What kind of data is it?
Situating sociolinguistic corpora in context

Workshop on sociolinguistic archive preparation, LSA 2012

Sali A. Tagliamonte
University of Toronto

In my presentation I will discuss how sociolinguistic corpora can be compiled so as to document and maximize access to the context of its collection. This is no doubt a murky issue for the coding and categorization enterprise, but it is as critical as demographic information if we are going to be able to compare data sets from different communities, eras or across research projects. However, how far does the researcher go in documenting this type of information? My goal in this presentation will be to outline what I have found to be 'best practice' in my own research while at the same time highlighting issues and problems I have encountered along the way. I build on the foundations of earlier corpus-building projects (1991; Sankoff & Sankoff, 1973; Sankoff & Cedergren, 1971; Thibault & Vincent, 1990, Poplack, 1989 #1063) and on data arising from my own fieldwork conducted in the UK and Canada between 1995-2011 (Tagliamonte, 1996-1998, 1999-2001, 2001-2003, 2003-2006, 2007-2010, 2010-2013).

The original fieldwork situation is a critical component of any corpus because it determines the nature of the linguistic data. How comparable are data samples? This can only be known, if substantial information about the fieldwork situation has been recorded, and is retrievable for later work. Indeed, this fact highlights the extreme importance of the original research goals and practice. At the outset of data collection, the nature of the discourse to be obtained must somehow be planned, and then controlled (to the best of the researcher's ability). For example, if vernacular data is the goal, then a concerted effort to "tap the vernacular" is required. If a study of quotatives (or the historical present) is the goal, then story-telling is imperative. Similarly, if a study of future temporal reference is the goal, then the fieldworkers should be instructed to ask questions about future plans and intentions. Situational information can be recorded in field notes and post-fieldwork observations that are documented in the meta-data files comprise the following:

1) Time/date/place of interview
2) Interviewing technique
3) Interviewer(s)
4) Participant(s)
5) Interview context, e.g. *what was going on, what was it like, what happened?*

The time/date/place of the interview, (1), permits the corpus to be situated more broadly (but very specifically) in time and space. This is particularly important in recent years when researchers are beginning to conduct large-scale cross-variety studies (e.g. Buchstaller & D'Arcy, 2009; Tagliamonte, to appear; Tagliamonte, Durham & Smith, 2009).

1) TIME: Take, for example, the case of a rapidly diffusing innovation (e.g. quotative *be like*). In this case precise information about the date each corpus was collected is necessary because the

difference between 1995 and 2001 has a major impact on frequency and patterning.[1] In Canadian English, for example, the frequency of *be like* increased from 13% to 63% in this short 6 year time span within the same sector of the population (Tagliamonte & D'Arcy, 2007). Moreover, the date of birth of every individual within the corpus must be known and contextualized because the change is diffusing so quickly a 12 year old's grammar of *be like* will be entirely unlike a 35 year old's. Furthermore, the geographic location of data collection must be taken into account because origin, rate of spread and current state of grammatical change is also relevant.

Even limiting the discussion to a social situation defined as an 'interview', coding for interviewing technique is a means to characterize the nature of the data in the audio-record. Was it collected using the standard Sociolinguistic Interview techniques documented in the field (Labov, 1971, 1972b, 1984)? Were specific types of questions used for specific purposes? If so, how successful was each interview? Who did each interviews and how, (3)? These are key aspect(s) of the situation. Differences in interviewing technique and interviewer and interviewee styles can have a major impact on the nature of the data. It is also well-known that even sociolinguistic interviews comprise different discourse styles, including story-telling, soapbox speech etc. (Labov, 1972a, 2001) and that the contrasts among these evince entirely different linguistic behavior (e.g. Paradis, 1996). What havoc for our linguistic explanations will ensue when we discover that many recently compiled corpora are made up of situations that are not (sociolinguistic) interviews at all, but some other type of interaction?[2] At the very least researchers should make clear what type of data they are using.

Moreover, the relationship between participants in a corpus must be disclosed. This is because a single individual will express him or herself quite differently depending on the interlocutor(s) (Cukor-Avila & Bailey, 2001; Douglas-Cowie, 1978; Watt, Llamas & Johnson, 2009, 2010). This is why a record of the participants is even more important, (4). An individual who is interviewed one-on-one with an out-group interviewer will produce a different type of interaction than one who is interviewed with a local 'facilitator', and both differ from the interactions between actual friends. These characteristics of a data set are often detailed in the description of the project and can be a key component of the interpretation of the results. For example, the discussion of the African Nova Scotian English fieldwork and data collection practices comprise nearly 30% of Poplack and Tagliamonte (1991:307-315) and form a critical background and foundation for the analysis and interpretation of the results that follow. The interview context, (5), is perhaps the most nebulous of the situational categories since it is dependent on the fieldworker's observations and conscientious recording of the nature of the interview situation. Nevertheless, even rudimentary field notes can provide indispensible information for later retrieval. For example: Ritter (2008) was conducting a comparative study on stutterer vs. non-stutterer behavior with respect to linguistic variation. However, being an uncommon

---

[1] For example, a 5 minute phone call with strangers from the 1980's may not be ideally compared with casual, hour long, interviews from the 1990's or 2001's particularly when vernacular and/or stigmatized, and/or discourse features, and/or rapidly changing features are under investigation.

[2] A wide range of sociolinguistic corpora was not collected using standard sociolinguistic interviews. These include oral histories, interviews which were recorded for a broadcast to a much larger TV or radio audience, e.g. the 7-UP series, public speaking and others (e.g. Hernandez-Campoy & Cutillas-Espinosa, in press; Kemp & Yaeger-Dror, 1991; Van de Velde, Hout & Gerritsen, 1997; Yaeger-Dror, Hall-Lew & Deckert, 2002).

characteristic — only 1% of the population are stutterers — it was difficult to recruit participants. A search of the meta-data files of the Toronto English Corpus revealed that three individuals were reported in the field notes as "stuttering a lot". Examination of these audio files exposed a bona fide stutterer thus providing an invaluable informant for the research study. Thus, it becomes critical for the principal investigator on a sociolinguistic project to encourage fastidious anthropological observation. In addition, annotation and observations are ideally added at the transcription phase when highly successful and unsuccessful sections of the audio record become blatant.

Once this type of situational information is entered into a relational database it can be searched and processed in any number of ways. For example, in a study of relative *who* in the Toronto English Corpus (D'Arcy & Tagliamonte, 2010), we discovered a high correlation of *who* with highly educated middle aged women. Upon further examination of the interviews, we noticed that the frequency of *who* seemed to increase when the interviewer was a woman. Because the interviewer for each interviewee was recorded in the metadata, a quick search of the database enabled us to re-code the data file according to the nature of the interview: that is, the relevant factor was not only the demographics of the speaker, but the interaction of the two interlocutors' demographics which was more accurately seen as an aspect of the social interactive situation. This, in turn, led to an innovative new perspective on relative pronoun variation. This highlights the importance of all the participants in the interaction whose relative age, sex, age and ethnicity undoubtedly play into the nature of the interaction.

Other situational features that are easy to code and may prove relevant later on include a description of the personality of the interviewer, relationship of interviewer to interviewee (if there is one), and relationships among interviewees in the corpus (if there are any); type of surroundings (living room vs. front porch; grandfather clock, bird, aquarium, etc.), and particularly successful parts of the interview. Furthermore any outstanding features of the context should be noted, such as a person who stutters a lot (noted above), someone's whose beard or dentures interfere(s) with the mike, an interview where alcohol was involved, etc.

In summary, although the context of data collection is actually a multi-faceted conglomeration of elements (social, geographical, psychological, etc.), it is decisive for analyzing and interpreting the results of any (socio)linguistic study. The task of recording this influential information in a sociolinguistic corpus begins in the planning stages of research, comes to fruition in the fieldwork setting as anthropological observation, is organized and documented in the research lab, and continues to evolve as the research enterprise advances. As analysis proceeds, the working data files produced when a particular linguistic feature is subjected to analysis becomes an invaluable database on which to build future research. For example, a datafile created for the study of adverb morphology (e.g. –Ø vs. –ly) can morph into a data file for the analysis of intensifying adverbs, which can in turn morph into a study of adjective variation. A data file created for the study of quotatives, can morph into a study of tense alternation in narrative. In my own work, data files are built with the future in mind and are often used as foundations for further study. As more and more research studies are completed, the details relevant to ongoing work on the corpora evolve as well. An individual that has a high rate of *like* as a discourse marker might be expected to have a high rate of *be like* as a quotative. But do they? Rates of usage of key features can be added to the database for future study. As linguistic variables are

studied in the corpora, they can be entered into the same database along with other details that have been discovered along the way. As researchers ask new questions, additional contextual information may become relevant. The database can then be modified and/or augmented. However, this can only happen if the researcher has made it a priority to record key aspects of the fieldwork situation in the first place.

References:

Buchstaller, Isabelle & D'Arcy, Alexandra (2009). Localized globalization; A multi-local, multivariate investigation of quotative *be like*. *Journal of Sociolinguistics* 13(3): 291-331.

Cukor-Avila, Patricia & Bailey, Guy (2001). The effects of the race of the interviewer on sociolinguistic fieldwork. *Journal of Sociolinguistics* 5: 254-270.

D'Arcy, Alexandra & Tagliamonte, Sali A. (2010). Prestige, accommodation and the legacy of relative *who*. *Language in Society* 39(3): 1-28.

Douglas-Cowie, Ellen (1978). Linguistic code-switching in a North Irish village: Social interaction and social ambition. In Trudgill, P. (Ed.), *Linguistic Code-Switching in a North Irish Village: Social Interaction and Social Ambition.* Sociolinguistic Patterns in British English: London. 37-51.

Hernandez-Campoy, J-M & Cutillas-Espinosa, J.A. (Eds.) (in press). *Style-Shifting Revisited.* Amsterdam and New York: John Benjamins.

Kemp, William & Yaeger-Dror, Malcah (1991). Changing realizations of A in (a)tion in relation to the front a-back a opposition in Quebec French. In Eckert, P. (Ed.), *ew Ways of Analyzing Sound Change.* New York: Academic. 129-184.

Labov, William (1971). Some principles of linguistic methodology. *Language in Society* 1(1): 97-120.

Labov, William (1972a). The isolation of contextual styles. In Labov, W. (Ed.), *The Isolation of Contextual Styles.* Sociolinguistic Patterns: Philadelphia. 70-109.

Labov, William (1972b). The study of language in its social context. In Labov, W. (Ed.), *Sociolinguistic patterns.* Philadelphia: University of Pennsylvania Press. 183-259.

Labov, William (1984). Field methods of the project on linguistic change and variation. In Baugh, J. & Sherzer, J. (Eds.), *Language in use: Readings in sociolinguistics.* Englewood Cliffs, N.J.: Prentice-Hall. 28-54.

Labov, William (2001). Style and Sociolinguistic Variation. In Eckert, P. & Rickford, J. (Eds.), Cambridge: Cambridge University Press.

Paradis, Claude (1996). Interactional conditioning of linguistic heterogeneity. In Guy, G. R., Feagin, C., Schiffrin, D. & Baugh, J. (Eds.), *Towards a social science of language, Volume 2: Social interaction and discourse structures.* Amsterdam and Philadelphia: John Benjamins. 115-133.

Poplack, Shana & Tagliamonte, Sali A. (1991). African American English in the diaspora: Evidence from old-line Nova Scotians. *Language Variation and Change* 3(3): 301-339.

Ritter, Michael A. (2008). Variable variation across populations: The case of stutterers. Toronto: University of Toronto.

Sankoff, David & Sankoff, Gillian (1973). Sample survey methods and computer-assisted analysis in the study of grammatical variation. In Darnell, R. (Ed.), *Canadian Languages in their Social Context.* Edmonton: Linguistic Research Inc. 7-63.

Sankoff, Gillian & Cedergren, Henrietta (1971). Some results of a sociolinguistic study of Montreal French. In Darnell, R. (Ed.), *Some Results of a Sociolinguistic Study of Montreal French.* Linguistic diversity in Canadian society: Edmonton.

Tagliamonte, Sali A. (1996-1998). Roots of Identity: Variation and Grammaticization in Contemporary British English. Economic and Social Sciences Research Council (ESRC) of Great Britain. Reference #R000221842.

Tagliamonte, Sali A. (1999-2001). Grammatical variation and change in British English: Perspectives from York. Economic and Social Science Research Council of the United Kingdom. Research Grant R000238287. York, UK: University of York.

Tagliamonte, Sali A. (2001-2003). Back to the roots: The legacy of British dialects. Research Grant. Economic and Social Research Council of the United Kingdom (ESRC). #R000239097.

Tagliamonte, Sali A. (2003-2006). Linguistic changes in Canada entering the 21st century. Research Grant. Social Sciences and Humanities Research Council of Canada (SSHRC). #410-2003-0005. http://individual.utoronto.ca/tagliamonte/.

Tagliamonte, Sali A. (2007-2010). Directions of change in Canadian English. Research Grant. Social Sciences and Humanities Research Council of Canada. #410 070 048.

Tagliamonte, Sali A. (2010-2013). Transmission and diffusion in Canadian English. Standard Research Grant #410-101-129. Social Sciences and Humanities Research Council of Canada. (SSHRCC)

Tagliamonte, Sali A. (to appear). Variation as a window on universals. In Siemund, P. (Ed.), *Linguistic Universals and Language Variation* Berlin/New York: Mouton de Gruyter.

Tagliamonte, Sali A. , Durham, Mercedes, and & Smith, Jennifer (2009). Grammaticalization in time and space: Tracing the pathways of FUTURE going to across the British Isles. *UK LVC 7 (Language Variation and Change). September 1-3, 2009.* Newcastle, England.

Tagliamonte, Sali A. & D'Arcy, Alexandra (2007). Frequency and variation in the community grammar: Tracking a new change through the generations. *Language Variation and Change* 19(2): 1-19.

Thibault, Pierrette & Vincent, Diane (1990). La collecte des données: caractérisation de l'enquête. *La Collecte des Données: Caractérisation de L'Enquête.* Un corpus de français parlé: Québec.

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

Watt, Dom, Llamas, Carmen & Johnson, Daniel 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.

Watt, Dom, Llamas, Carmen & Johnson, Daniel E. (2010). Levels of linguistic accommodation across a national border. *Journal of English Linguistics* 38(3): 270-289.

Yaeger-Dror, Malcah, Hall-Lew, Lauren & Deckert (2002). It's not or isn't it? Using large corpora to determine the influences on contraction strategies. *Language Varation and Change* 14(1): 79-118.