

A syntactically annotated corpus of Appalachian English

Christina Tortora

Data Management Plan

This project proposal aims to create an on-line, freely accessible, ~1,000,000-word syntactically annotated (or “parsed”) corpus of Appalachian speech, using the *Penn Treebank* method of syntactic annotation (<http://www.ling.upenn.edu/hist-corpora/annotation/>). The proposed parsed corpus will be accompanied by a full set of digitized, text-searchable recordings of the speech from which the corpus is transcribed, in the form of .wav files. The .wav files on which the parsed corpus is based will be text-searchable as a result of force-aligning the transcripts with the speech signal, using the *PPL Forced Aligner Online Processing System* developed by Jiahong Yuan of the University of Pennsylvania, and made available to the public (<http://martinet.sas.upenn.edu/PPLClient/>). The final product will thus afford an unprecedented approach to the analysis of synchronic English dialect data for all kinds of scholars. I will use resources at the College of Staten Island (of the City University of New York) to archive the files of the parsed corpus. All component files will be made available to the public, and searchable and manipulable with *CorpusSearch2* (corpussearch.sourceforge.net) and *Praat* (fon.hum.uva.nl/praat/), both freely accessible.

Here I discuss [A] the files that will be stored and made available to the public, and [B] a guarantee from the College of Staten Island that it has the appropriate resources to house the database.

[A] Files to be stored

In order to make the corpus fully functional in the ways described above, I will use resources at the College of Staten Island (CUNY) to archive the various component files of the corpus. The following are the component files which will constitute the parsed corpus; these will all be freely available to the public:

- (1) All of the .wav files of the recordings of Appalachian speech, from which the 1,000,000-word corpus is transcribed. These .wav files will be created through a digitization process, using either *Audacity* (<http://audacity.sourceforge.net/>) or *Praat* (fon.hum.uva.nl/praat/). The 1,000,000 words of digitized speech may be distributed over roughly 500-800 .wav files. The number may turn out to be smaller, depending on the techniques I develop to unify .wav files.
- (2) All of the “text grid” files, which represent the text from the transcripts aligned with the speech signal in the .wav files. In a normal text editor, these files look like lists of words with time stamps and some other forms of coding (see Appendix 2, in *Supplementary Documents*). When opened up in tandem with their matching .wav file in *Praat*, however, they constitute a grid of text which is lined up with the speech signal (see Appendices 1A and 1B, in *Supplementary Documents*). This is what will allow the user to do specific word searches that will take him/her to the desired points in the .wav files. For example, if a phonetician wants to view and simultaneously listen to the speech signal associated with all instances of the string “born” in the 1,000,000-word speech signal, this can be done via garden-variety text searches.
- (3) Ordinary text files of syntactically annotated text, in the style of the *Penn Parsed Corpora* (see <http://www.ling.upenn.edu/hist-corpora/annotation/index.html>, and then click on “Introduction” on the left, to see what these text files look like). These files will be searchable using *CorpusSearch2*, which is open-source and fully accessible at corpussearch.sourceforge.net.
- (4) A full, non-annotated, ordinary text version of the corpus (in the form of individual files that will likely be organized according to region of origin), for users who are not interested in the parsed version of the corpus, and who would find searching a normal text file of greater relevance.

(5) A catalogue, in the style of the *Penn Parsed Corpora*, laying out how to use the corpus, what the known problematic issues are, plans for updating and improving it, an invitation to users to contact me regarding mistakes and problems, etc. In terms of the question of intellectual property, I would ask that anyone who produces published work based on data gleaned from the corpus would cite it appropriately; otherwise, my intention is to make all of these archived files freely available.

I will also pursue various ways of making the linguistics community aware of this project's final product, by presenting at various conferences.

[B] Resources for housing the database

The College of Staten Island will be the optimal place to archive these files for combined use, as it is home to the City University of New York's *High Performance Computing Center* (HPCC). The HPCC hosts a number of large HPC server systems, as well as smaller servers, database systems, and web servers. I have already received a commitment letter from the College of Staten Island, stating that the CUNY HPCC will host the database and web servers required to support my proposed work (see tech support letter, in *Supplementary Documents*). I will be provided with the necessary networking, technical, and security support, including systems administration, database management, and web support. The CUNY HPCC will provide for daily back-ups and offsite storage in accordance with established HPCC policy to provide for long term data preservation.

The CUNY HPCC and the proposed systems required to support my project are network-connected via a dedicated one gigabit per second line to the CUNY point of presence and from there to Internet2, NYSERnet, and the commercial Internet, assuring for rapid query access to the corpus of Appalachian speech.