Towards Best Practices in Sociophonetics:

Robust, Digital, Empirical, Reproducible Sociolinguistic Methodology

Christopher Cieri, Stephanie Strassel

Linguistic Data Consortium

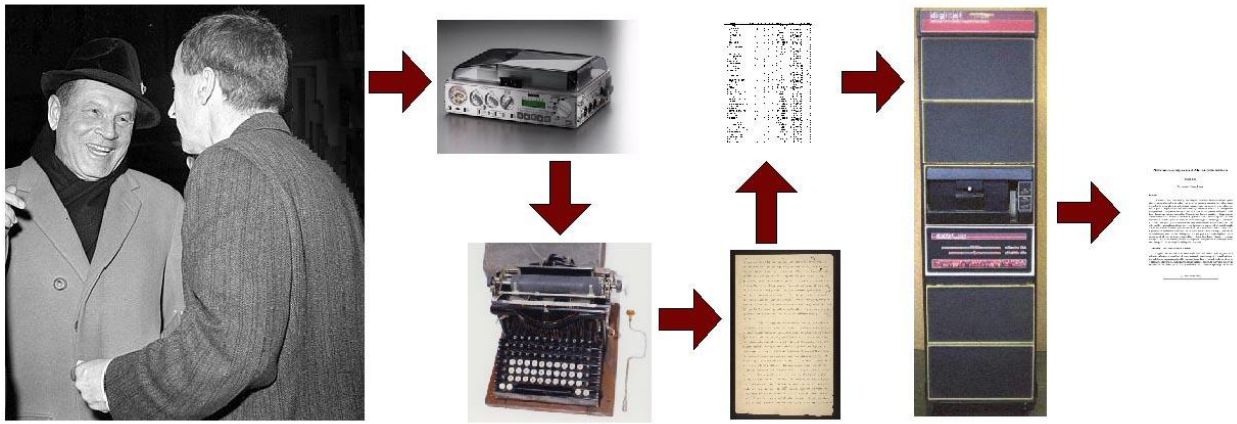
Linguistic Data Consortium

History

- 1963 Quantitative study of variation & change in speech community intensively corpus based since inception
- 1971 Montreal Group's first computer corpus for speech community study
- 1999 Gregory Guy's workshop on publicly available corpora
- 2001 LDC DASL project,–t/d deletion study
- 2002 William Labov's SLx Corpus and the DASLTrans
- 2003 Workshop at Penn of robust sociolinguistic methodology
- 2007 DiPaolo & Yaeger-Dror workshop with USSS, MIT-LL, Phanotics
- 2009 Update on methodology, Resulting paper

Evolution?

1963



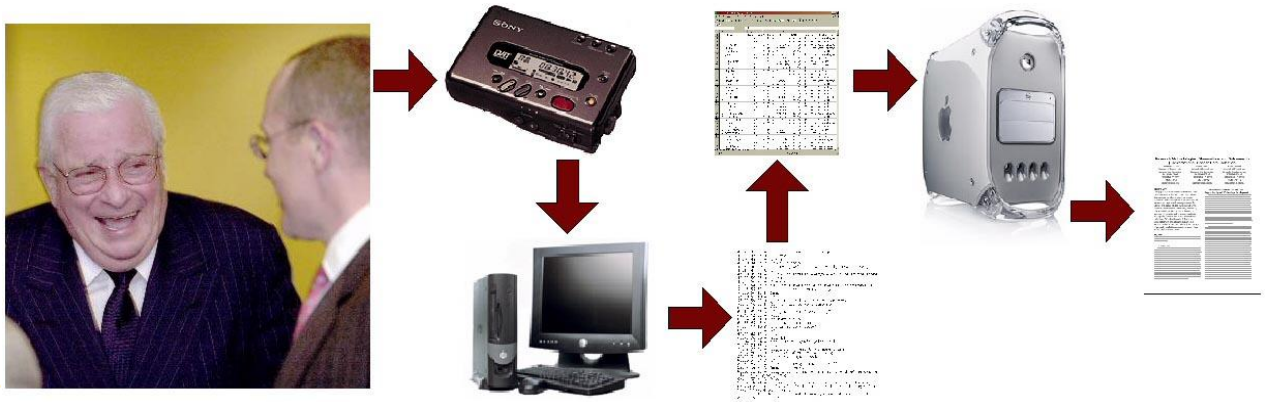
Interviews are recorded but not always transcribed; when transcribed, transcripts are often only partial.

Analytical tools are not integrated.

The presentation is an independent artifact.

After nearly 40 years of technological advance, our use of data is largely unchanged; only the components differ.

2003



Methods

- Original
  - listen to recording for interesting tokens, possibly digitize them
  - code tokens marking on score sheet
  - reformat data for statistical analysis
  - analyze
  - write-up citing examples where appropriate
- Proposed
  - digitize entire session, integrate other sources of data
  - segment, transcribe, align
  - integrate dictionary and demographic information
  - query transcript for tokens
  - code and analyze
  - write-up including direct citations to original and coded data

Linguistic Data Consortium

Suboptimal Methods

- slow & labor intensive
  - thus discouraging
- susceptible to distraction
  - missed tokens
  - unbalanced view of corpus
- redundant coding
  - of independent variables based on word class
- lose sequence and time of utterances, events
- ignore the style profile of an interview
- effort for reanalysis nearly equal to effort for original
- only limited opportunities for re-use or sharing

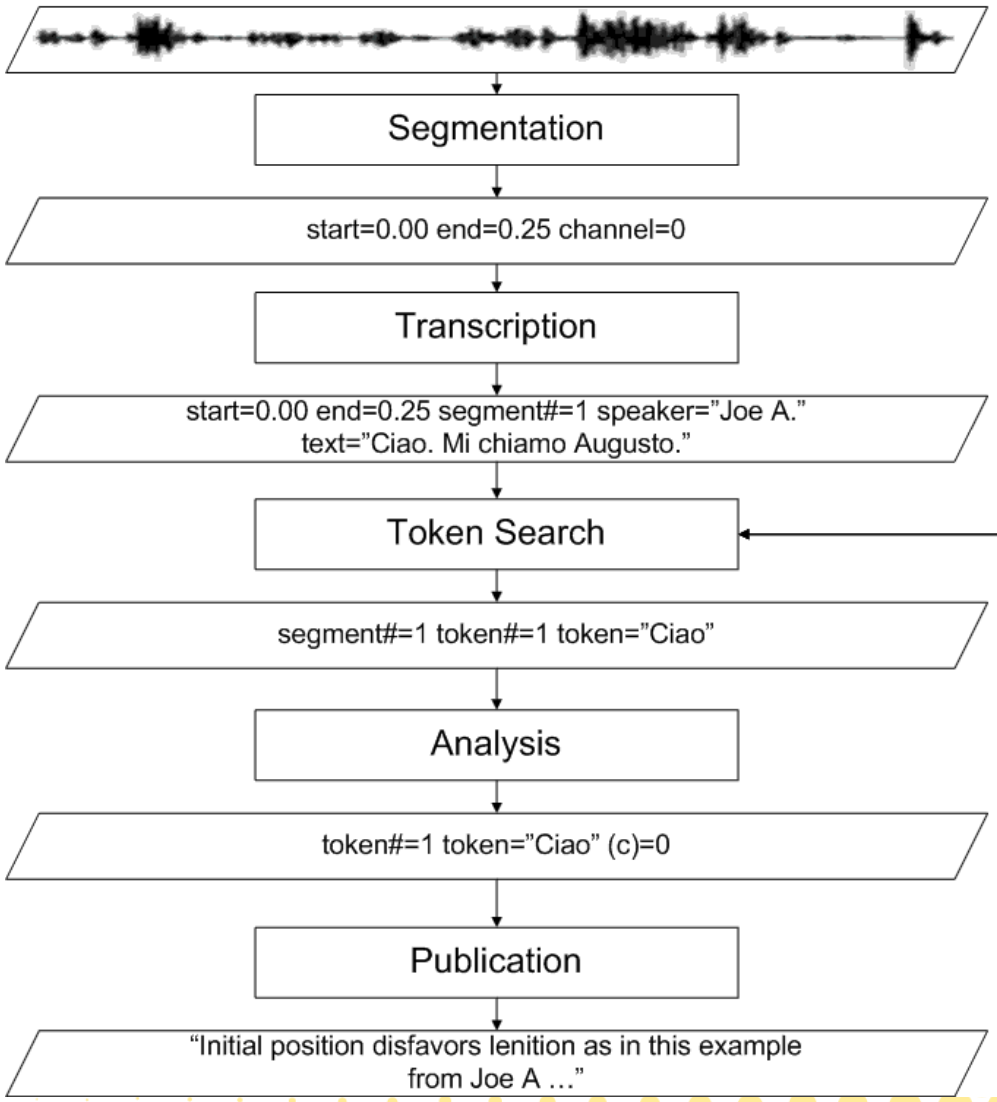
Optimal Methods

- make coding efficient allowing researchers to
  - consider greater percentage of tokens/variable
  - investigate more variables
- minimize misses
  - improve accuracy and balance
- improve consistency
- retains accurate time and sequence information
- retains mapping among sound, transcript, tokens, coding, analysis and examples in publication
- encourages re-use of data
  - each additional pass requires less effort than original
  - re-use & reanalysis profits from previous preparation

Goal

- ☑ raw data – text, audio, video – are 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
- ☑ data publicly accessible for education, research and technology development
  - Michelle Minnick-Fox, Nationwide Speech Project, NECTE Corpus

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

Linguistic Data Consortium

Build or Borrow?

- Original fieldwork will always be necessary, providing
  - valuable researcher training and experience
  - appreciation for the challenges of fieldwork
  - in-depth knowledge of the speech community
  - coverage of new communities and language varieties
  - new methodological perspectives
  - potential new contributions of data to public archive

- Today we'll talk mostly about building
- But note that LDC now offers data at $0 cost to
  - impecunious students
  - with a bona fide need

Linguistic Data Consortium

Build or Borrow?

- Corpus-based approaches complement first hand fieldwork
  - replication of methods, stable benchmarks for
    - competing approaches
    - comparison of results across studies & over time
  - re-annotation and reuse for new purposes
  - reduces impediments facing new researchers
    - exploration prior to fieldwork
    - lower cost, greater accessibility
  - allows established scholars to tackle broader issues
  - demonstrates best practice in corpus creation
    - serves as a teaching tool
    - measurement of inter-annotator consistency
  - allows for multi-site collaboration
  - greater volume in case of rare phenomena
  - new perspective

Specifications

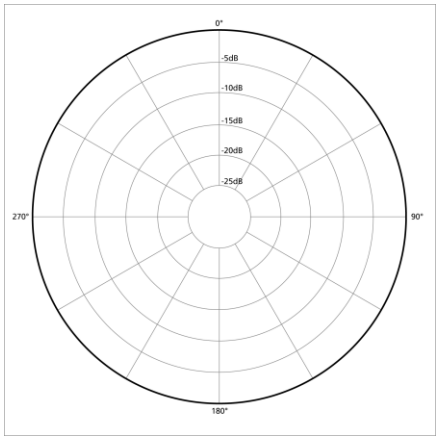
- Linguistics = Language Science
  - Sciences are supposed to be reproducible
  - In order for a study to be reproducible, method must be carefully documented!
  - difficulty to achieve perfectly explicit guidelines even when working on well-studied variable
- DASL -t/d deletion study
  - goal: compare corpus-based approaches to previous work 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
- Differences in coding inhibits direct comparison of results
- Some categories unmentioned - how were these coded?
    - What constitutes a pause?

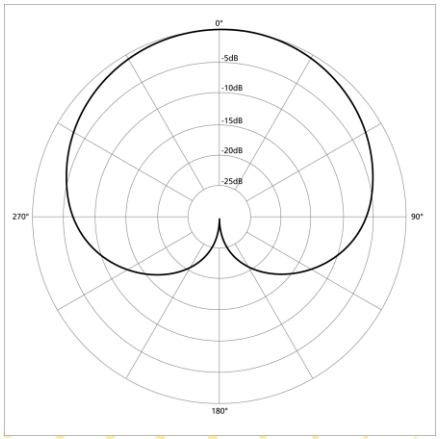
Collection

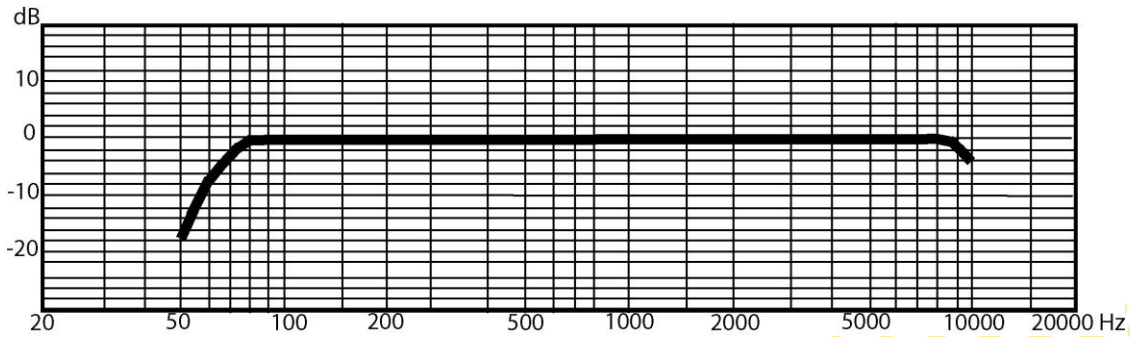
- Imponderables
  - temperature, medium treated as fixed
  - speakers not selected for ability to sit still and speak clearly
- Sometimes Controllable
  - external noise
  - reflection
  - distance
    - subject to microphone
    - subject to interviewer

Collection

- Controllable
  - microphone type: probably condenser
  - polar pattern: omni-directional versus cardioid
  - form factor/mounting: probably lavaliere
    - ≤20cm, ≥15cm if directional
    - on the lapel, not the collar or placket
    - not in the shadow of the chin
    - not directly in front of the mouth
  - frequency response







Recorders

- Desiderata
  - adequate quality @ affordable price
    - standard digital format, ≥16-bit samples, ≥16kHz sampling
    - uncompressed, nonproprietary allowing universal random access
  - standard data interface for moving speech files to computer
  - small, unobtrusive, very portable
  - simple to use
  - adequate storage and battery life for 1 entire day in the field
  - monitors for battery life, remaining storage, level, clipping
  - 2 channels with separate adjustments
  - solid-state
  - compatible with the microphones
    - connector type (trs, xlr), power protocol (plug-in, phantom)

Recorders

- Sampling Rate
  - ≥16kHz
- Sample Size
  - ≥16 bits if appropriate given source, e.g. less needed for telephone
- Compression
  - Why risk it?
- Storage
  - sampling rate * sample size/8 per second
    - 96,000 * 24/8 * 60 * 60 = ~1GB/hour
- Analytic Software Requirements

Recorder Test

- single TIMIT sentence with 25dB gain
- played through speaker at consistent volume
- same room, same time of day in each case
- microphones placed at
  - 8”: lavaliere
  - 12”: table top near subject
  - 36”: table top near interviewer
  - 144”: window sill
- recorders on factory default settings
  - Zoom H2 & H4, Marantz PMD620, Tascam DR-100
  - Built-in mic
  - Sound Pro SP-CMC-2 (dual AT-831) wired lavalier cardioid electret
  - Shure 183 omnidirectional, cardioid

Recorders

H2

H4

DR-100

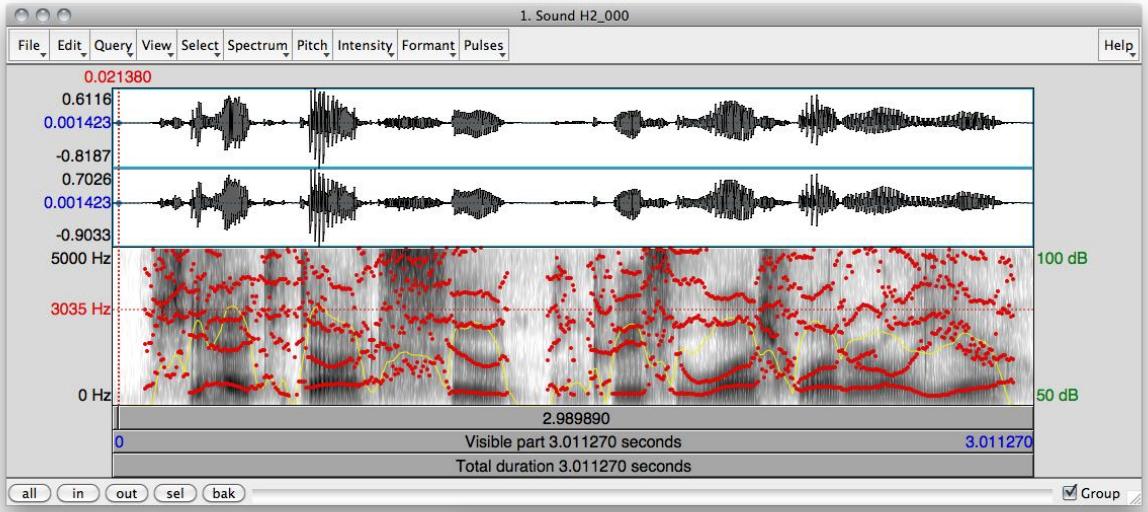
PMD620

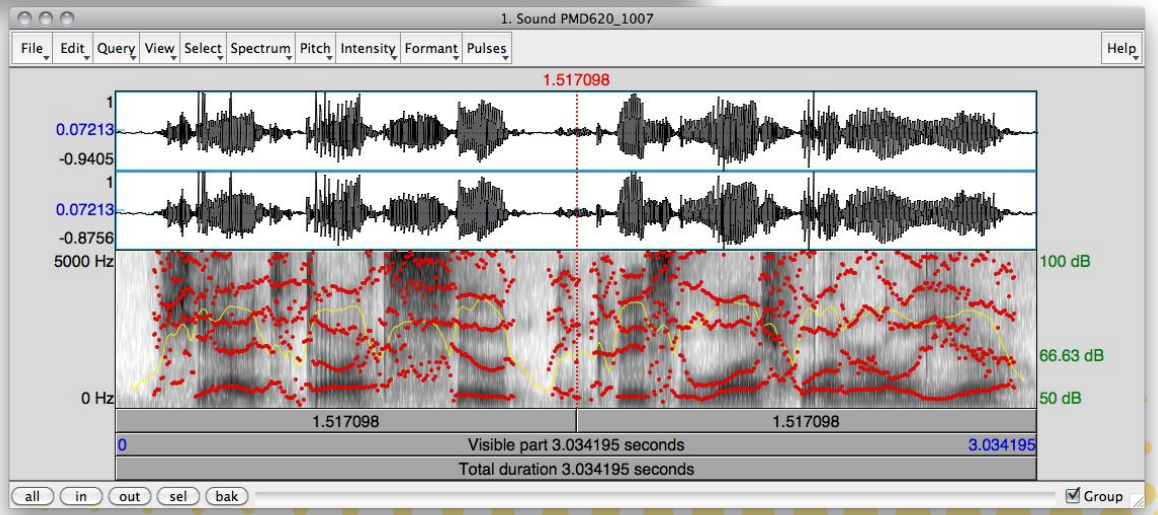
Recorder Test Results

- quality generally very good
- factory settings slightly too sensitive for test case
  - some clipping

Recorder Test Results

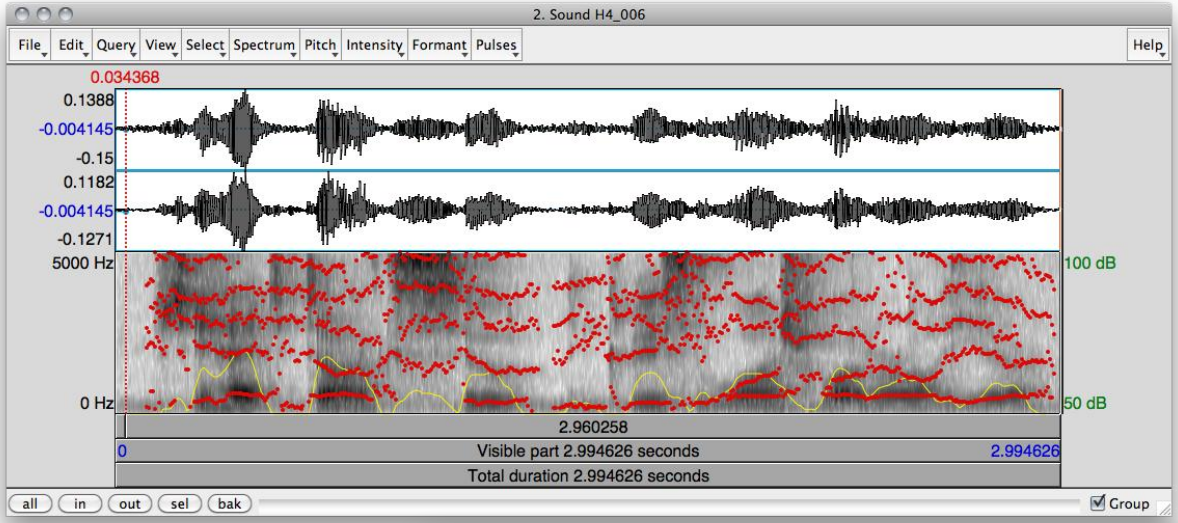
- inexpensive recorders, well placed produce good results

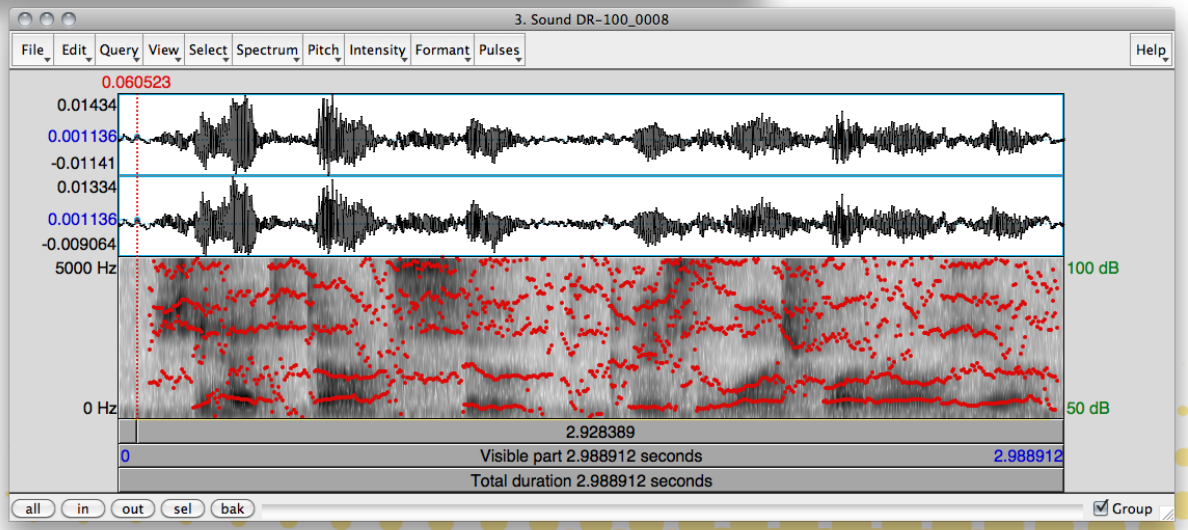




Recorder Test Results

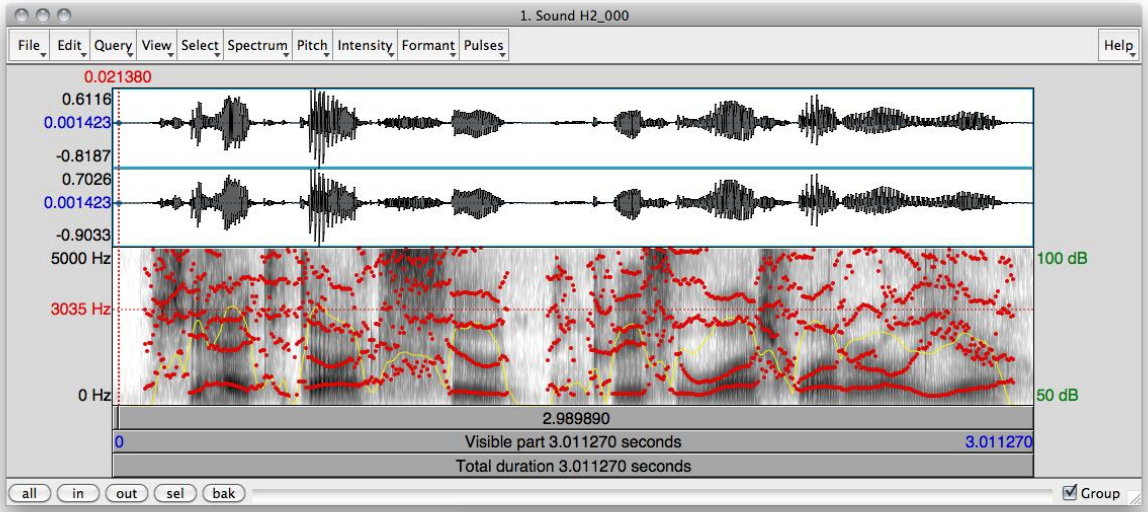
- expensive recorders poorly placed produce poor results

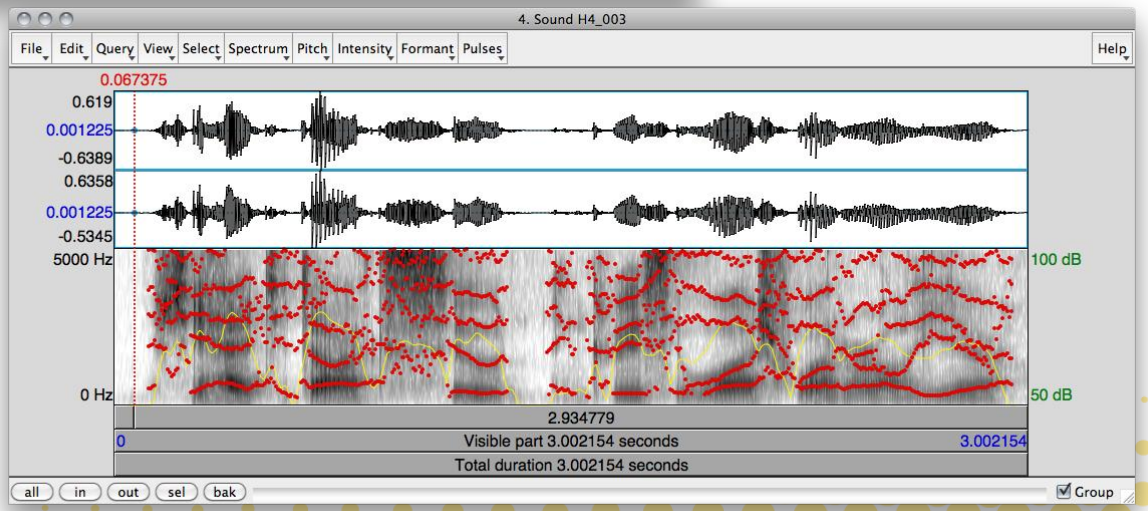




Recorder Test Results

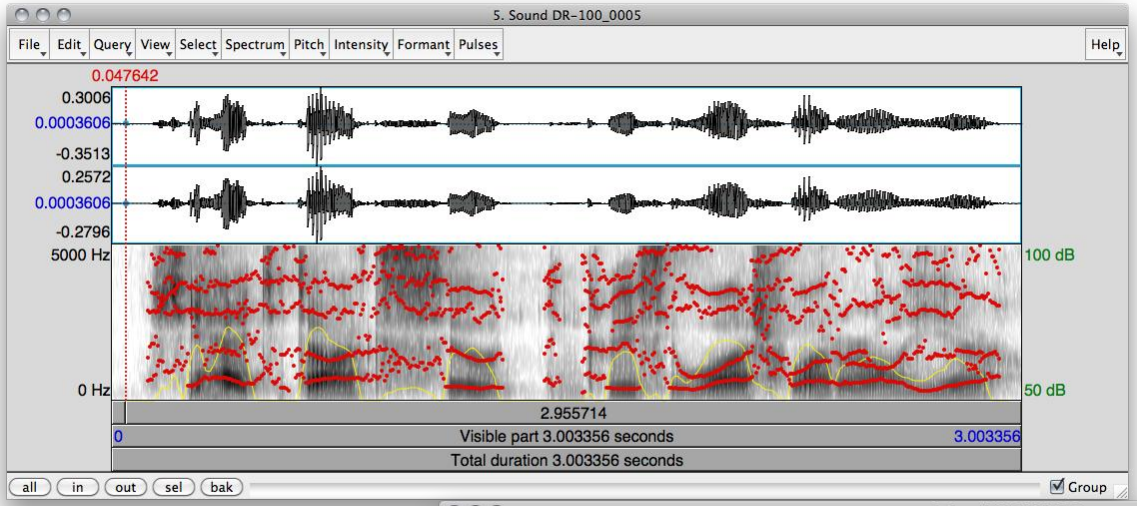
- expensive recorders may not warrant extra cost

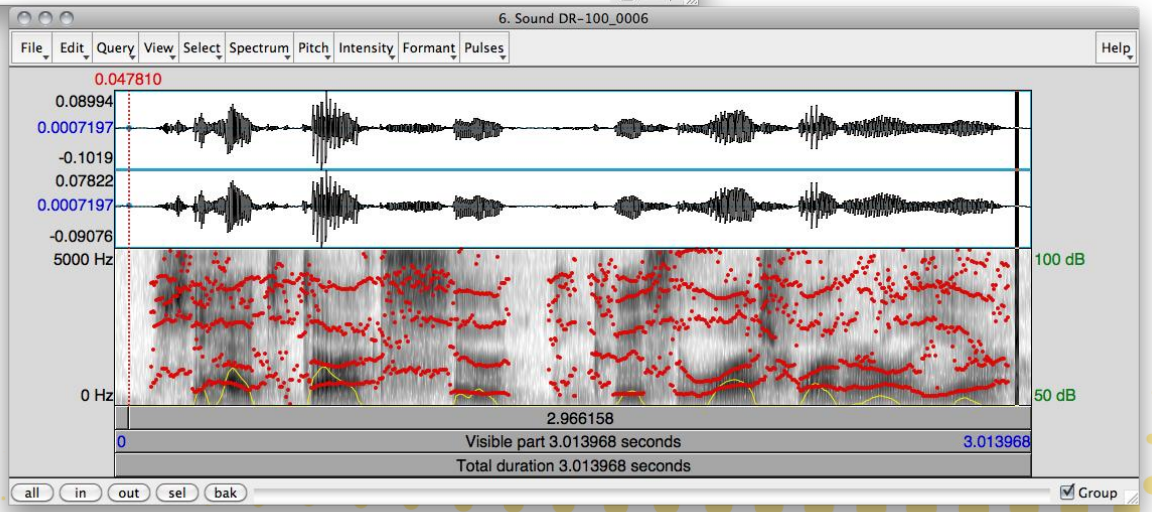




Recorder Test Results

- difference between unidirectional and omnidirectional slight





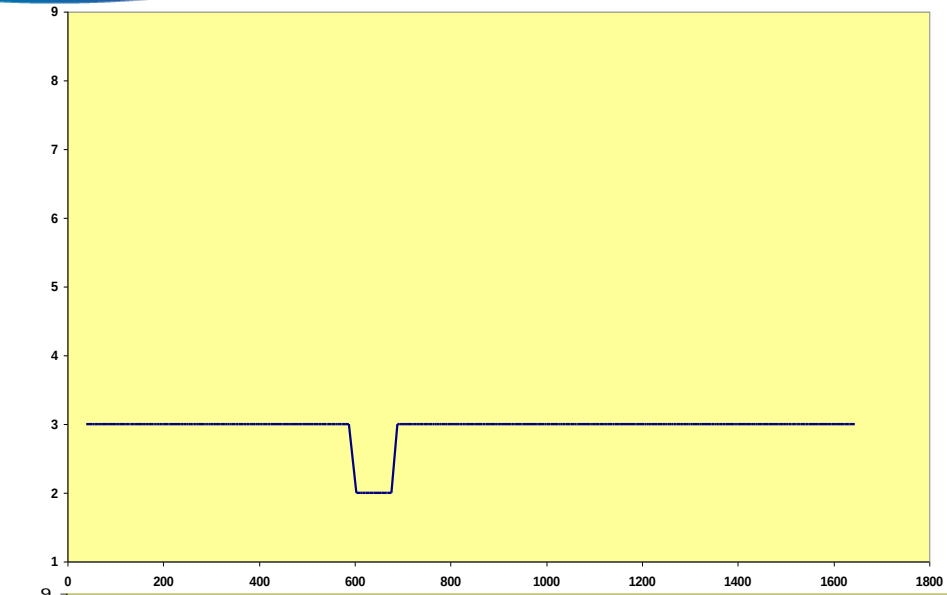
Segmentation

- Divides corpus into manageable units
  - indicates structural boundaries in recording
  - provides time-alignment for transcripts and other annotations
    - transcript becomes index to audio
  - simplifies subsequent transcription, token selection, processing, analysis
    - ≤8 seconds for transcription, FA runs better, Praat can display
- Preserve integrity of original signal
  - virtual, not actual, chopping of digital signal
  - allows multiple segmentations of the same event
- Speech Activity Detection (SAD) technology
  - exists for some audio types (LDC has telephone, BUT has broadcast)
  - segments by pause group
  - need training material (segmented, representative sociolinguistic data)

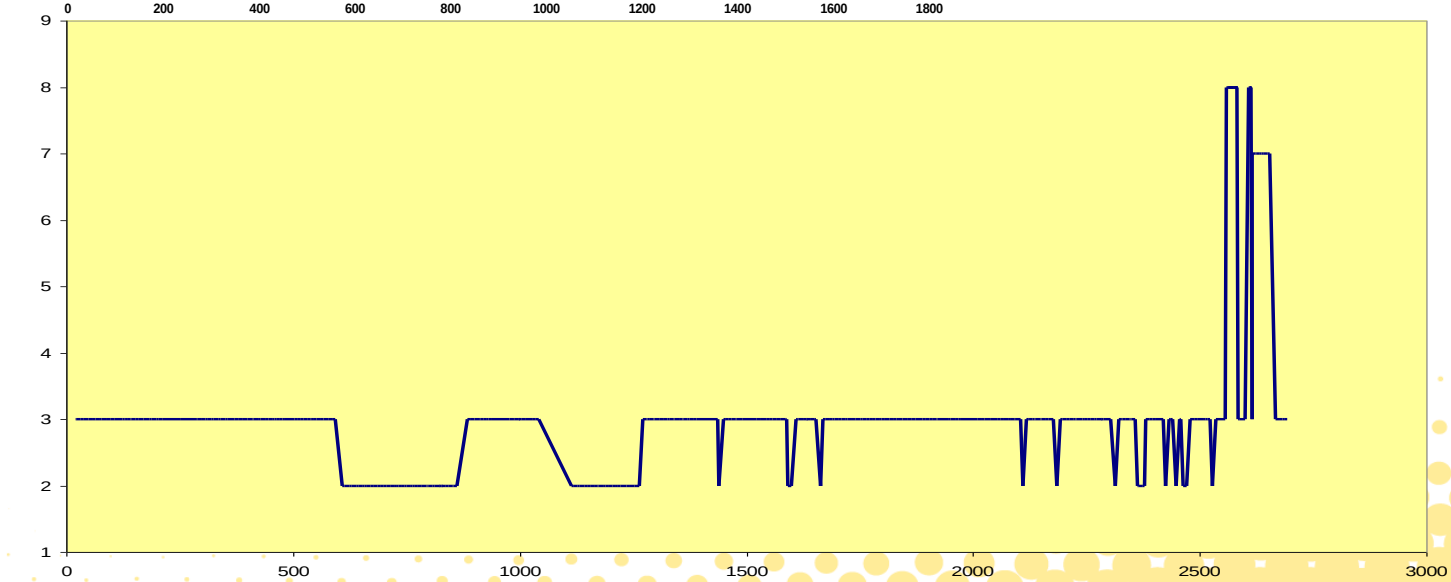
Segmentation

- Segmentation for a specific purpose
  - speaker turn, breath/pause group (1xRT), utterance, SU (≥5xRT)
  - word level, phone level best handled as additional pass
    - imparts additional level of analysis
    - more difficult/costly, requires specialists
    - “free” with forced alignment
- Issues
  - levels of granularity
  - multiple speakers on one channel
  - overlapping speech even across channels
  - how long is a pause?
  - additional features: background, non-speaker noise, SID, style

Time as Variable



Time is on the horizontal axis.
Conversational situation (style) is on the vertical.
Larger numbers mean greater formality.
4+ are elicited styles
3 is the default interview situation
2 is for narratives and extended descriptions
1 is for speech to another party
The longer interview clearly provides greater opportunities to study style shifting!



Transcription

- Stoker ’97 provides early justification for transcription in related field

Transcription

- Stoker ’97 provides early justification for transcription in related field

> *He accordingly set the phonograph at a slow pace, and I began to typewrite from the beginning of the seventeenth cylinder.*
>
> *He thinks that in the meantime I should see Renfield, as hitherto he has been a sort of index to the coming and going of the Count. I hardly see this yet, but when I get at the dates I suppose I shall. What a good thing that Mrs. Harker put my cylinders into type! We never could have found the dates otherwise.*
>
> *Stoker, Bram (1897) Dracula*

Transcription

- Why transcribe?
  - index to audio, intermediary to later coding
  - searchable
- How to transcribe?
  - verbatim
  - no “correction”
  - standard orthography, punctuation
  - conventions for
    - unintelligible speech
    - non-standard variants
    - speaker restarts, disfluencies, hesitations
  - 7-10xRT using Transcriber, Xtrans

Linguistic Data Consortium

Transcription

- Multiple passes focusing on different tasks
  - limit cognitive load of any one pass
  - tasks
    - basic text
    - disfluencies
    - conversational situation
    - dialect phenomena
    - personal identifying information
    - phonetics (inter-annotator agreement 70-90%)

Automatic Speech Recognition (ASR)

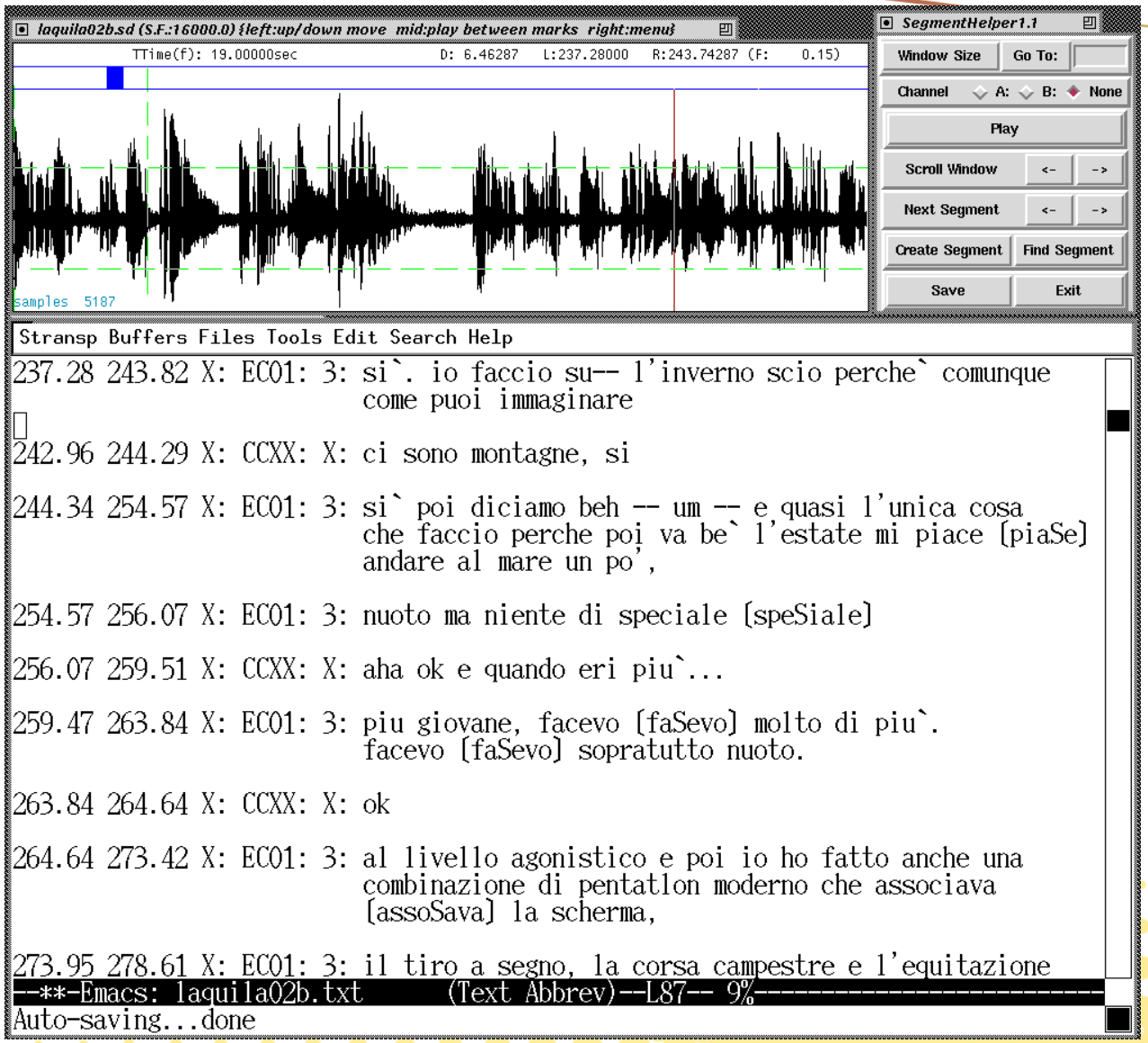
- ASR Mediated Transcription experiment
  - native speaker trained Dragon Naturally Speaking Italian
  - listened to tapes via foot-pedal controlled device
  - repeated each utterance to Naturally Speaking & corrected its mistakes

	ASR	**Manual**
Experiment 1	13.1xRT	13.4xRT
Experiment 2	11xRT	7.8xRT

- ASR
  - sensitive to channel
  - need to be trained for linguistic variety
  - targets of sociolinguistic study typically not those of ASR
  - See Speech Processing: Interactive Creation and Evaluation Toolkit
    - http://cmuspice.org/, Prof. Tanja Schutz, CMU

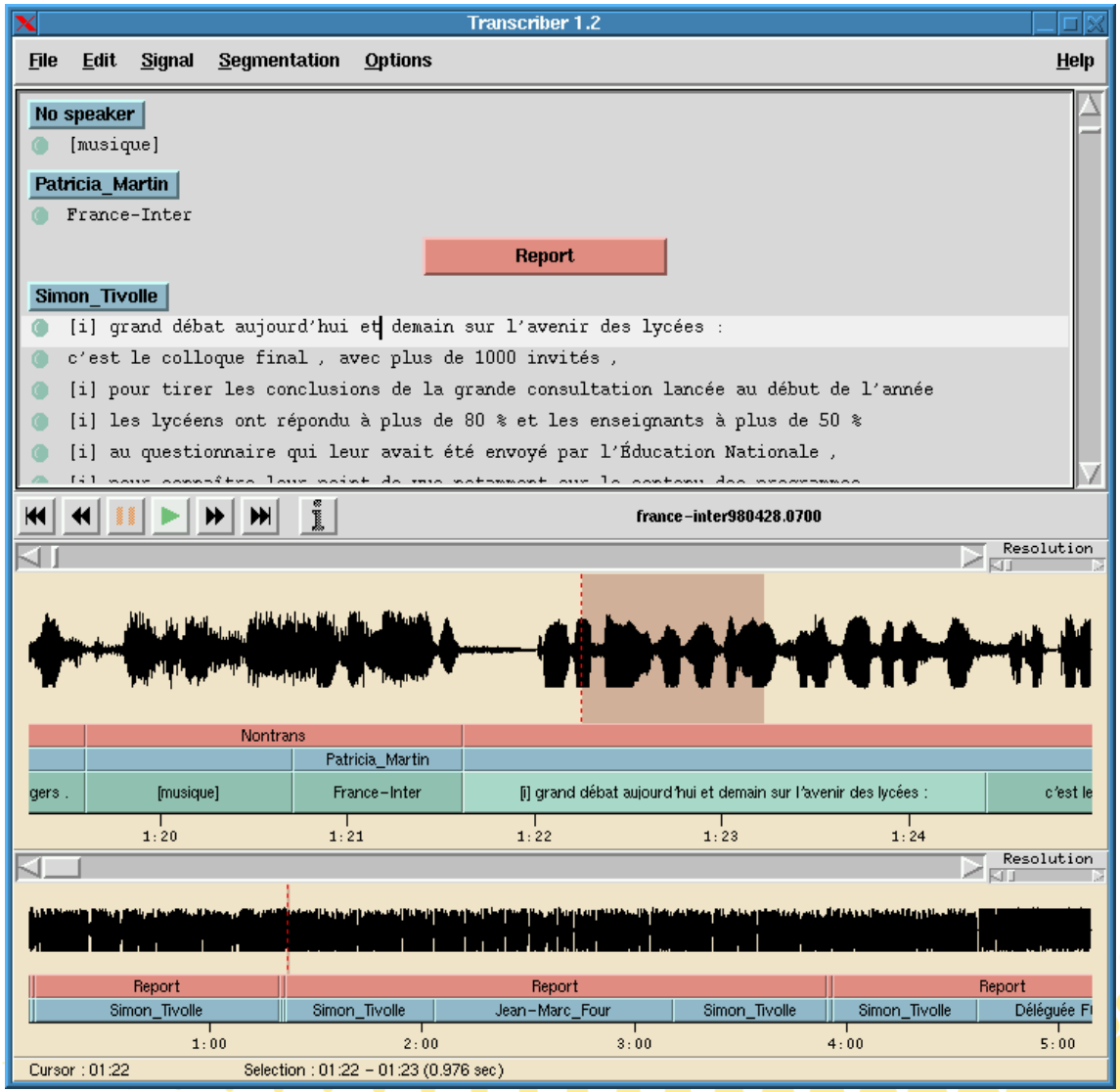
Linguistic Data Consortium

Strans +



Transcriber

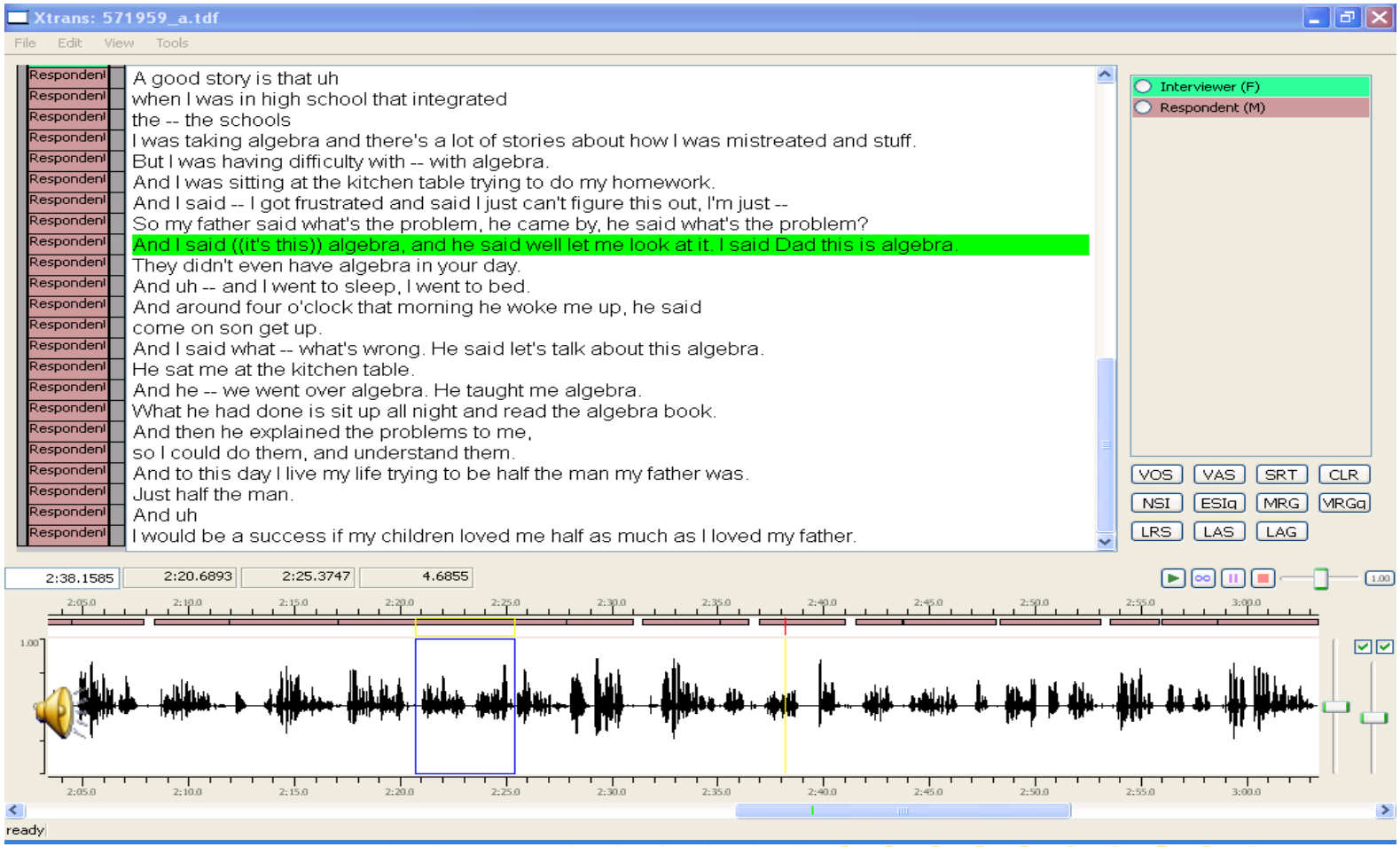
http://trans.sourceforge.net/en/presentation.php



- fastest segmentation
- More user friendly than strans
- Linux, Windows, OSX
- open-source
- multiple audio, text formats
- requires full segmentation of audio
- built for single-channel broadcast news
- handling of overlapping speech

XTrans

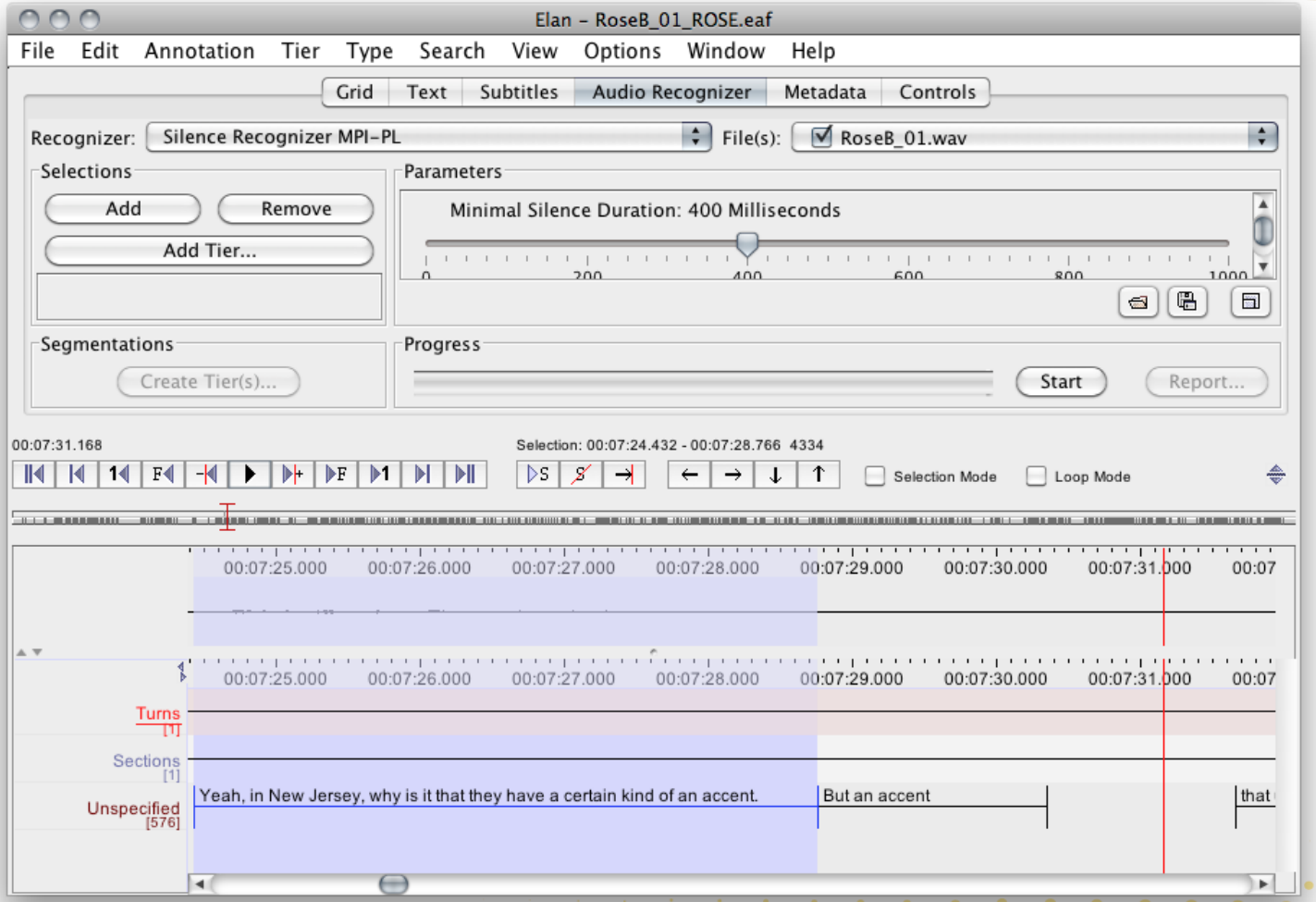
http://www.ldc.upenn.edu/tools/XTrans/



- fast segmenting, multi-channel, -speaker, overlaps, reads Transcriber, SPH
- Linux, Windows, OSX (in emulation)

Elan

http://www.lat-mpi.eu/tools/elan



- video, reads Transcriber, SPH, interacts with Praat, Linux, Windows, OSX
- segmentation complex

Token Selection

- What parameters drive token selection?
  - phonological, morphological, lexical, syntactic
  - balance across extra-linguistic features
  - But are there hidden parameters?
    - Convenience
    - Time
    - Fatigue
- Incomplete coverage, lack of balance damages research
- Variation across studies reduces ability to compare results
- Pronouncing dictionaries can mediate token selection
- What do we know about time as independent variable?

Linguistic Data Consortium

Token Selection

- Selection of tokens for analysis can be automated to large extent
  - concordance to identify tokens of interest
  - string matching or regular expressions
  - lexicons to mediate
  - filter to remove additional non-tokens
- In DASL –t/d deletion Study
  - p_{token} in TIMIT 2.9%, smart token selection removed 99% of non-tokens
  - p_{token} in Switchboard 0.8%, smart token selection removed 99.4% of non-tokens
  - Smart token selection all these two large corpora to be coded for –t/d delection **in their entirety**
- substantially reduces overall effort
- ensures desired coverage

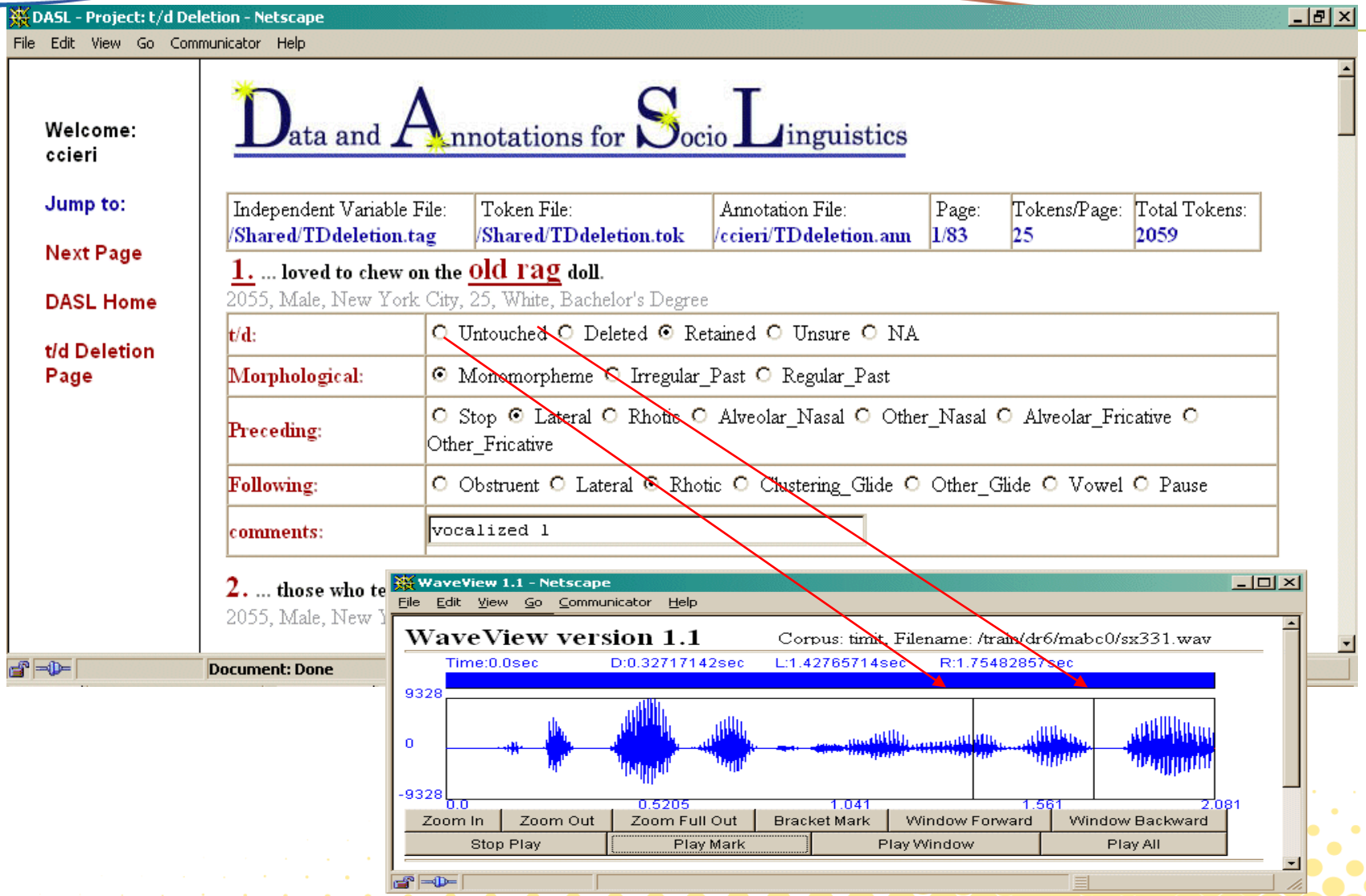
Coding

- Careful data preparation
  - segmentation
  - transcription
  - pre-selection of candidate tokens

enables efficient coding

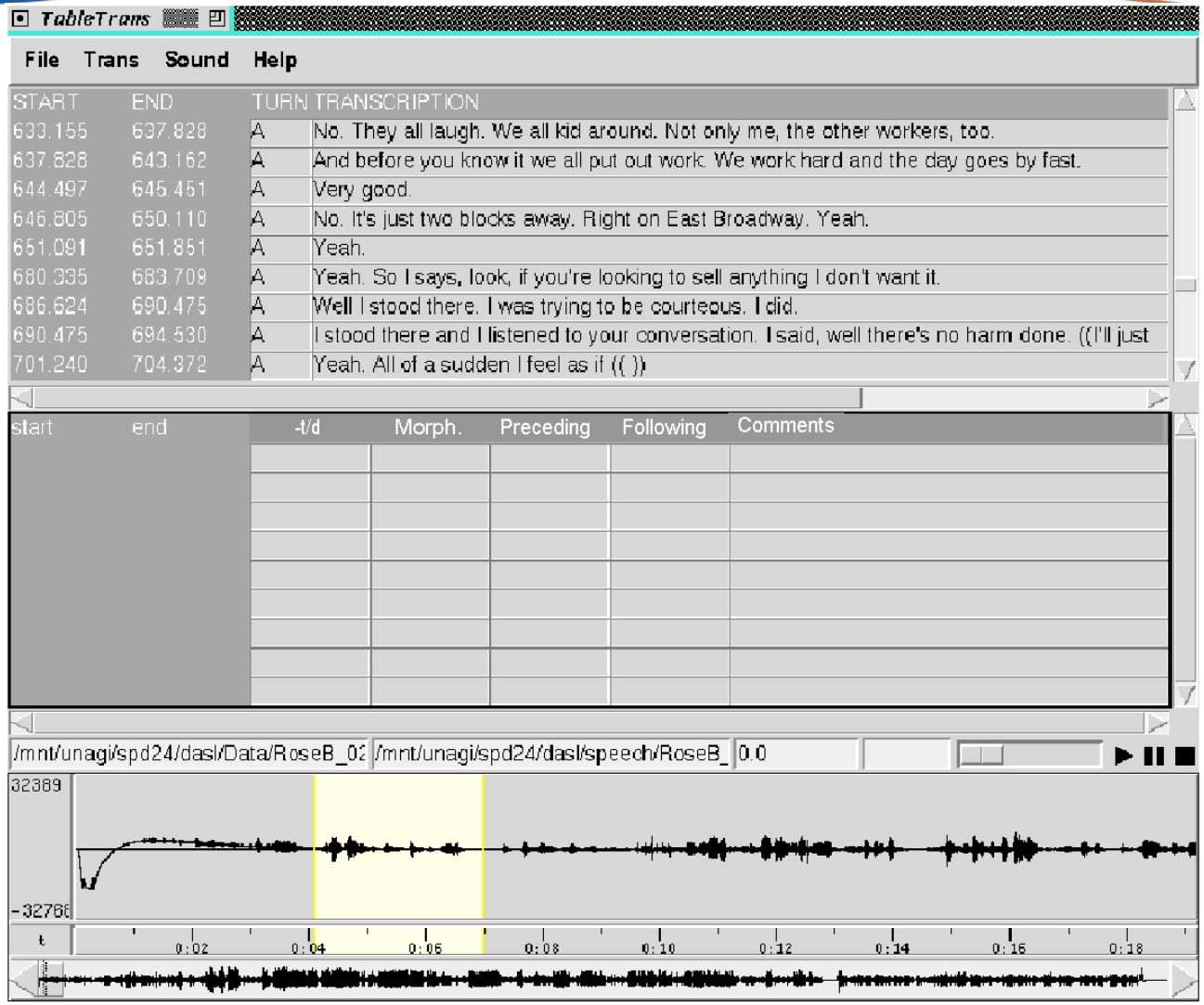
- Attention directed at a single task: how is this variable realized in this batch of tokens
- Coding decisions connected back to transcript and audio

DASL –t/d Deletion Coding

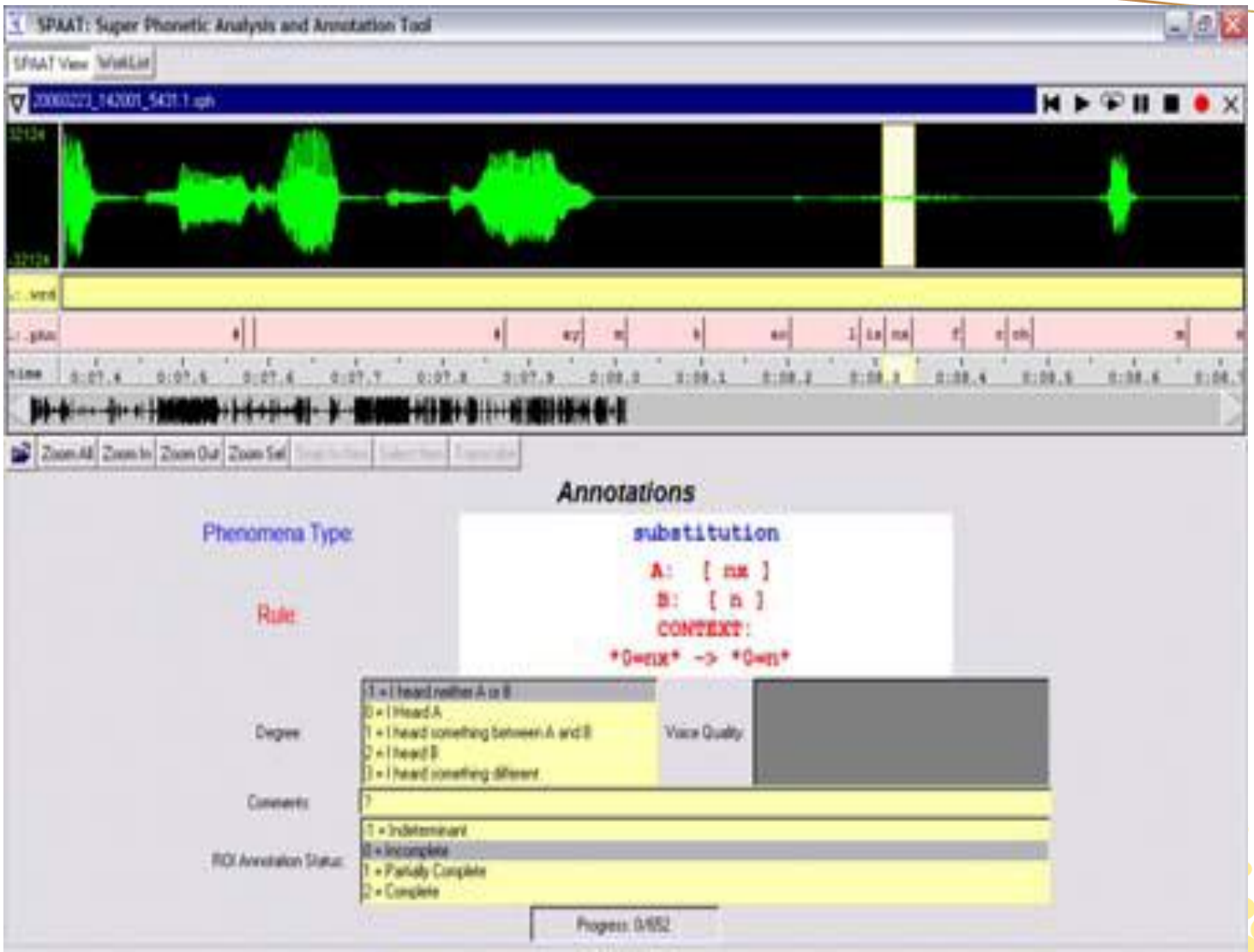


Linguistic Data Consortium

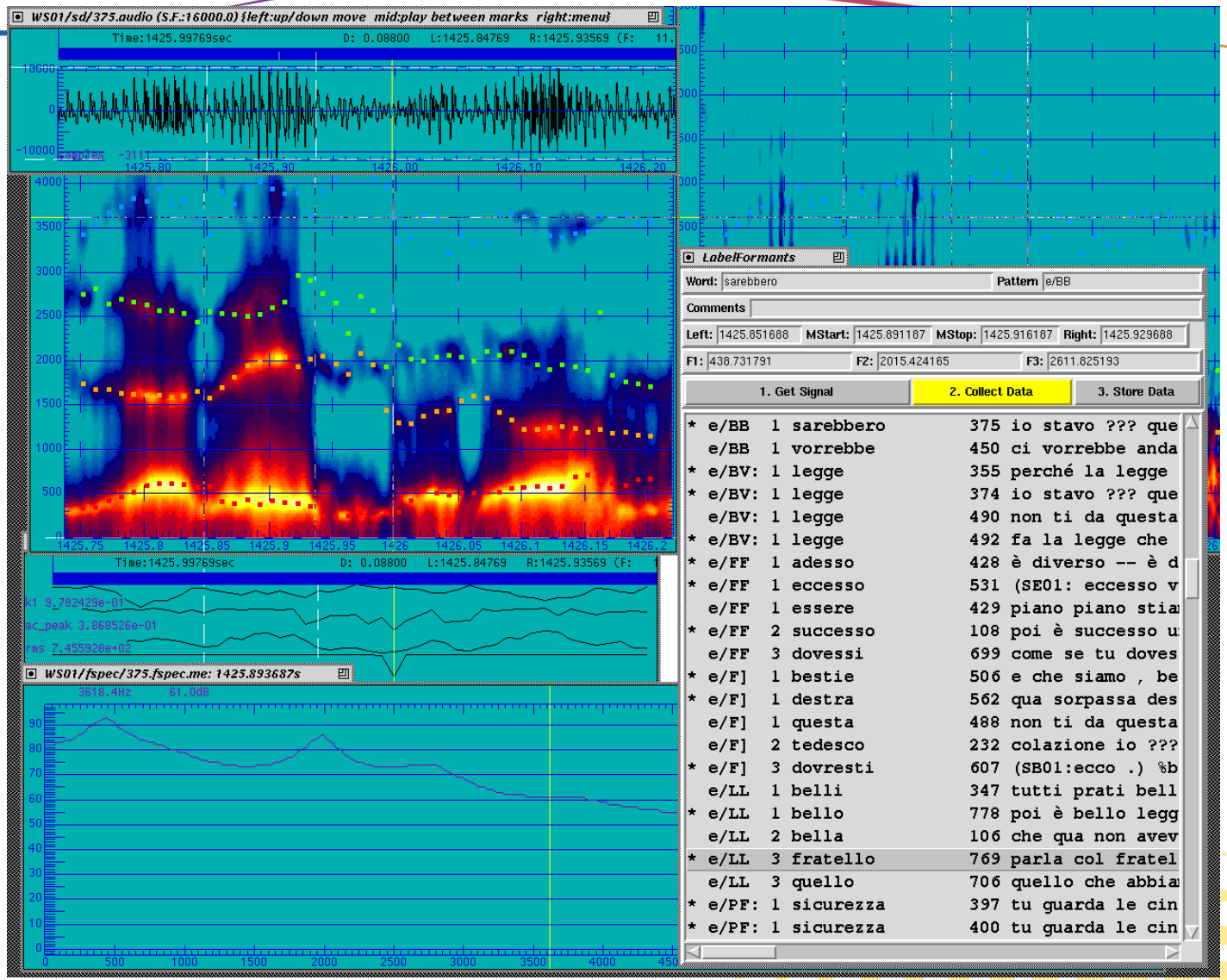
TableTrans



SPAAT (Super Phonetic Annotation & Analysis Tool)



Formant Analysis



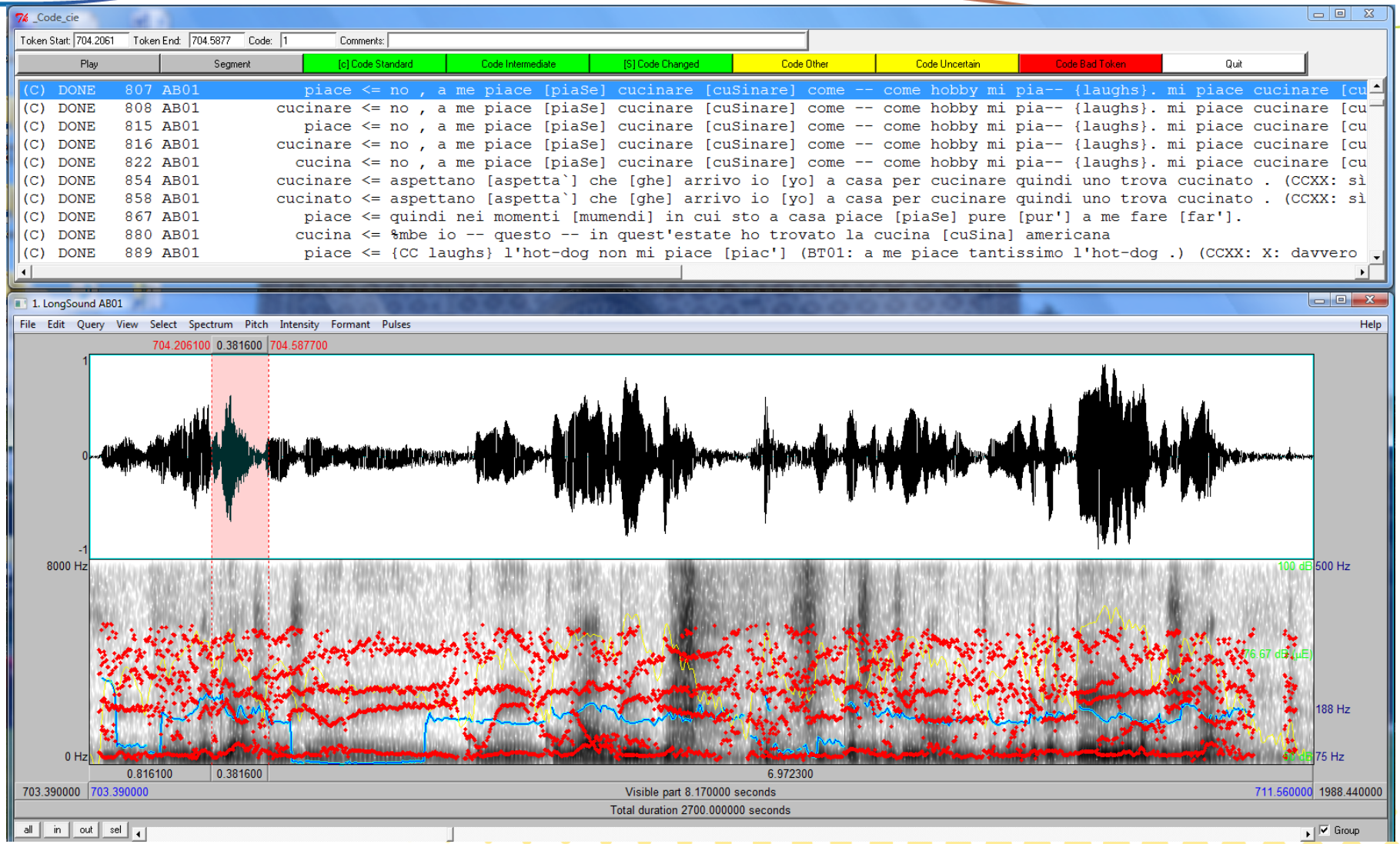
Token Selection

Vowel Segmentation

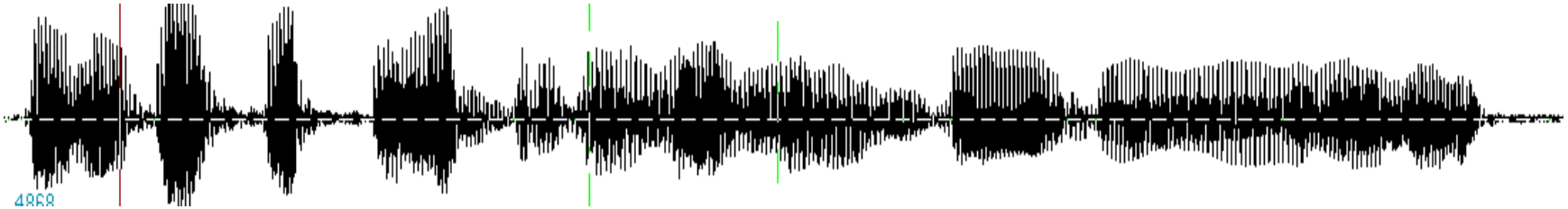
Identification of central tendency of word stressed vowel

Hand checking of formant tracker values for F1 and F2

Impressionistic Coding



Annotations



U1	U2	U3						U6	U7	
			U4:	una	donna		bella	U5		
					H1:		bella			
					S1:		E			
						F123				

Relations

		Hit		Segment	Analysis
		Hit # →	→	Hit # →	Hit #
	Utterance	Pattern		Segment	F1
	Utterance # ←	Utterance #	Lexicon	S Start Time	F2
	U Start Time	Word →	Word	S Stop Time	F3
	U Stop Time	W Start Time	Expected Pron		
Subject	Channel	W Stop Time	Stressed Vowel		
Speaker ←	Speaker	Actual Pron	Preceding Env		
Age	Situation		Following Env.		
Sex					
Ed Level					
Profession					
Region					
Location					

Format Needed

```
speaker=MC01    situation=8    channel=X

hitnum=1267                    uttnum=376
word=gabbia                    pattern=a/BB
utterance=gabbia               comments=""

mstart=2610.823500 mstop=2610.848500
sstart=2610.740000     sstop=2610.908000
wstart=2610.710000     wstop=2611.533687
ustart=2610.71         ustop=2611.54

F1=891.1739  F2=1706.9408  F3=2337.6178
```

Managing Data

- How can we manage data all through the coding and analysis process?
- In the case of Praat
  - scripting language
    - SLAAP Vowel Capture Script (http://ncslaap.lib.ncsu.edu/tools/)
    - Josef Fruehwald's Vowel Logging System
  - menus and buttons
  - control from outside
    - Plotnik/Praat (Labov, Rosenfelder, this conference)
  - interaction through file formats
    - Transcriber □ Praat TextGrid (http://ncslaap.lib.ncsu.edu/tools/)
    - lcf2txt.pl: Xtrans .lcf □ Text (for forced aligner)
    - lcf2TextGrid.pl: Xtrans .lcf □ Praat TextGrid
    - Penn Phonetics Lab Forced Aligner (http://www.ling.upenn.edu/phonetics/p2fa/) □ Praat TextGrid

Linguistic Data Consortium

Annotator Consistency

- Measure of success for coding specification
  - Can coding be re-applied by independent annotator with high agreement?
- Determining inter-annotator agreement and consistency
  - For both dependent and independent variables
  - Raw percentages aren't enough – some agreement just due to chance
  - More robust measures, e.g. Kappa scores
- Why bother?
  - Reveals ambiguities and unstated assumptions in spec
  - Necessary for comparison of results across studies and over time

Linguistic Data Consortium

Publishing

- development, production methods fully documented
- complete audio available in standard format uncompressed or with lossless compression
- transcripts in XML or other standard, non-proprietary platform-independent and application-independent format
- consistent naming conventions for audio, transcriptions and any annotations
- all data formats specified and confirmed
- inter-annotator agreement measured and published
- coding practice fully documented
- results shared
  - not just findings but raw data and annotations

Linguistic Data Consortium

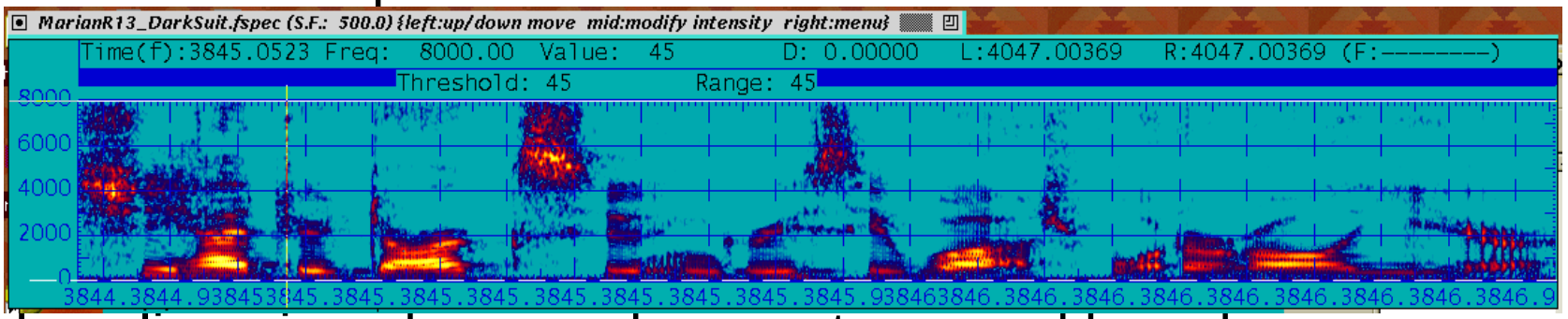
Fine

Coding Spec Best Practices

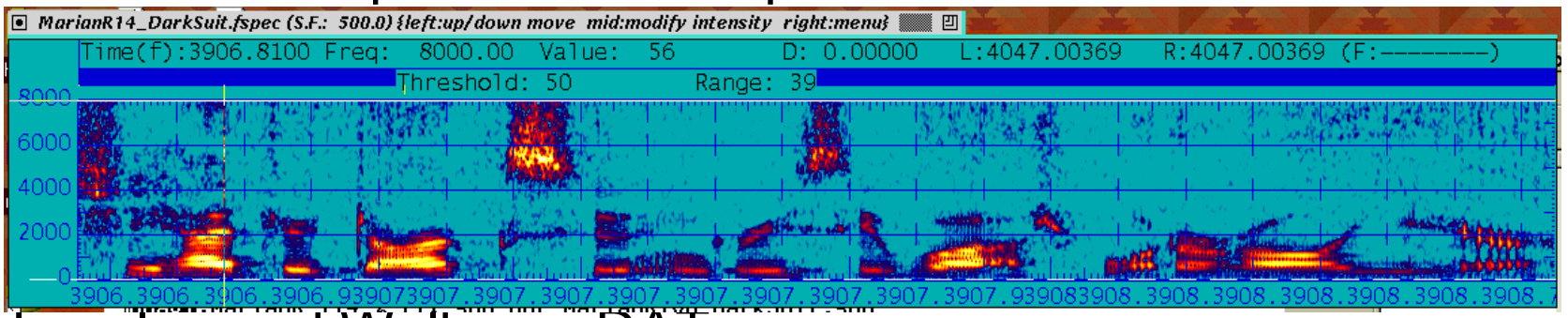
- Formal annotation/coding specifications promote coder reliability and direct comparison of results
- Developed iteratively over several rounds of pilot labeling including analysis of inter-coder reliability, via (double-blind) dual coding
  - Consider removal, merging of rules/categories with low consistency
- 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

Recording Quality

- Lavalier microphone and minidisk



- Lavalier microphone and computer sound board



- Lavalier and Walkman DAT