Parsed corpora of vernacular speech: challenges and prospects for the study of syntax

Christina Tortora (work with Anthony Kroch and Beatrice Santorini)
contact info: ctortora@gc.cuny.edu

This talk will address various issues in the design and use of parsed corpora of vernacular speech, with two goals in mind: (a) to understand the challenges and potential pitfalls in transcribing and grammatically annotating speech data, and (b) to illustrate that this form of data brings with it unique opportunities for advancing syntactic theory, despite the challenges. These points will be illustrated with examples and pilot studies based on the *Audio-Aligned and Parsed Corpus of Appalachian English* (AAPCAppE; Tortora, Santorini, & Blanchette, in progress; csivc.csi.cuny.edu/aapcappe/).

Regarding the challenges in creating parsed corpora of vernacular speech, we discuss the problems of *transcription as theory* and *annotation as theory*. As noted by Ochs (1979), transcriptions of speech reflect the transcriber's theory of speaker intentions. For example, in the string *he went* [ə] *hunting*, a transcriber might take the schwa to be the hesitation *uh*, but the acoustic signal might instead reflect the speaker's use of the *a*-prefix (*he went a-hunting*). Likewise, English orthographic conventions may mask the true morpho-syntactic nature of particular linguistic entities. For example, the written string *they should have left* obscures the fact that the form commonly written as *have* is always pronounced [əv] following a modal, raising the question of whether *have* is an auxiliary at all, in these contexts. Related to this is ambiguity of form, relevant to annotation. For example, bare verb forms in vernacular English are often used with a past tense function, as in *they bring it yesterday*. However, in the absence of the time adverbial or other contextual clues, the choice of Part of Speech tag (past or present) becomes less clear.

We show, however, that despite these issues, corpus creators can devise transcription conventions which make the ambiguities transparent to the corpus user, thus maximizing the potential for care in use of the data. In addition, using the AAPCAppE as an example, we show that the challenges of creating such corpora are far outweighed by their benefits as tools for advancing syntactic theory.

First, because it is based on vernacular speech, the AAPCAppE consists of data which are not generally found in writing (or if found in writing at all, not with sufficient frequency to study them). One example involves data typical of Appalachian regional speech, namely, the elision of the form [əv] in the context of modals and infinitival-*to* (e.g., *they should __ left* and *they ought to __ left*). Our pilot study on these constructions reveals a number facts not previously known. For example, elision of [əv] occurs more frequently with infinitival-*to* than it does with modals. In addition, infinitival-*to* structures with missing [əv] exhibit a temporal/aspectual interpretation not always equivalent to the "same" structures with overt [əv]. Furthermore, this latter interpretive difference in infinitival-*to* contexts is not found with [əv] ~ ø variation embedded under modals. The variation and contextual differences found with this phenomenon in the AAPCAppE thus give a window onto the fine structure of the tense and aspectual functional domain of English which is otherwise not available. Kayne's (1997) theory of *of* (pronounced [əv]) as a complementizer offers a promising avenue for investigating the question; the parsed corpus itself provides an ideal means for doing so, also because it is paired with the underlying speech signal, which allows us to study — in unreflecting vernacular speech — the phonological properties of the morpho-syntactic entity in question. This gives us a more accurate picture of the nature of an element which has previously been analyzed as an auxiliary verb (with the exception of Kayne 1997). It is also important to understanding the true nature of the input for learners, which is speech (and not written text).

Another example involves data typical of English vernacular speech more generally, namely, sentence amalgams like *that's what bothers me is he can't help us* or *that's the only thing they do is fight* (O'Neill 2015). A clearer understanding of the nature of the various amalgam types is important to our theory of clausal architecture, but these structures are not found with enough frequency in writing to make a rigorous study possible. And as O'Neill (2015) reveals, there are limitations on the usefulness of grammaticality judgment tasks with amalgams; for example, it has thus far been impossible to isolate their prosodic features in an experimental context. But their prosodic features are directly relevant to the proper analysis of their syntactic structure. It is precisely in this way that a parsed corpus of vernacular speech like the AAPCAppE provides opportunities for research that are not otherwise available: as noted

above, the corpus is paired with a digital recording of the speech signal, which allows us to study the prosodic features of amalgams. Again, here, because of its properties, a parsed corpus of vernacular speech provides data which would not be found either in writing or through experimentation, and as a result this form of data provides opportunities to pursue important theoretical questions. In the case of amalgams, the AAPCAppE gives us a window onto the question of whether it is necessary to broadly allow exocentric structures or not.

In sum, like well-known historical parsed corpora (such as the *Penn Parsed Corpora of Historical English*, see references), a corpus like the AAPCAppE allows for large-scale quantitative studies that can reveal patterns of variation that would not otherwise be revealed through experiments like grammaticality judgment tasks. This becomes even more true the more we build such corpora, as the potential for cross-dialectal comparison increases. For example, the AAPCAppE can be used as a tool for comparison with other parsed corpora of English, both historical (PPCHE) and contemporary, such as the in-progress *Corpus of New York City English* (CoNYCE; nycerg.commons.gc.cuny.edu/conyce/). Our ultimate point is not that parsed corpora of vernacular speech should supplant experimentation or written corpora as a means for testing hypotheses. Rather, we wish to show that, despite the data problems presented by such corpora, they serve as tools essential to progress in linguistic theory.

References:

Kayne, R. 1997. "The English Complementizer *of*," *Journal of Comparative Germanic Linguistics* 1: 43-54. also in Kayne 2002, Parameters & Universals, OUP.

Kroch, A. & A. Taylor. 2000. *The Penn-Helsinki Parsed Corpus of Middle English* (PPCME2). Department of Linguistics, University of Pennsylvania. CD-ROM, second edition, http://www.ling.upenn.edu/hist-corpora/

Kroch, A., B. Santorini, & L. Delfs. 2004. *The Penn-Helsinki Parsed Corpus of Early Modern English* (PPCEME). Department of Linguistics, University of Pennsylvania. CD-ROM, first edition, http://www.ling.upenn.edu/hist-corpora/

Kroch, A., B. Santorini, & A. Diertani. 2010. *The Penn-Helsinki Parsed Corpus of Modern British English* (PPCMBE). Department of Linguistics, University of Pennsylvania. CD-ROM, first edition, http://www.ling.upenn.edu/hist-corpora/

Ochs, E. 1979. "Transcription as theory," in E. Ochs & B. Schieffelin (eds.) *Developmental Pragmatics*, pp. 43-72. New York: Academic Press.

O'Neill, T. 2015. *The domain of Finiteness: Anchoring without Tense in copular amalgam sentences*. PhD dissertation, The CUNY Graduate Center.

Tortora, C. & G. Johnson. 2016. We should expected this: the (non-)periphrastic past with modals and *to* in Appalachian English. Ms. CUNY.

Tortora, C., B. Santorini, & F. Blanchette. in progress. *Audio-Aligned and Parsed Corpus of Appalachian English* (AAPCAppE). http://csivc.csi.cuny.edu/aapcappe/