, UWO, August.

Niedzielski, N. & Dennis Preston 2000. *Folk Linguistics*. NY: De Gruyter.

Poplack, S. 2007. Foreword. In J. Beal, K. Corrigan & H. Moisl, H. eds. *Creating and digitizing language corpora.* Houndmills : Palgrave-Macmillan. ix-xiii.

___ 2011. Paper presented at Methods in Dialectology 14. UWO, London, Ont.

Rickford, John in press….

Rose, Y., Hedlund, G., Byrne, R., Wareham, T., & MacWhinney, B. 2007. Phon 1.2: A Computational Basis for Phonological Database Elaboration and Model Testing. In P. Buttery, A. Villavicencio & A. Korhonen eds., *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition*, 45th Annual Meeting of the Association for Computational Linguistics, pp. 17-24. Stroudsburg: ACL.

Sagae, K., Davis, E., Lavie, E., MacWhinney, B., & Wintner, S. 2007. High-accuracy annotation and parsing of CHILDES transcripts *Proceedings of the 45th Meeting of the Association for Computational Linguistics*. Prague: ACL.

Sagae, K., Davis, E., Lavie, A., MacWhinney, B., & Wintner, S. 2010. Morphosyntactic annotation of CHILDES transcripts. *Journal of Child Language* 37:705-729

Sankoff, David, H. Cedergren, W. Kemp, P. Thibault, & D. Vincent 1989. Montreal French: language, class and ideology. In ed. R. Fasold & D. Schiffrin eds., *Language Change and Variation*. Amsterdam: John Benjamins. 107-18.

___ and Gillian Sankoff. 1973. Sample survey methods and computer-assisted analysis in the study of grammatical variation. In R. Darnell (ed.), *Canadian Languages in their Social Context*. Edmonton, Alberta: Linguistic Research. Pp. 7-64.

Scanlon, M. and A.Wassink 2010. African American English in Urban Seattle. *American Speech* 85:163-184.

Simons, Gary F. 2006. Ensuring that digital data last: The priority of archival form over working form and presentation form. SIL Electronic Working Papers 2006-003. Dallas: SIL International. Online: http://www.sil.org/silewp/abstract.asp?ref=2006-003.

Sykes, Robert. 2011. An acoustic description of Utah English vowels. 161st meeting of The Acoustical Society of America. Seattle, WA.

Tagliamonte, S. 2011a Peaks and trends in linguistic change. Penn colloquium.

___2011b. The joys and perils of building, analyzing and sharing corpora. Paper presented at Methods in Dialectology 14, UWO, August.

Thibault, Pierrette 1990. Questionnaire. In Thibault, Pierrette and Diane Vincent. *Un corpus de français parlé*. Laval University: Québec. Available online at ERIC: http://www.eric.ed.gov/PDFS/ED348875.pdf, with the questionnaire in an appendix, pp110-121. Partially republished as: La langue en mouvement. *LINX* 25: 79–92.

___ & Michele Daveluy 1989. *Quelques traces du passage du temps dans le parler des montréalais*. *Language Variation & Change* 1:19-45.

Van de Velde, Hans, Roeland van Hout & Marinel Gerritsen. 1997. Watching Dutch change: A real time study of variation and change in standard Dutch pronunciation. *Journal of Sociolinguistics* 1(3). 361-391

Van Hofwegen, Janneke and Walt Wolfram 2010. Coming of age in African American English: A longitudinal study. *Journal of Sociolinguistics*. 14/4: 427–455,

Watt, Dominic, C. Llamas, D. E. Johnson 2010. Levels of linguistic accommodation across a national border. *JEngL* 38:270-289.

Wolfram, Walt 2007. Sociolinguistic folklore in the study of African American English. *Language and Linguistic Compass* 1: 292-313.

___ 2010. Why we care about the development of AAE. Nwav39.

___, Clare Dannenberg, Stanley Knick, and Linda Oxendine. (2002) *Fine in the World: Lumbee Language in Time and Place*. Raleigh:NC State Humanities Extension Publications.

___ & Natalie Schilling-Estes. 2006. *American English: Dialects and Variation*. Oxford: Blackwell.

Wong, Amy 2007. Qualifying paper on field work for study of Chinese Americans in New York City. NYU.

___2011. New York City English and second generation Chinese Americans. *English Today* 26: 3-11.

Yaeger-Dror, Malcah & T.C. Purnell, eds. 2010a. *Accommodative Tendencies in interdialect communication*. Special issue of *JEngL* 38(3).

___ & ___, eds. 2010b. *Accommodation by Second Language Learners*. Special issue of *American Speech* 85(2).

___ and Erik R. Thomas, eds. 2010. *African American English Speakers and Their Participation in Local Sound Changes: A Comparative Study* . Publication of the American Dialect Society # 94. Durham, NC: Duke University Press.