Closer Still to a Robust, All Digital, Empirical, Reproducible Sociolinguistic Methodology

Christopher Cieri, Stephanie Strassel

University of Pennsylvania, Linguistic Data Consortium
ccieri@ldc.upenn.edu, strassel@ldc.upenn.edu
www.ldc.upenn.edu/Papers

History

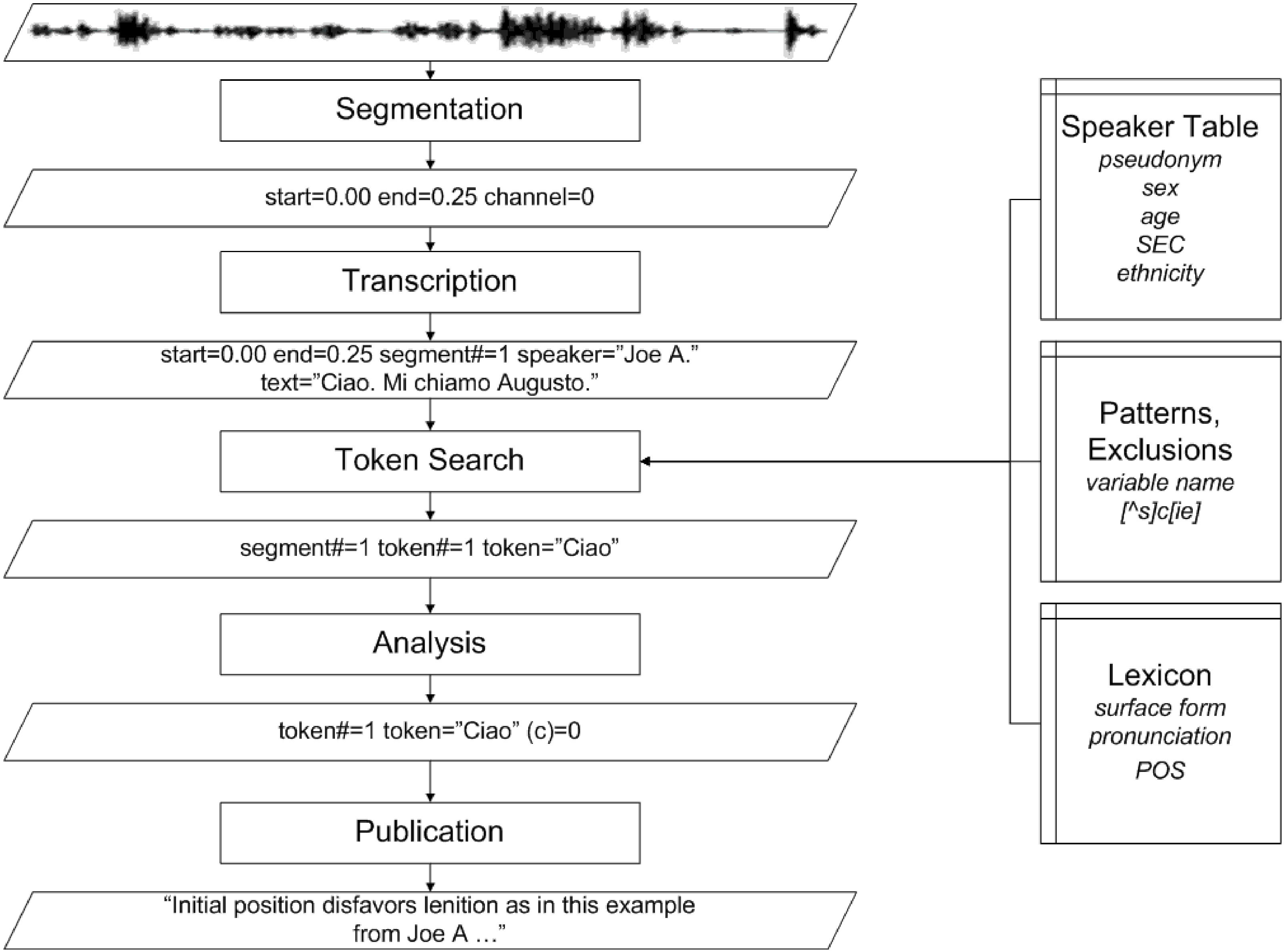
- 1963 Quantitative study of variation and change in the speech community has been intensively corpus based since inception
- 1971 Montreal Group began to create first computer based corpus for speech community study
- 1999 Gregory Guy convened a workshop on publicly available corpora, invited us to present on LDC corpora of potential use to sociolinguistics
- 2001 presented on corpus based sociolinguistics, our DASL project and the –t/d deletion study
- 2002 presented with William Labov on the SLx Corpus of classic sociolinguistic interviews and the DASLTrans
- 2003 organized Workshop at Penn of robust sociolinguistic methodology
- 2007 Malcah Yaeger-Dror convened workshop, invited Reva Schwartz, and MIT-LL and LDC to present on transcription practice and Phanotics project
- 2009 today we are very close to the realization of this ideal

Vision

- raw data – text, audio, video – is digital as are annotations, specifications
- transcripts other annotations are linked back to the original, raw data
  - time stamped for speech, linked via word offsets for text
- raw data or transcript proxy is computer searched for target variables
  - lexicons, speaker tables, other data external to recordings consulted as needed
- coding decisions are still made by humans
  - though the potential for partial automation exists
- variables, coding practice described to permit replication by others on the same or comparable data
- coding strings, examples in a paper, dots on a scatter plot or tracked backed to original recordings
- ideally data also publicly accessible.

Model

Model



Segmentation
start=0.00 end=0.25 channel=0
Transcription
start=0.00 end=0.25 segment#=1 speaker="Joe A."
text="Ciao. Mi chiamo Augusto."
Token Search
segment#=1 token#=1 token="Ciao"
Analysis
token#=1 token="Ciao" (c)=0
Publication
"Initial position disfavors lenition as in this example
from Joe A …"
Speaker Table
pseudonym
sex
age
SEC
ethnicity
Patterns, Exclusions
variable name
[^s]c[ie]
Lexicon
surface form
pronunciation
POS

Segmentation

- Virtually divide digital audio stream into manageable units
- Greatly facilitates downstream transcription, token retrieval, coding, analysis
- Can also indicate structural boundaries in recording
- Variable segment granularity to meet project needs
  - Maximum segment duration of 5-8 seconds makes downstream transcription and coding considerably more efficient
    - Sentence units (SU), breath/pause groups are convenient first-order units
    - Turns, discourse units, word, phones, etc. as optional second pass
- With right tools, SU or breath group segmentation can be performed in under 1.2x real time
  - Automatic segmentation, forced alignment with manual verification can also save time

Transcription

- Why a full transcription?
  - Index to speech, searchable
  - Provides stable basis for subsequent tasks
- Transcription specification to document conventions for orthographic representation
  - Use of standard orthography facilitates subsequent searching, retrieval of tokens, reanalysis
  - Specify treatment of common phenomena like disfluencies, non-standard forms, mispronunciations, transcriber uncertainty
- Transcription can be quite efficient given right tools combined with short audio segments

Comparison of Methods

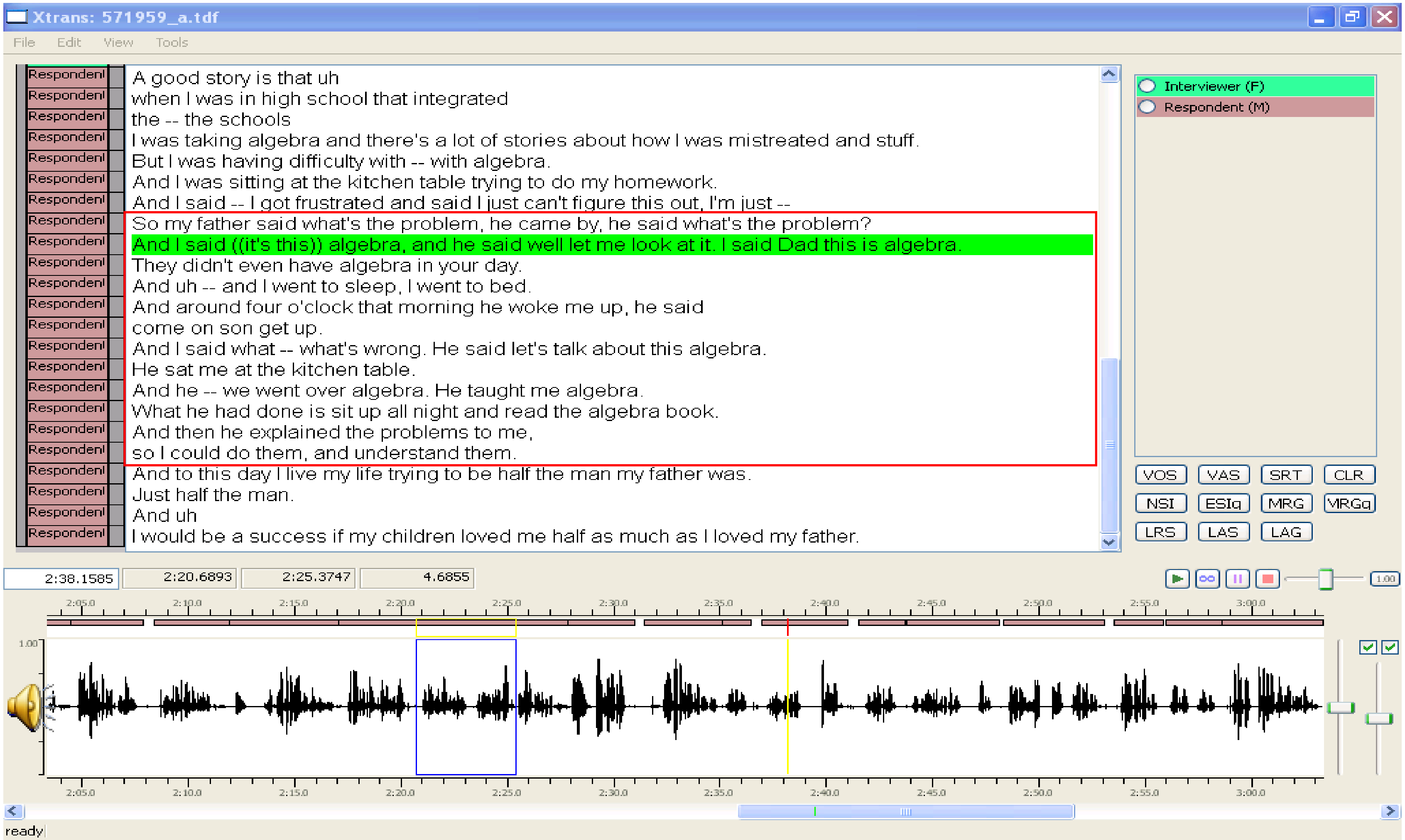
	Quickest			Most Careful
Segmentation	Automatic	Auto w/ verification	Manual	Manual w/ verification
Completeness	Content words	Add partial words, disfluencies	Add partial words, disfluencies	Add verification pass
Filled Pauses	Optional	Incomplete	Exhaustive	Exhaustive w/ verification
Disfluencies	None	Incomplete	Exhaustive	Exhaustive w/ verification
Transcriber Uncertainty	Flag and skip	Flag and best guess	Flag and best guess	Flagged best guess w/ verification
Feature Marking	None	Minimal	Full	Accurate, complete w/ correction
Speaker, Backgrnd Noise	None	Minimal	Exhaustive	Exhaustive w/ verification
Manual Passes	1	1-2	2-3	4+
Approx. Cost (x Real Time)	5 x	15 x	25 x	50 x

Comparison of Methods

	Quickest			Most Careful
Segmentation	Automatic	Auto w/ verification	Manual	Manual w/ verification
Completeness	Content words	Add partial words, disfluencies	Add partial words, disfluencies	Add verification pass
Filled Pauses	Optional	Incomplete	Exhaustive	Exhaustive w/ verification
Disfluencies	None	Incomplete	Exhaustive	Exhaustive w/ verification
Transcriber Uncertainty	Flag and skip	Flag and best guess	Flag and best guess	Flagged best guess w/ verification
Feature Marking	None	Minimal	Full	Accurate, complete w/ correction
Speaker, Backgrnd Noise	None	Minimal	Exhaustive	Exhaustive w/ verification
Manual Passes	1	1-2	2-3	4+
Approx. Cost (x Real Time)	5 x	15 x	25 x	50 x

Approximately 10x including segmentation

Quick Transcription Example



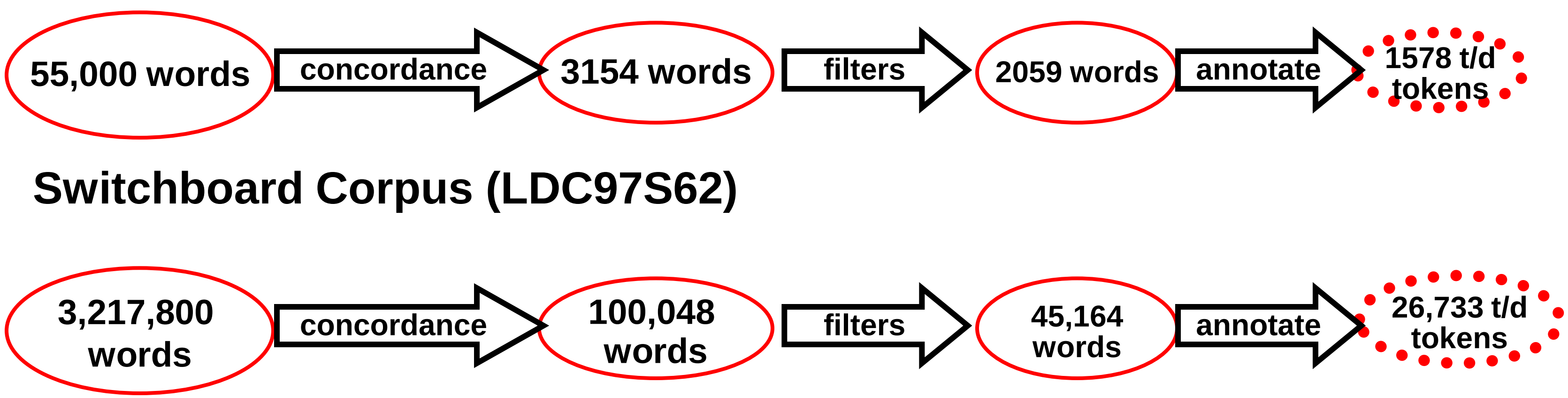
http://www.ldc.upenn.edu/XTrans

Token Selection

- Selection of tokens for analysis can be automated to large extent
  - Concordance to identify tokens of interest
    - Using string matching, regular expression queries
  - Filters to remove additional non-tokens
- More robust than manual selection, which might miss or overlook tokens
- Implemented in DASL t/d study

TIMIT Corpus (LDC93S1)

Switchboard Corpus (LDC97S62)



Coding Spec Challenges

- Difficulty of achieving perfectly explicit guidelines
  - Even when working on well-studied variable
- In DASL t/d deletion study, goal was to investigate comparability of corpus-based approaches with previous studies involving sociolinguistic interview data
- But previous t/d coding specs not typically published
  - Had to resort to personal communication with authors, detective work, reverse engineering from results
- Variation in coding for some factor groups inhibits direct comparison of results
  - Morphological factors, e.g. passives ("I was frightened")
- Some categories unmentioned - how were these coded?
    - Nasal flaps? Glottalized segments? What constitutes a pause?

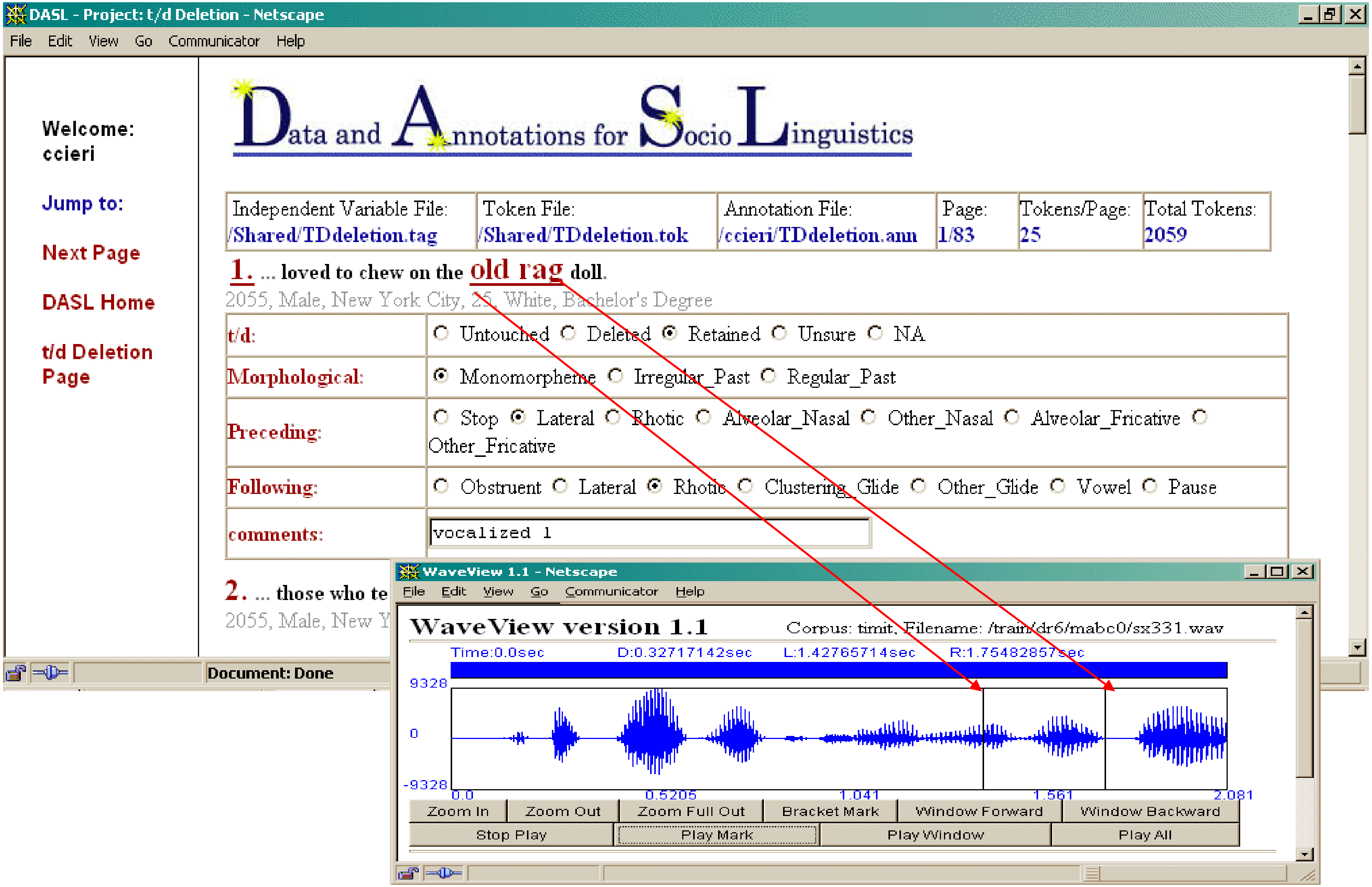
Coding Spec Best Practices

- Formal annotation/coding specifications promote coder reliability and direct comparison of results
- Developed iteratively over several rounds of pilot labeling including analysis of inter-coder reliability, via (double-blind) dual coding
  - Consider removal, merging of rules/categories with low consistency
- Written guidelines include
  - Title, date, version number
  - Introduction with framing/contextual info and general description of rule syntax
  - Screenshots of annotation/coding interface
  - Multiple examples for each rule
    - Including some difficult cases as well as counter-examples
  - Embedded sound files to illustrate application & non-application of rule
  - Appendix, glossary
  - Rules of thumb to promote consistent labeling
  - Can't tell, difficult decision flags
- (Link to) guidelines published along with results

Coding

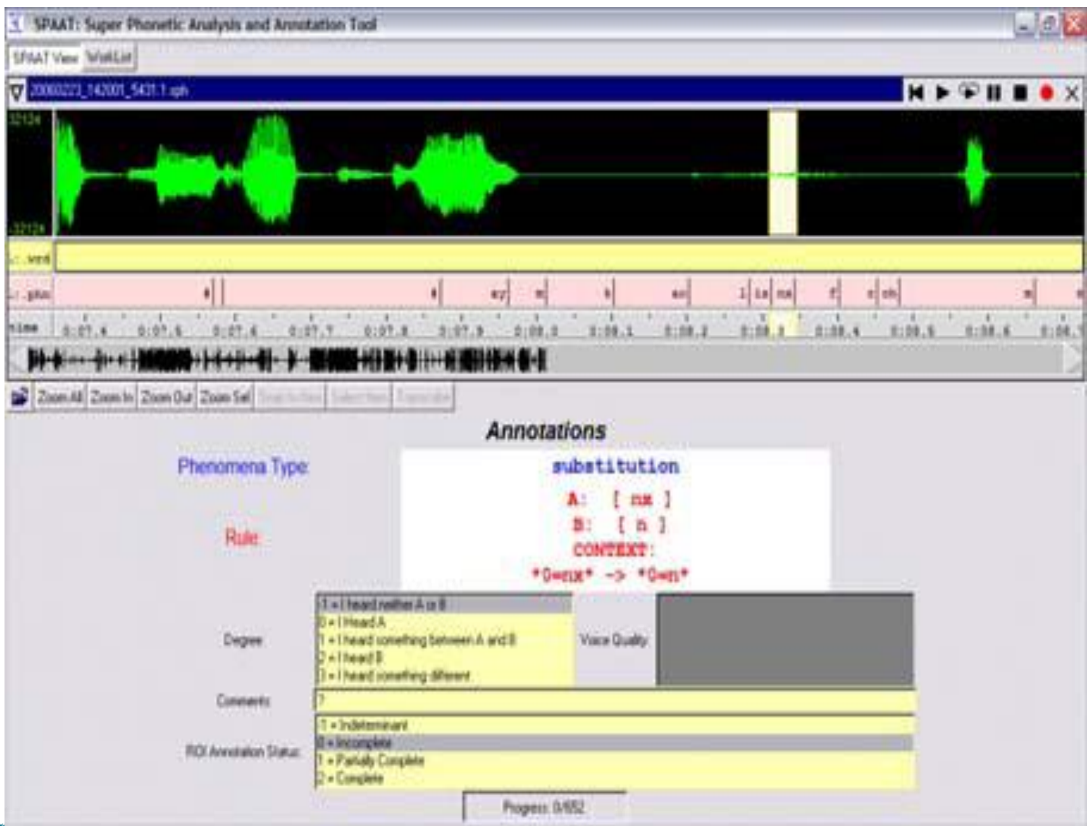
- Careful data preparation (segmentation, transcription) and pre-selection of all candidate tokens enables efficient coding
- "Regions of interest" already identified
- Attention directed at a single task: how is this variable realized in this batch of tokens
- Some customization of coding tools can increase efficiency further still

DASL t/d Coding Tool

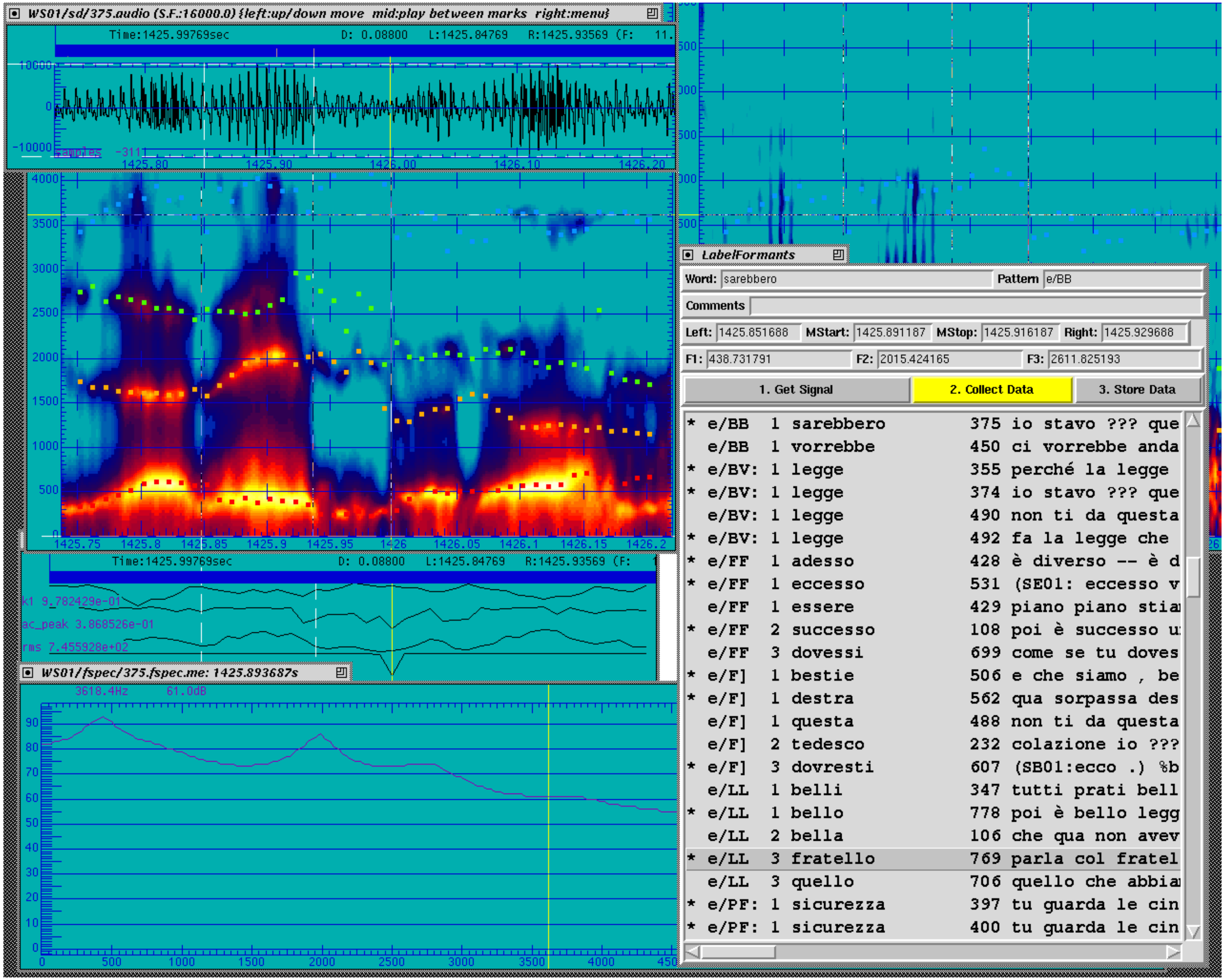


SPAAT (Super Phonetic Annotation & Analysis Tool)

- One variable, one ROI at a time
- Average of 250 judgments/hour, up to 400+ for experienced labelers



Formant Analysis



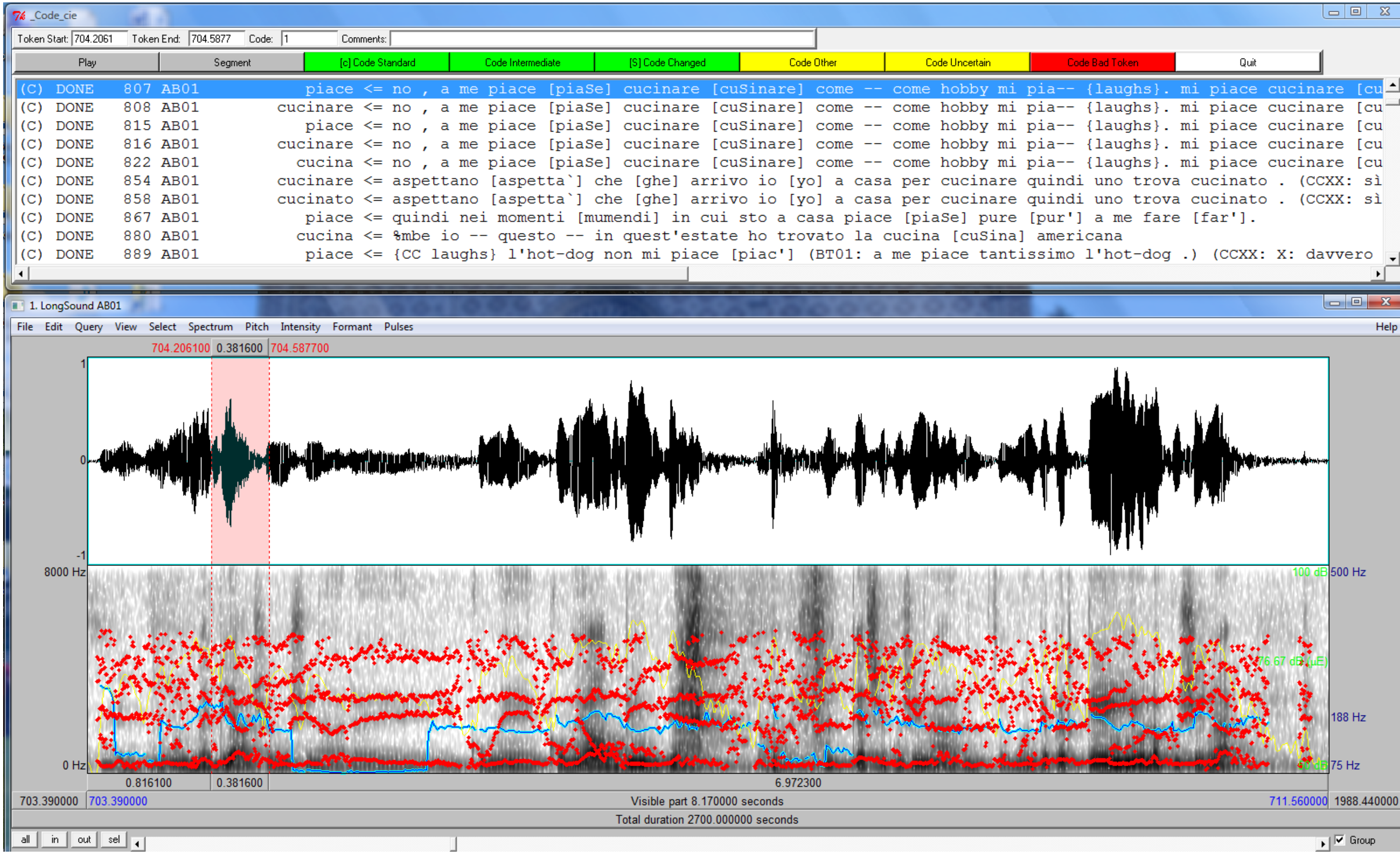
Token Selection

Vowel Segmentation

Identification of central tendency of word stressed vowel

Hand checking of formant tracker values for F1 and F2

Data Management



Vision

- ☑ raw data – text, audio, video – is digital as are annotations, specifications
- ☑ transcripts other annotations are linked back to the original, raw data
  - Xtrans, Praat, various Concordancers
- ☑ raw data or transcript proxy is computer searched for target variables
  - Ottawa Workshop, Montreal Project, SPAAT
- ☑ coding decisions are still made by humans
  - though the potential for partial automation exists
  - Yuan's Forced Aligner, Evanini's formant extractor
  - Other HLTs: ASR, Universal Phonetic Decoders, Energy Detectors, POS Taggers
- ☑ variables, coding practice described to permit replication by others on the same or comparable data
  - DASL Project, SLx,
- ☑ coding strings, examples, points on a graph tracked to original recordings
  - HTML <a> tags, Stefan Dollinger's Bank of Canadian English, Tom Veatch's 1993 dissertation
- ☑ ideally data also publicly accessible
  - Michelle Minnick-Fox, Nationwide Speech Project, NECTE Corpus

Journal of Experimental Linguistics

eLanguage Home About Log In Register Search Current Archives Announcements

Home > **Journal of Experimental Linguistics**

Journal of Experimental Linguistics

JEL is an interdisciplinary journal of reproducible research on topics related to speech and language. Regular publication will begin towards the end of 2009. For further details, see the announcements below.

The *Journal of Experimental Linguistics* is part of the Linguistic Society of America's eLanguage initiative. Like the rest of eLanguage, JEL is an Open Access online journal.

JEL is a linguistic "journal of reproducible research", that is, a journal of reproducible computational experiments on topics related to speech and language. These experiments may involve the analysis of previously published corpus data, or of experiment-specific data that is published for the occasion. Other relevant categories include computational simulations, implementations of diagnostic techniques or task scoring methods, methodological tutorials, and reviews of relevant new publications (including new data and software).

In all cases, JEL articles will be accompanied by executable recipes for recreating all figures, tables, numbers and other results. These recipes will be in the form of source code that runs in some generally--available computational environment.

Although JEL is centered in linguistics, we aim to publish research from the widest possible range of disciplines that engage speech and language experimentally, from electrical engineering and computer science to education, psychology, biology, and speech pathology. In this interdisciplinary context, "reproducible research" is especially useful in helping experimental and analytical techniques to cross over from one subfield to another.

Publication is in online digital form only, with articles appearing as they complete the review process. A rigorous but rapid process of peer review, designed to take no more than 4-6 weeks from submission to publication, will be supplemented by a vigorously-promoted system for adding moderated remarks and replies after publication.


LDC


SLX Corpus of Classic Sociolinguistic Interviews
Disc 1 of 1
slx
Linguistic Data
Consortium
LDC2003T15
1-58563-273-2
<PPortions © 2003 Trustees of the University of Pennsylvania. IPA93 Fonts © 2003 SIL International
LDC