The DASL Project: a Case Study in Data Re-Annotation and Re-Use

Christopher Cieri, Stephanie Strassel

University of Pennsylvania and Linguistic Data Consortium
3615 Market Street, Philadelphia, PA 19104-2608 U.S.A.
{ccieri, strassel}@ldc.upenn.edu

Abstract

It is well known and often repeated that publicly available digital data encourages basic and collaborative research including the comparison of results across studies, the measurement of inter-annotator consistency and the use of stable data as a benchmark with which to compare new models and methodologies. Instances of such reuse abound. The reuse and re-annotation of the Switchboard and TDT corpora was described in detail during LREC 2000 (Graff and Bird 2000). Unfortunately, very few studies have actually focused on the issues surrounding re-use and re-annotation of data. The LDC project to develop Data and Annotations for Sociolinguists (DASL) encourages data sharing and the re-annotation and reuse of published data as an important complement to first-hand fieldwork. DASL annotators use a tool, developed for the project, that gives linguists access to the four corpora via the Internet and allows simultaneous annotation at multiple sites. In addition to the empirical study of linguistic variation among the speakers represented, this project will address methodological issues in the corpus re-use and in team based annotation of linguistic data. The paper will describe the tools, data and data formats developed for DASL, outline the challenges we have faced in re-annotating the data using a team approach and summarize the results to date.

1. Introduction

The project in Data and Annotations for Sociolinguistics investigates best practices in the use of digital speech corpora to address problems in sociolinguistic theory. The quantitative study of linguistic variation is necessarily based upon empirical observation and statistical description of linguistic behavior. Collecting and annotating databases plays a crucial role in quantitative linguistics. The current state of computing technology encourages the collection, annotation, analysis and even summarization and presentation of linguistic behavior wholly within the digital domain. The facile sharing of digital data encourages a whole range of positive practices. However, the use of speech corpora in quantitative linguistics also raises questions both theoretical and methodological. The goal of the DASL Project is to begin to address these issues via a case study analysis of a well-documented sociolinguistic variable via several large well-documented speech corpora, the development of new resources and the distribution of previously unavailable resources.

2. Value of Shared Linguistic Data

The ability to easily share digital data encourages collaboration via

- use of stable data to benchmark new or competing models and methodologies
- reannotation and reuse of existing data for new purposes
- measurement of interannotator consistency
- reduction of impediments facing new participants in the research community

Sharing data does not however diminish the value of ongoing data collection. Both the research and the research community benefit from new contributions. The researcher gains new skills and a unique appreciation of the subject pool while the research community gains not only a new data set but also new perspectives and new methodological approaches.

The DASL Project hopes to encourage data sharing and the re-annotation and reuse of published data as an important complement to first-hand fieldwork.

3. Motivation

Even having accepted the premise that data sharing is a good thing, one may still ponder why sociolinguistics was chosen as a field in which to experiment with data distribution. As a field, quantitative sociolinguistics is about 40 years. The community is sizeable; the annual conference of quantitative sociolinguists draws many more applicants than the 150-170 presentation slots it typically accommodates. Since the very beginning it has embraced field-work, empirical observation and quantitative analysis. Most of the researchers in this community collect data and produce research results that are potentially relevant to a broad community of researchers including speech and language engineers. Unfortunately local practice in that community does not encourage resource sharing, reannotation or re-use.

4. DASL Overview

This project in data sharing and reannotation investigates the well-documented process of t/d deletion in four large digital speech corpora: TIMIT, Switchboard-1, CallHome American English and Hub-4 English Broadcast News. These will be described below. A team of annotators is coding the corpora for t/d deletion so that interannotator agreement can be measured. The interface used to conduct the annotation allows linguists to interact with the corpora, both text and speech, via the worldwide web so that this project can generalize to include multiple sites.

In addition to the empirical study of t/d deletion and the methodological questions concerning the use of speech

corpora in sociolinguistics this project will address several other questions:

- How do the corpora used in this study relate to the data most commonly used in quantitative sociolinguistics, namely recordings of interviews?
- Do the insights gained from the large scale study of a geographically diffuse subject pool differ qualitatively from speech community studies?
- What is the rate of interannotator consistency for the task of coding t/d deletion?
- Can studies of similar variables be organized on a large scale with teams of non-specialist annotators?

One of the most interesting issues surrounding the use of the proposed corpora for sociolinguistics is that of style. Speaking style plays an important role in the stratification of many sociolinguistic variables. A great number of quantitative studies of variation rely on data collected during sociolinguistic interviews that may combine conversation, story telling and question answering with more formal interactions such readings and word games. The corpora used herein involve somewhat different interactions. How do these interactions fit among the constellation of styles already studied? Do these interactions have correlates in everyday life?

5. The Variable

The DASL Project began with the analysis of -t/d deletion in English. -t/d deletion is a well-understood, stable variable common in multiple varieties of English. This variable shows similar patterns of stratification within many diverse speech communities in which it has been studied. The variable is easily coded, making it an idea first choice in a study where inter-annotator agreement is a.

6. The Data

The data for this project come from four corpora, each created for a purpose other than sociolinguistics but capable of being reannotated to serve our purposes. The data has already been transcribed and segmented so that individual speaker turns can be retrieved separately. Within long speaker turns, individual pause groups are segmented.

The TIMIT corpus contains over 600 speakers each reading a set of 10 phonetically rich sentences selected from a larger pool. DASL investigates how this method and, more importantly, the resulting data compare to the reading selection and word lists elicitations common in sociolinguistic interviews?

The Hub-4 corpus contains many hours of broadcast news. Network anchor people produce most of these utterances though there are also man-on-the-street interviews. What can we learn about variation in this speech that is so familiar to American TV viewers.

In the Switchboard corpora, speakers participate in multiple, short, telephone conversation with each other. Does the resulting data pattern like the early part of a sociolinguistic interview when the interviewer is looking for topics of common interest?

The CallHome data, perhaps the most interesting, contains 30 minute conversations among family members and close friends. Although the participants know they are being recorded, it is clear that they often forget or ignore that fact. How does this data compare to the styles already well studied in quantitative sociolinguistics.

DASL will provide analyses of each of the corpora. Because the data is available in digital form with transcripts, it can be annotated and analyzed for multiple variables with relative efficiency.

7. Sampling

Before annotation begins we first search the corpus for word of potential interest. Because the corpora are segmented at the speaker turn or pause group level, locating the speech corresponding to a -t/d token is simple. During this first effort we searched the orthographic transcripts of the speech files. As a result the queries used are perhaps more complicated than necessary. The annotation tool accepts a query on the form of a regular expression. We used a regular expression to find any consonant followed by a "t" or "d" at the end of a word where the following word does not begin with a "t" or "d". Regular expressions and English orthography do not combine perfectly. Left alone this query would return many words that we still prefer to avoid. While we could train the annotators to ignore cases that look erroneously like candidates for -t/d deletion, the time required to reject these is significant. To ameliorate this problem we used a series of filters that removed the "false hits" from consideration. In the case of TIMIT, initial queries reduced a corpus of 54,387 words down to a review list of 2059 words of which 1578 were actual -t/d tokens.

For subsequent efforts we plan to use a pronouncing lexicon as an intermediary to the search. In other words, the pronouncing lexicon would provide the list of English words susceptible to -t/d deletion. The interface would then search for those words in the transcripts that are in turn time-aligned to the audio.

8. The Annotation

Once the corpora have been concordanced, filtered and prepared for annotation, they are exhaustively annotated.

8.1. Annotation Specification

The variable under study is -t/d deletion. For each token, the annotator makes judgments with respect to four factor groups: status of the dependent variable; morphological category; preceding segment and following segment.

8.2. Tools

Using a customized sociolinguistic annotation tool, users can query each corpus to select just those tokens of potential interest, greatly reducing the time needed to code data. An interactive web-based display allows annotators to view each token, listen to the utterance and view the corresponding waveform, access demographic data and code linguistic factors. The annotator can simply click on the word to hear it spoken.

Following each token, the interface displays the factors to be coded. Each factor is shown as a radio button, and coding a token entails clicking on the button corresponding to the relevant factor within each factor group.

A comment field also appears after each token for the annotator to record notes. Results are easily exported to a spreadsheet or statistical analysis package.

9. References

Guy, Gregory. 1991. "Explanation in variable phonology: An exponential model of morphological constraints." Language Variation and Change 3: 1-22.

Guy, Gregory. 1980. "Variation in the group and the individual: The case of final stop deletion." Locating language in time and space. Academic Press: New York.

Labov, William. 1989. "The child as linguistic historian." Language Variation and Change 1: 85-94.

Labov, William. 1975. "The quantitative study of linguistic structure." Pennsylvania working papers on linguistic change and variation, vol. 1, no. 3.

Neu, Helene. 1980. "Ranking of constraints on /t,d/ deletion in American English: A statistical analysis." Locating language in Time and Space. Academic Press: New York.

Roberts, Julie. 1995. The acquisition of variable rules: t,d deletion and –ing production in preschool children. University of Pennsylvania dissertation.

Romaine, Suzanne. 1984. "The sociolinguistic history of t/d deletion." Folia Linguistica Historica 2: 221-225.

Wolfram, Walt. 1969. A sociolinguistic description of Detroit Negro speech. Center for Applied Linguistics: Washington, D.C.

Figure 1: Data sets currently being annotated

Corpus	ISBN	Minutes	Type of Data
TIMIT	1-58563-019-5	630	Phonetically Rich Sentences
Switchboard-1	1-58563-121-3	12000	Short Conversations with Constrained Topics among Strangers
CallHome American English	1-58563-111-6	1200	Long Conversations with Free Topics among Intimates
American English Broadcast News	1-58563-109-4	6240	Broadcast News

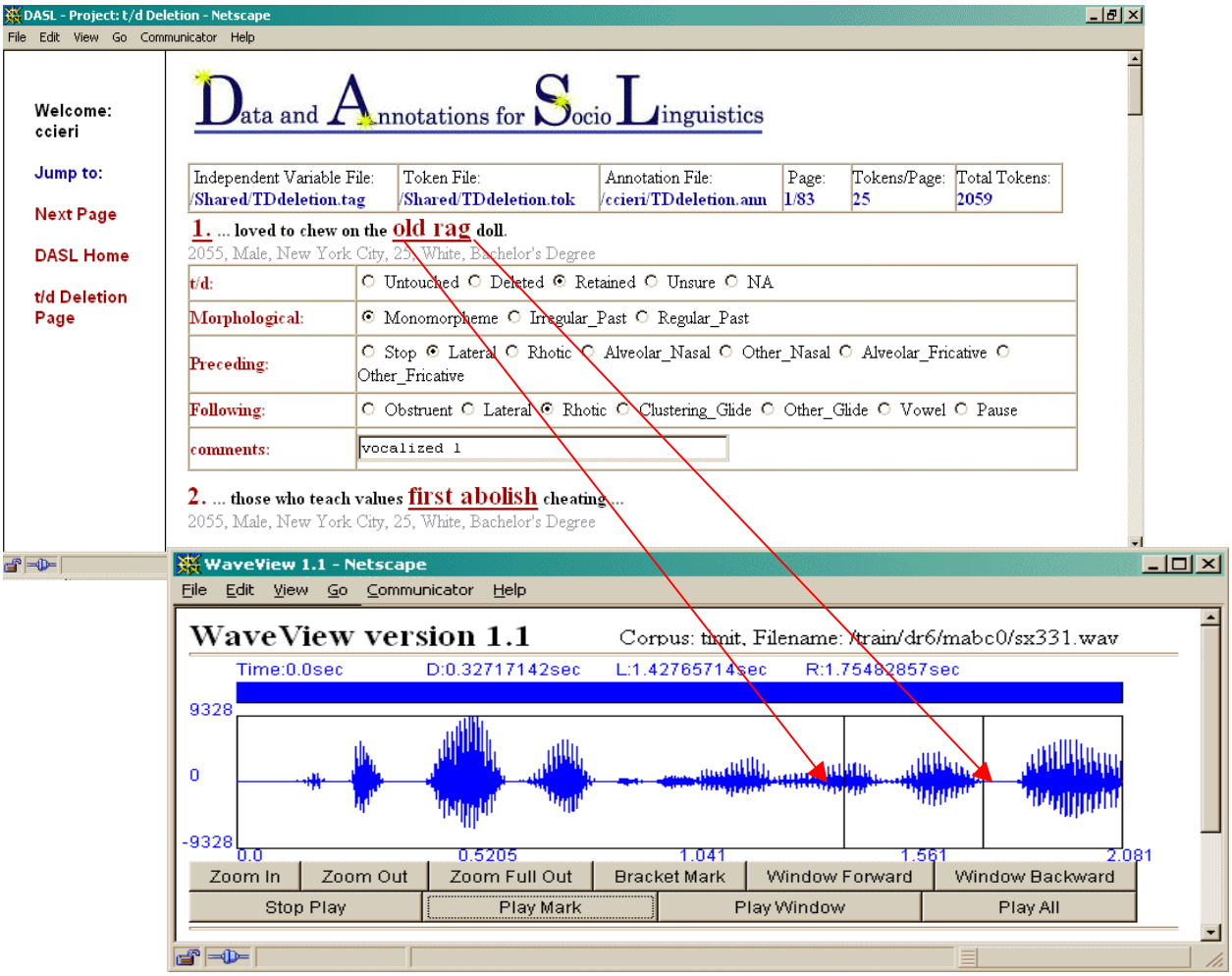

Figure 2: Web-based Annotation Interface