Data and Annotations for SocioLinguistics: A Corpus-Based Approach to Sociolinguistic Research

Stephanie Strassel and Christopher Cieri
University of Pennsylvania/
Linguistic Data Consortium
{strassel, ccieri}@ldc.upenn.edu

www.ldc.upenn.edu/Projects/DASL

Background

- Methodology for Quantitative Analysis of Variation
  - Established in late 60's; has undergone multiple refinements:

    Labov (1966, 1972), Labov, Yaeger, Steiner (1972), Sankoff 1980, Guy (1980, 1991)
    - speech community model
    - individual data collection, annotation, archiving(?) effort
    - high costs to individual researcher (or reduced effort, cutting corners)
  - Technological advances enable, encourage another update of methodology
    - wholly digital collection, analysis and presentation
    - shared resources
- Linguistic Data Consortium creates and shares language resources across a broad range of disciplines

Shared Resources

- Shared data resources and tools encourage
  - the comparison of results across studies, over time
    - replication of Labov's NYC department store study by Fowler (1986)
  - stable data as benchmark for competing theories (Labov 1997)
  - the re-annotation and reuse of existing data
    - Although not a substitute for first hand data collection, stable data permits broad and comparative investigations.
  - the measurement of inter-annotator consistency
    - variation in coding of -t/d deletion
  - the reduction of impediments facing new researchers or established scholars tackling broader issues

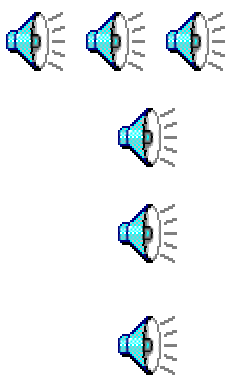
DASL Overview

- Currently
  - quantitative sociolinguistics is necessarily data-driven
  - huge stores of data exist, but most not publicly accessible
  - demands on individual researchers sometimes too high; corners are cut
  - current technology makes sharing data more attractive than ever before
    - speech community data can be compared with reasonable effort
    - broader investigations (multiple speech communities, regions) are possible

- Investigation of best practices in use of computer-based data & tools to support linguistic inquiry and documentation
  - multiple sites
  - large annotated data sets with platform-independent tools for access
  - encourage data sharing and related issues
    - inter-annotator agreement
    - data banks
  - case study

Case Study: Data

- Four LDC Corpora, created for linguistic technology development
- All data already transcribed, segmented to provide fine-grained access
- Basic speaker demographic information available (gender, age, education, region)



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

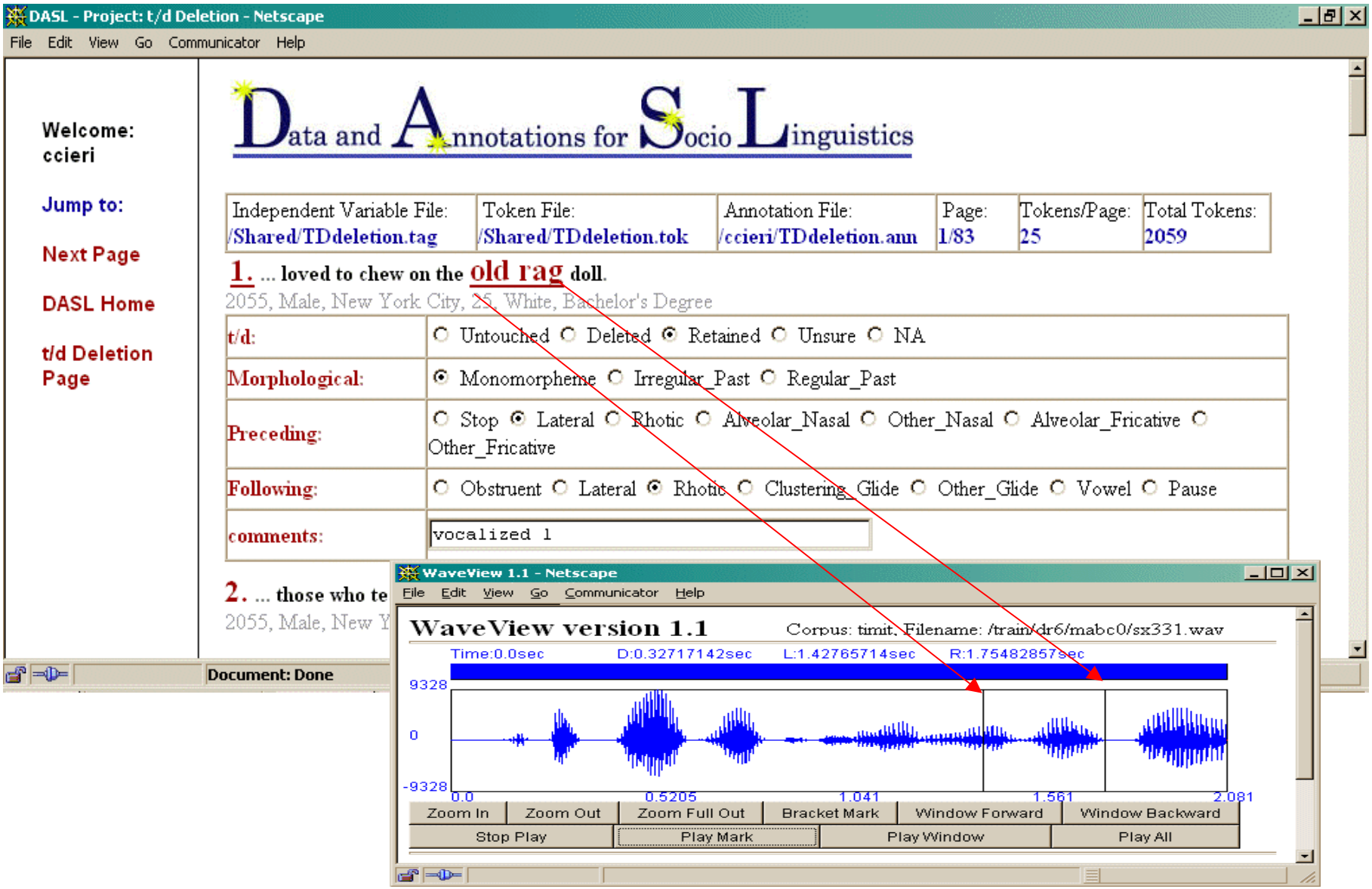
Case Study: Variable

- English -t/d deletion
  - best plans ~ bes' plans
- Well-documented and well understood, stable indicator
- Linguistic factors
  - morphological
  - preceding segment
  - following segment
  - stress, target segment, cluster complexity, word frequency, etc.
- External factors
  - education, age, region
  - style
- How does this data compare to traditional studies' results?

DASL Technology

- Create concordance
  - -regular expression search of corpus
- Create tag set
  - -specify which factors to code
- Create annotation file
  - -combines data with tag set
- Annotate using web browser
  - -play each example, tool supports common audio formats
  - -code factors in each factor group, adding comments when needed
  - -demographic information displayed
- Save results and output to text file
  - -can be imported to Excel Spreadsheet, Varbrul package, etc.

DASL Technology



- Penn Linguistics Colloquium - March 2001

Results: Overview

- TIMIT Corpus Overview
  - Corpus contains 6300 sentences; 54,387 words
  - Regular expression, unfiltered, produced 3154 tokens for consideration
  - With filters, 2059 tokens
  - Of these, 1578 were annotated for -t/d deletion (others were cases of N/A)

- Annotation (coding) specification
  - Roughly follows Guy (1980)
  - Linguistic
    - morphological, preceding & following phonological environments
  - Social
    - age, gender, education, region, race

Results: VARBRUL

- Summary
  - Tokens deleted: 518 (32.8%)
  - Tokens retained: 1060 (67.2%)
- First Run
  - difficulties with defining morphological factors
  - age, gender, region not selected
- Second Run
  - substantially similar to previous studies' results

Group	Factor	Factor weight	% Deleted	N
Morph	monomorpheme	0.535	38	1024
	irregular	0.531	20	41
	regular past	0.428	23	513
Preceding	alv nasal	0.756	53	432
	alv fricative	0.635	42	391
	other fric	0.433	25	73
	stop	0.426	23	244
	other nasal	0.390	16	25
	lateral	0.240	16	161
	rhotic	0.161	9	252
Following	obstruent	0.767	53	607
	rhotic	0.650	48	56
	clust glide	0.645	42	105
	lateral	0.380	29	17
	other glide	0.330	21	14
	pause	0.305	18	252
	vowel	0.245	14	527
Race	black	0.753	51	67
	other	0.552	27	15
	white	0.489	32	1455
	unknown	0.433	39	41
Education	unknown	0.752	58	31
	associates	0.616	43	56
	high school	0.524	36	207
	bachelors	0.514	33	876
	masters	0.436	29	350
	phd	0.357	22	58

Future Plans

- Dual Annotation
  - 5% of TIMIT re-coded by new annotator working independently
- Continue with annotation of SWB, other corpora as time/funding permits
  - additional factors
  - modify interface
- Other issues
  - categorizing style in four corpora
  - expand to include multiple sites
  - new data contributions from sociolinguists
  - new variables
  - feedback on methodology, tool
  - new data collections guided by insights from DASL project
- Follow progress at website
  - http://www.ldc.upenn.edu/Projects/DASL