

Robust Sociolinguistic Methodology: Tools, Data and Best Practices

Christopher Cieri, Stephanie Strassel
{ccieri, strassel}@ldc.upenn.edu

University of Pennsylvania
Linguistic Data Consortium and Department of Linguistics
3600 Market Street, Philadelphia, PA 19104 U.S.A.

www.ldc.upenn.edu

Background

- **National Science Foundation**

- **TalkBank: (www.talkbank.org)** an interdisciplinary research project funded by a 5-year grant (BCS-998009, KDI, SBE) to Carnegie Mellon University and the University of Pennsylvania.
- The TalkBank coordinators are Brian MacWhinney (CMU) and Christopher Cieri (Penn). Co-P.I.'s are Mark Liberman (Penn) and Howard Wactlar (CMU). Steven Bird (Melbourne) consults.
- Foster fundamental research in the study of human and animal communication. TalkBank will provide standards and tools for creating, searching, and publishing primary materials via networked computers.
- 15 disciplinary groups were identified in the TalkBank proposal; six have received focused efforts: Animal Communication, Classroom Discourse, Conversation Analysis, Linguistic Exploration, Gesture, Text and Discourse and Technical Development. In 2002, Sociolinguistics added as the seventh area on the strength of the DASL project



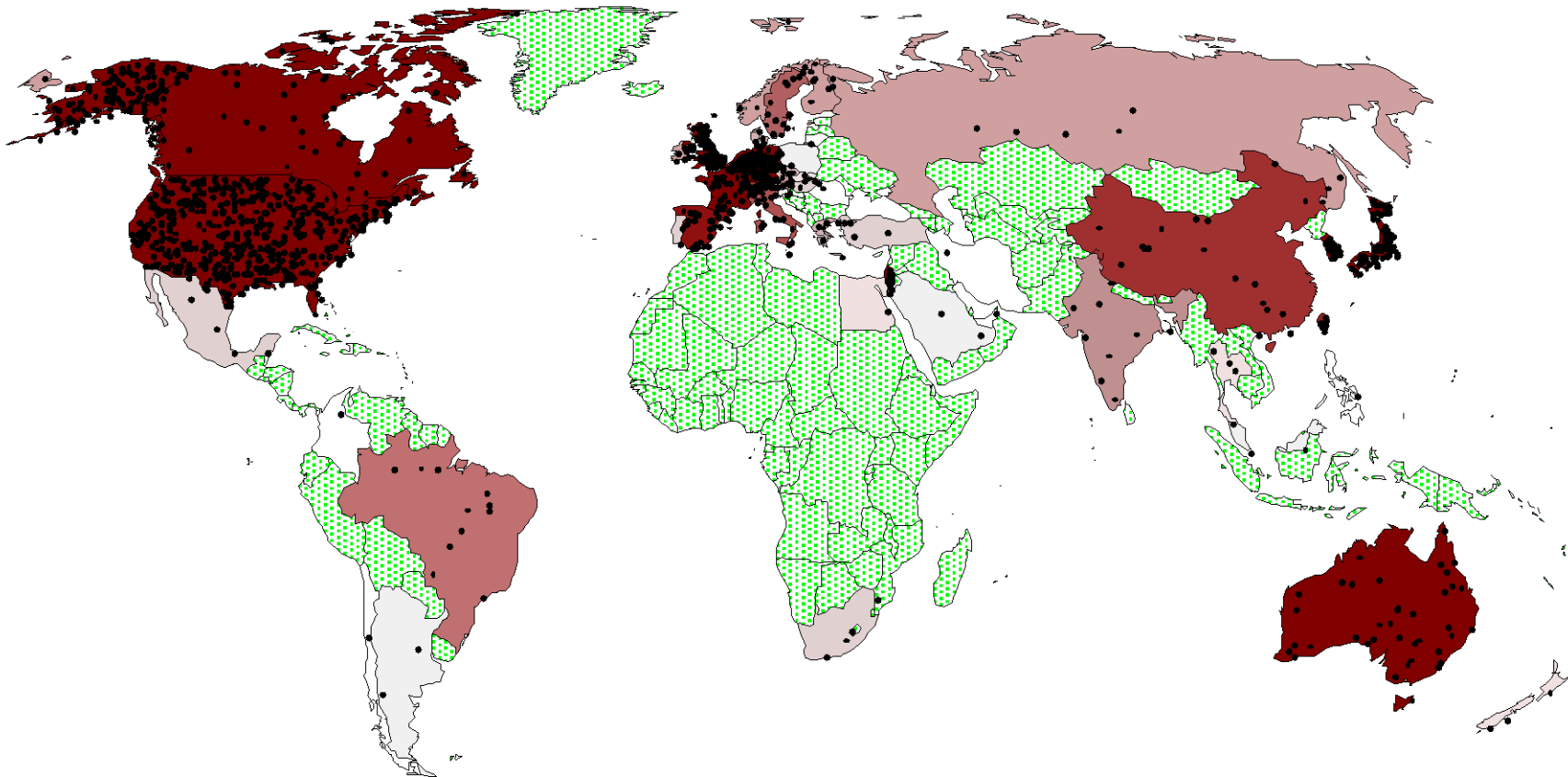
- **Linguistic Data Consortium**
 - a not-for-profit activity of the University of Pennsylvania
 - serving researchers, educators and technology developers in language-related fields
 - by creating and collecting, archiving, distributing
 - language resources, including data, tools, standards and best practices
- **Data Distribution**
 - organizations join per year receiving ongoing rights to all data released that year
 - data from funded projects at LDC or elsewhere, community or LDC initiatives
 - broad data distribution across research communities
 - funding agencies avoid distribution costs
 - users receive vast amounts of data while avoiding enormous development costs
- **Data Collection, Annotation, Research Projects**
 - support NSF, DARPA programs
 - other government and commercial technology development programs
 - all results distributed through LDC



Who/What is LDC

In operation 11 years, 36 FT Staff
248 Corpora + 2/month
>15,000 copies to 468 members +
1197 organizations in 57 countries

N/S America	784
Europe	518
Asia	184
ME/Africa	53
Aus/NZ	41



- Investigate best practices in use of digital data and tools to support empirical linguistic inquiry and documentation. Now a Talkbank activity.
- Vision for empirical, quantitative research that is
 - robust – tackles new challenge conditions
 - accountable – documents relationship between method and result
 - repeatable – shares data, tools methods to allow comparison
 - collaborative – encourages researchers to build upon each others' work
- Analysis of –t/d deletion in the published TIMIT (*isbn:1-58563-019-5*) and Switchboard (*isbn:1-58563-121-3*) corpora
- Web based annotation tool
- SLX Corpus of Classic Sociolinguistic Interviews conducted by William Labov and his students
- SLX Corpus toolkit
- This workshop

- **Corpus** – a body of records of linguistic behavior collected and annotated for a **specific** purpose
 - audio and video recordings of speech and gesture
 - written text
 - collected under naturalistic or experimental conditions
- **Annotation** is any process of adding value to a corpus
 - through the application of human judgment or
 - (semi)automatic processing based upon human judgment or previous annotation
- **Segmentation and Transcription** are special kinds of annotation
 - segmentation defines the scope and granularity of future annotations
 - transcription encodes subtle human judgements about what was said, who said it and what was intended
- **Coding of sociolinguistic variables** is annotation

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.

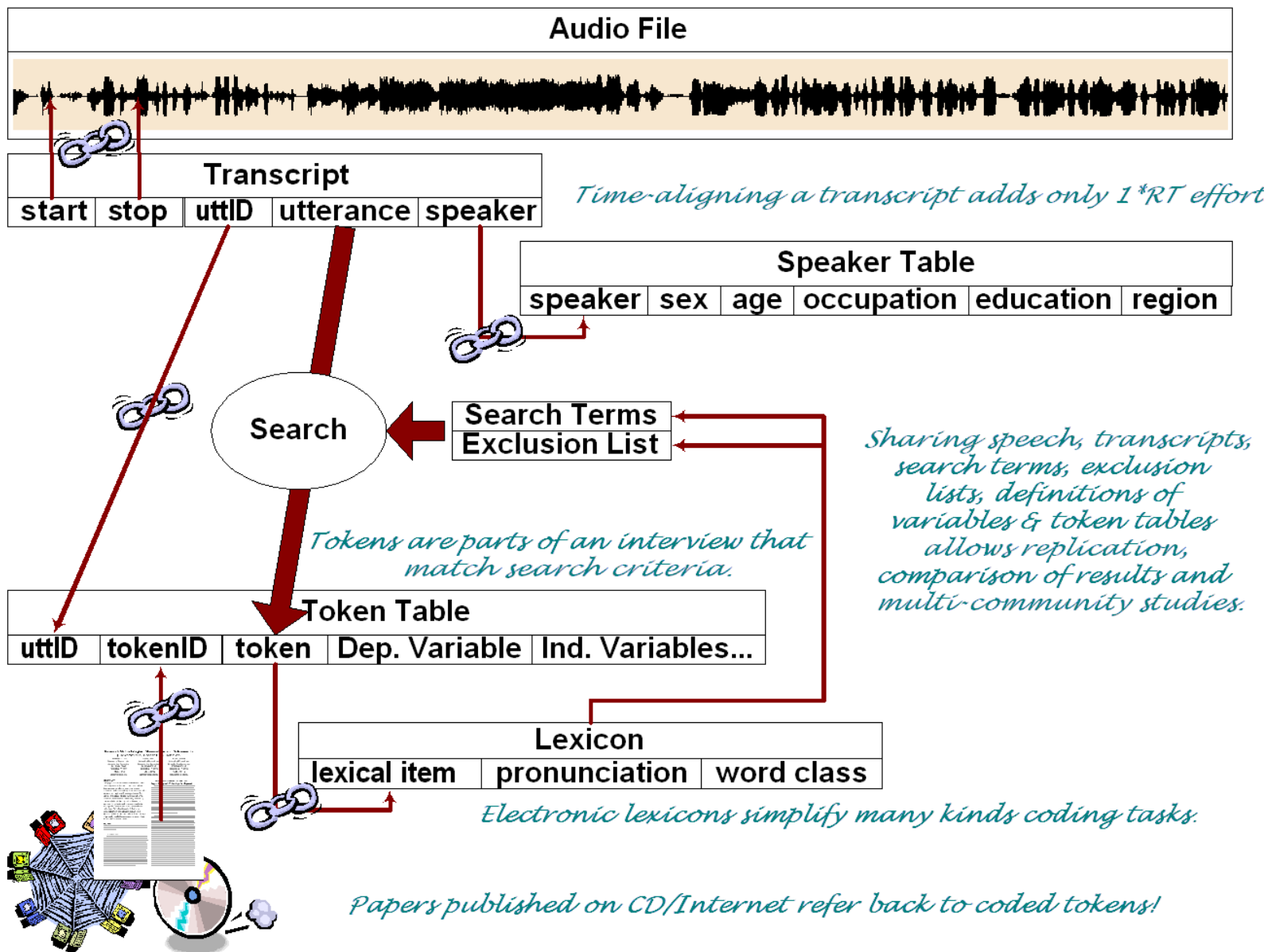
After 40 years of technological advance, our use of data is largely unchanged; only the components differ.

2003



- **Suboptimal methodologies lose information**
 - miss tokens, give an unbalance view of corpus
 - code information redundantly
 - lose sequence and time of utterances, events
 - ignore the style profile of an interview
- **Optimal methodology**
 - simplifies work so that researchers can address current topics more completely and with balance and can approach new topics
 - improves consistency
 - retains time and sequence information
 - retains mapping between sound, transcript, selected tokens, their coding, the analysis and examples in publication
 - encourages re-use of data
 - » each additional pass requires less effort than original

2003-



Case Study

- **Is the phonological variation observed better modeled as a small number of varieties with inherent variation or a larger number of invariant varieties?**
- **Vowel system of a Regional Italian influenced by Standard Italian and two local dialects**
- **Data**
 - 80 subjects stratified for age, gender, socioeconomic background
 - Interviewers both native and non-native
 - Subjects typically interviewed in pairs
 - Multiple conversational situations (styles)
 - Style as a function of time in the interview
 - Objective and subjective analyses:
 - » vowels system, intervocalic /v/, “c” before high vowels
- **Need Tools, Formats**
 - Collect and Annotate data
 - Manage layers of analysis
 - Summarize and Present results

Before

- Listen to tape for interesting tokens
- Digitize individual tokens
- Code tokens (using software where appropriate)
- Mark tokens on score sheet
- Reformat data for statistical analysis
- Problems
 - slow, labor intensive
 - high risk of missed tokens
 - tokens typically unbalanced, representation of styles poor
 - time measured poorly
 - effort for reanalysis nearly equal to effort for original
 - only limited opportunities for re-use

- Digitize entire interview & check audio quality.
- Transcribe, segment & check format.
- Query system for items of possible interest.
- Where appropriate, preprocess for segmental analysis.
- Label and analyze segments of interest.
- Summarize.
- **Advantages**
 - fewer misses
 - balanced coverage
 - time measured accurately
 - re-use & reanalysis profits from previous preparation

- Recorded on audio cassette using Sony Walkman Pro stereo recorder and two lavalier microphones.
 - each subject on separate mike, interviewer typically off-mike
- Digitized as **two channel**, 16 bit, 32KHz files via Sony DAT recorder; down-sampled to 16KHz and transferred to computer via a Townshend DAT Link; saved in Entropic .sd format
 - .wav and .sph formats also possible
- Demultiplex, check signal levels & remove empty or clipped channels
- Confirm recording length, trim beginning & ending silence

- **Time align transcript to audio file**
 - allows transcript to serve as index into audio
 - focuses attention on units smaller than interview
- **One long file instead of many small files**
 - preserves integrity of original event, allows later re-segmentation
 - preserves time
- **Levels**
 - Initial Segmentation
 - » at each speaker turn
 - » within long turns at ~8 seconds
 - » segmented into breath groups where convenient
 - Further segmentation refines domain of analysis
 - » word level, phonetic segment level (for vowels)

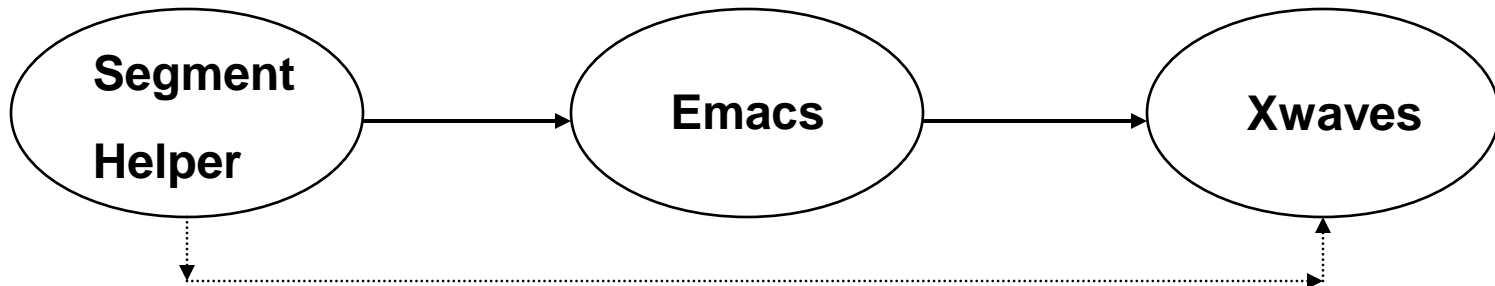
- **To transcribe or ...**
 - fewer misses
 - balanced coverage
 - re-use & reanalysis
- **Automatic or manual transcription?**
- **Segmentation before Transcription**
- **Orthographic transcription with interesting items & features transcribed phonetically**
- **Who does 1st and 2nd pass?**

- **Strans**

- Emacs with menus modified and macros added to support transcription talking to Xwaves through “send_xwaves”

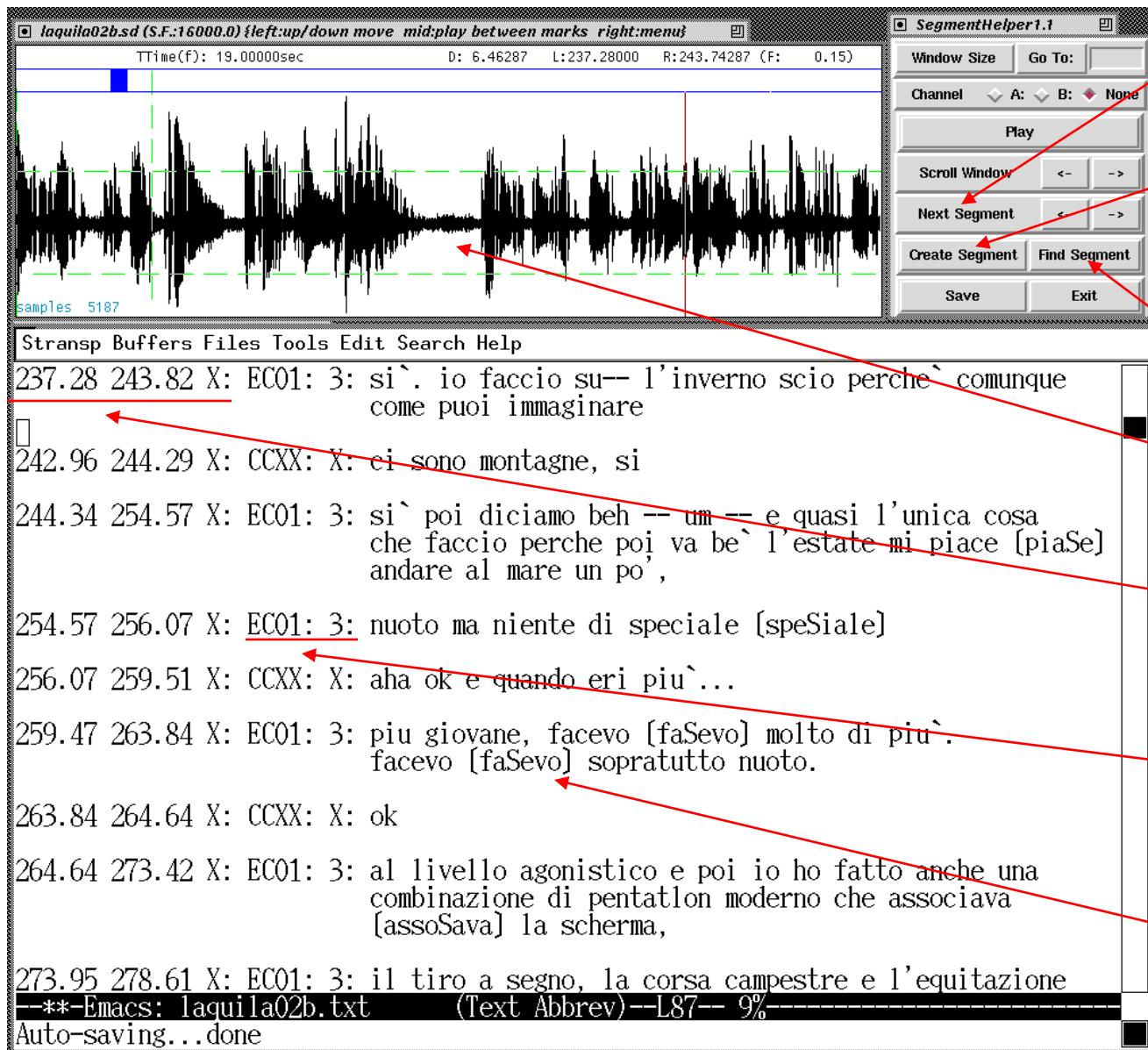
- **Segment Helper**

- Emacs running in server mode
- Client writes all commands to stdout where Emacs either acts on them immediately or passes them onto Xwaves.



- Segment Helper & all utilities hereafter written in PerlTK -- free, available on Unix and NT, merges the TK GUI capacity with Perl's flexibility and flow control.
- Now **Transcriber** does it all!

Strans +



■ Next Segment - shifts display so that 10% of last segment shows

■ Create Segment polls Xwaves for left, right cursor positions and writes those as time stamps with channel marker in text

■ Find Segment finds position in waveform of segment defined in text

■ Monoaural recording with subject on single mike; interviewer off mike.

■ Segment defined by start & stop times plus channel marker and written by software based on cursor positions.

■ Speaker ID written by human and later normalized. Situation code written semiautomatically and checked by human.

■ Interesting feature transcribed phonetically.

- **Features**

- Editing signal: - -
- Non-lexemes: %m (English & Italian spelled differently)
- Truncation: n- non
- Non-Standard pronunciation: usciti [usci'i]
- Code switching: <English Where are you from?>
- Overlap/Back-channel: (CCXX: %mhm)
 - » favor subject over interviewer, turn-holder over others

- **ASR 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

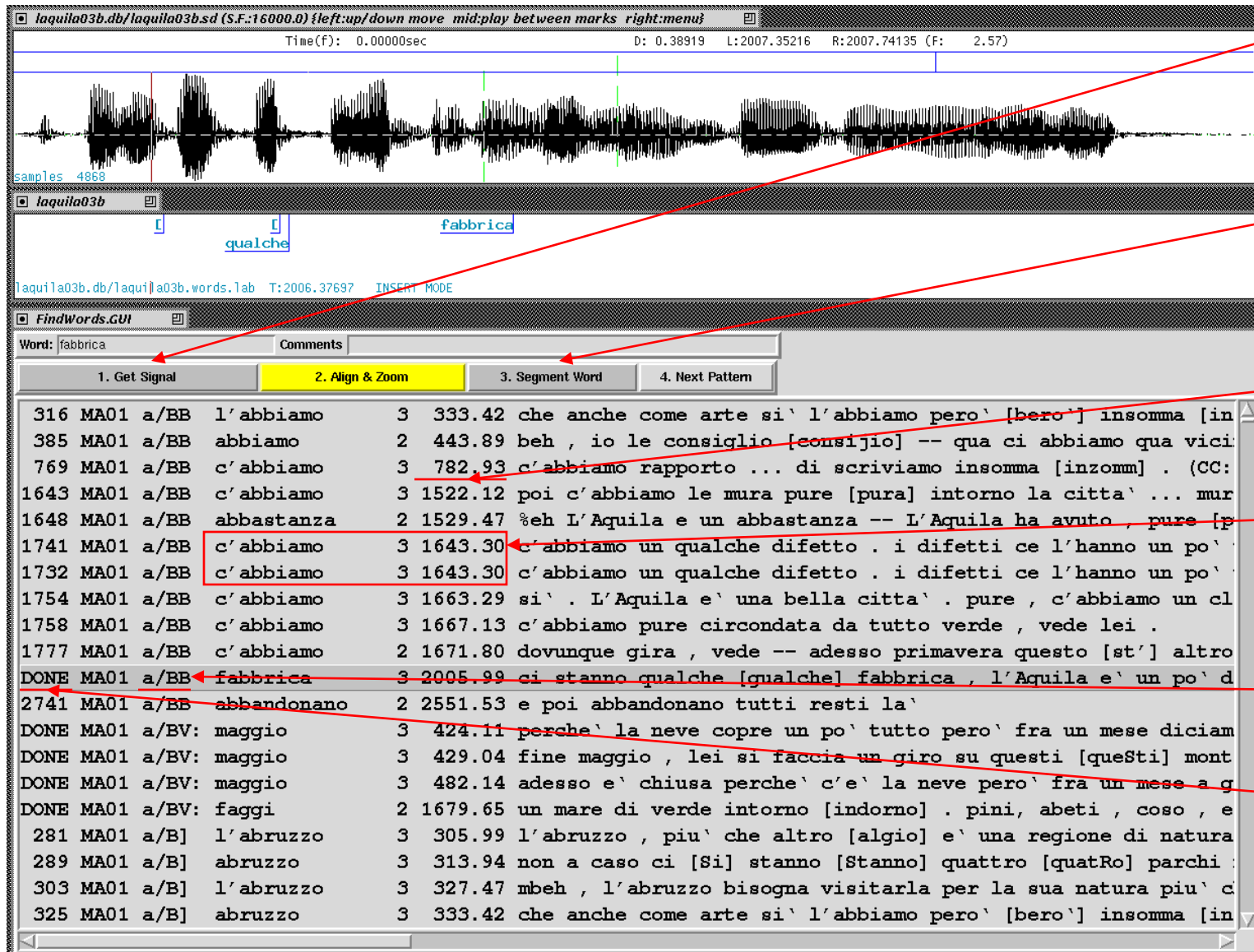
- **After Segmentation and Transcription, files are checked by a second transcriptionist for**
 - bad segmentation
 - » too much silence in segment
 - » segment boundary too close to signal
 - » signal not contained within segment
 - inaccurate transcription
 - inaccurate situation code
 - misspellings
 - inaccurate phonetic transcription within []
- **Format**
 - 628.67 633.94 X: MC01: 2: e m- -- a mezzanotte siamo rientrati %e -- in albergo

Syntax Check

- After last human QC pass use automatic process
 - segments that are too long
 - time stamps out of order or internally inconsistent
 - impossible channel marker, speaker ID or situation code
- QC catches human formatting errors.
- System controls all subsequent processing avoiding most kinds of human error.
- Format
 - uttnum=77 speaker=MC01 situation=2 channel=X
 ustart=628.67 ustop=633.94
 utterance=e m- -- a mezzanotte siamo rientrati
 %e -- in albergo

- Software looks up each word in pronouncing lexicon to enable phonetic query, categorization.
- Software searches reformatted transcript, identifies and numbers any words matching query. Each hit word is presented to user in context as text and audio
- Software guesses location of word in utterance based on simple assumption that all syllables are of roughly equal length -- does surprisingly well
- Linguist adjusts word boundaries in waveform display, zooms and iterates until satisfied.
- **Format**

```
- hitnum=276 pattern=e/R] word=albergo
  wstart=632.934813 wstop=633.778312
  uttnum=77 speaker=MC01 situation=2 channel=X
  ustart= 628.67 ustop= 633.94
  utterance=e m -- a mezza notte siamo rientrati %e -- in
  albergo comments=""
```



■ GetSignal locates and plays utterance, guesses word position and sets cursors

■ SegmentWord writes segmentation to new file and marks hit as done.

■ Retaining times allows user to balance samples over corpus

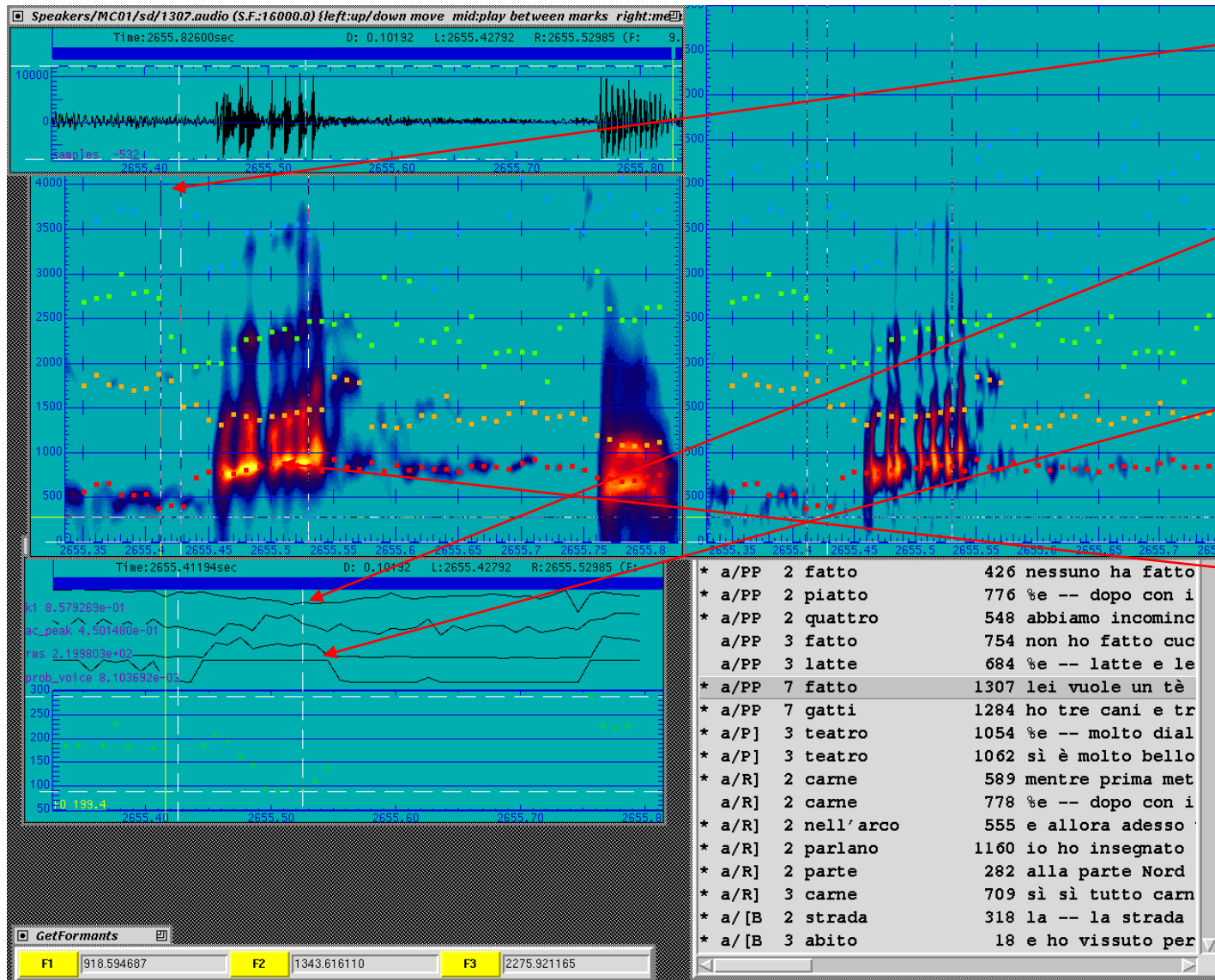
■ Lexical Item matching search. May be more than one per utterance

■ Abstract Label for Search Pattern

■ Unique Hit Number

- **Automatically create analytic files for each token**
- **Accepts word start and end times from previous step**
- **Finds corresponding audio**
- **Creates**
 - **Wide band spectrogram**
 - **Narrow band spectrogram**
 - **Maximum entropy (LPC) spectrogram**
 - **Formant tracks**
 - **F0 analysis**
- **Saves all files for later use by human annotator.**

Label Formants



■ Time Aligned displays of waveform, F0 and spectrograms

■ Software guesses position of segment within word.

■ User adjusts segmentation and saves to file.

■ Software estimates formant values automatically. User selects or corrects.

■ All sound files, spectrograms, and F0 files processed ahead of time in batch and saved for later redisplay.

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

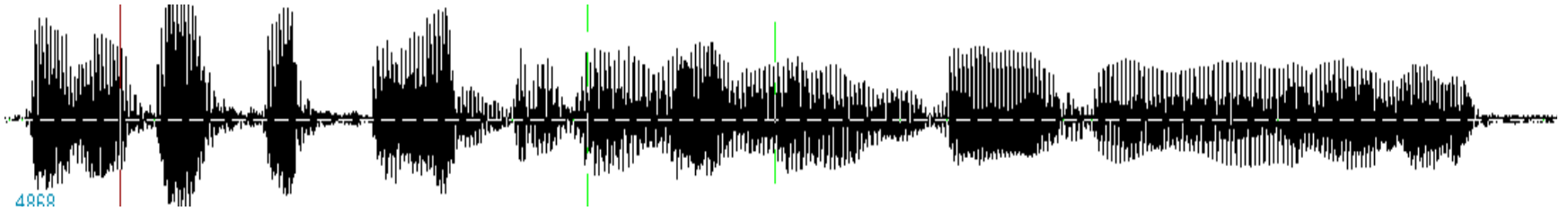
wstop=2611.533687

ustart=2610.71

ustop=2611.54

F1=891.1739 F2=1706.9408 F3=2337.6178

Annotations



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

		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		
	Channel	W Stop Time	Stressed Vowel		
Subject	Speaker	Actual Pron	Preceding Env		
Speaker	← Speaker		Following Env.		
Age	Situation				
Sex					
Ed Level					
Profession					
Region					
Location					

- **Software flattens relations and exports to analytical software; R in this case.**

Best Practices for Digital Methodology: Collection

Speakers utter phonetically rich sentences under a variety of circumstances.

	1	2	3
Is "dark" r-ful?			
Is fricative in "greasy" voiced?			
Is there intrusive-r in "wash"?			
What's the vowel in "water"			
How confident are you?			

- **Commonly used: small portable recorder and lavalier microphone**
 - High quality is possible
 - Cost is generally low
 - Unobtrusive
 - Highly portable
- **Obtrusiveness and quality are variables that can be managed.**
- **Data collected under other conditions may be natural and valuable.**
 - Examples from CALLHOME, Switchboard, ROAR



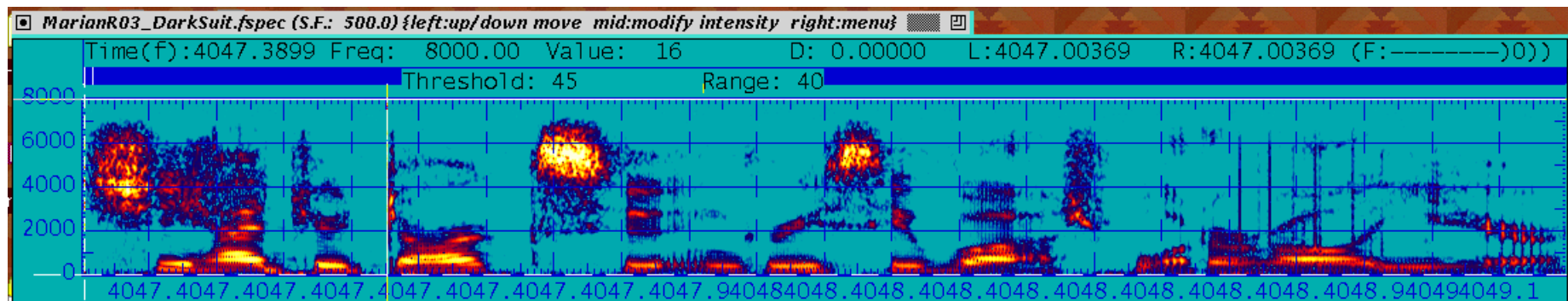
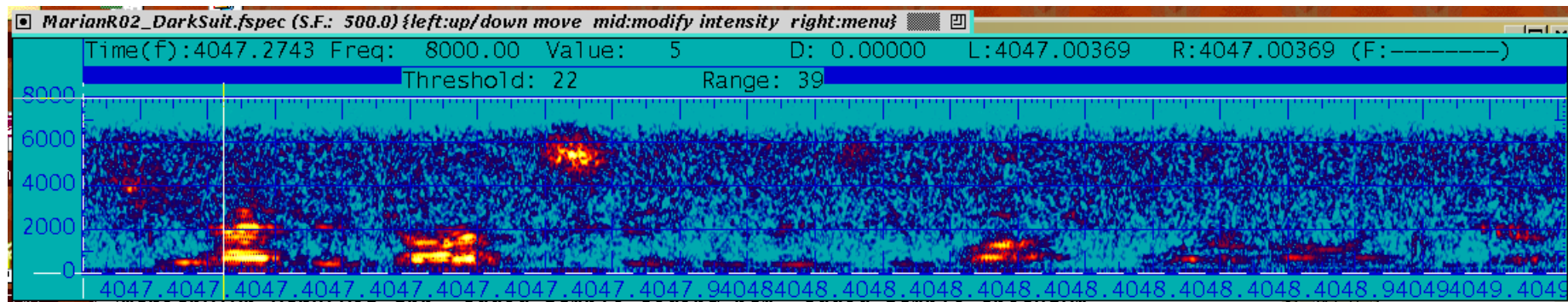
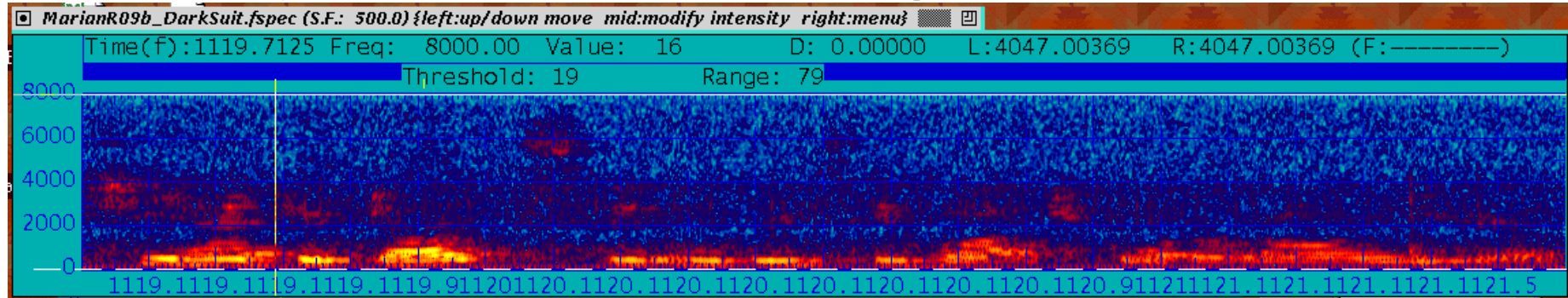
UDC Recording Experiment

- Two subjects in sociolinguistic interviews with semantic differentials, phonetically rich sentences, word list.
- Microphones and recording devices co-varied.

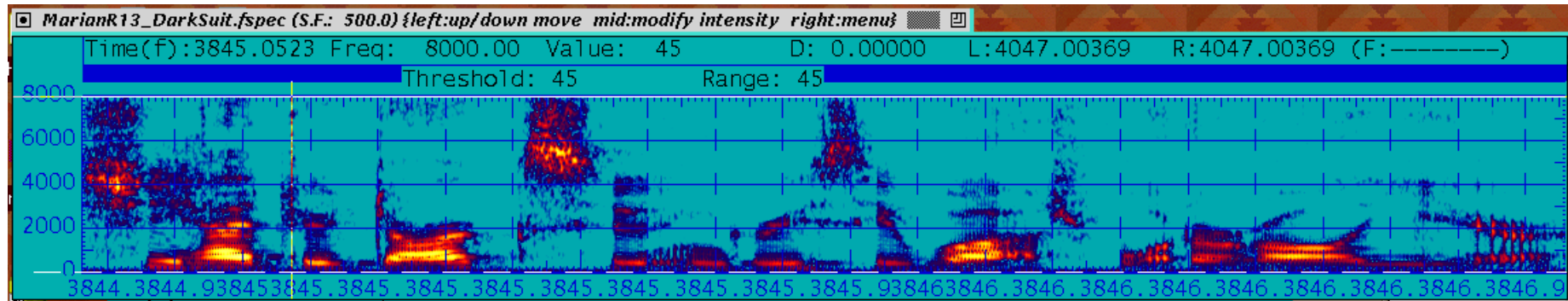
#	Microphone	Recorder	Comments
1	PZM on Subject's Chair	Studio System	Low Frequency Hum
2	Wireless, Cardioid Lavalier on Interviewer	Studio System	Nearly Inaudible
3	Hypercardioid, Head Mounted	Studio System	Very Little Noise
4	Lavalier	Studio System	Very Little Noise
5	Cardioid Lavalier	Studio System	Very Little Noise
6	Dynamic Studio on Stand	Studio System	Faint Hiss
7	Studio on Stand	Studio System	Low Frequency Hum
8	Shotgun (Hypercardioid) on Boom	Studio System	High Frequency Noise
9	Built-in on Table	Panasonic RQ-A70	Low Signal, High Noise
10	Lavalier	Sony Walkman Pro	Low Frequency Hum
11	Lavalier	Sony TCM5000EV	Faint Low Frequency Hum
12	Lavalier	Sony Walkman DAT	Faint Low Frequency Hum
13	Lavalier	Sony M2-R50 Minidisk	Low Signal, No Hum
14	Lavalier	Computer	Hiss

- **Variables**
 - Really poor choices can affect coding of even highly salient variables.
- **Distance from mouth to microphone**
 - Low frequency is affected by even small differences.
 - Room noise becomes more obvious with greater distances.
- **Unobtrusive collections**
 - Very unobtrusive microphones can still produce very useful recordings.
- **Motor Hum**
 - Recorders with motors
 - But compare minidisk and TCM5000EV
- **Interference**
 - Recording from laptop's sound board.

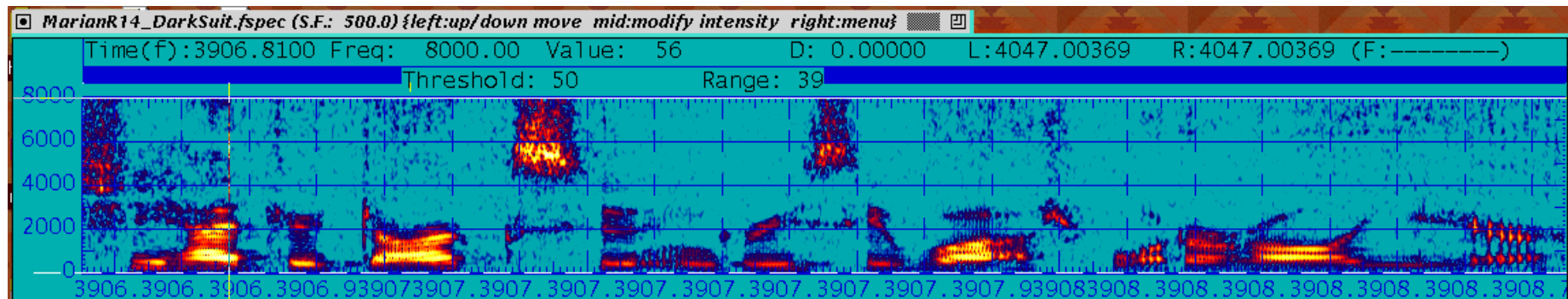
- Two very poor choices and one good



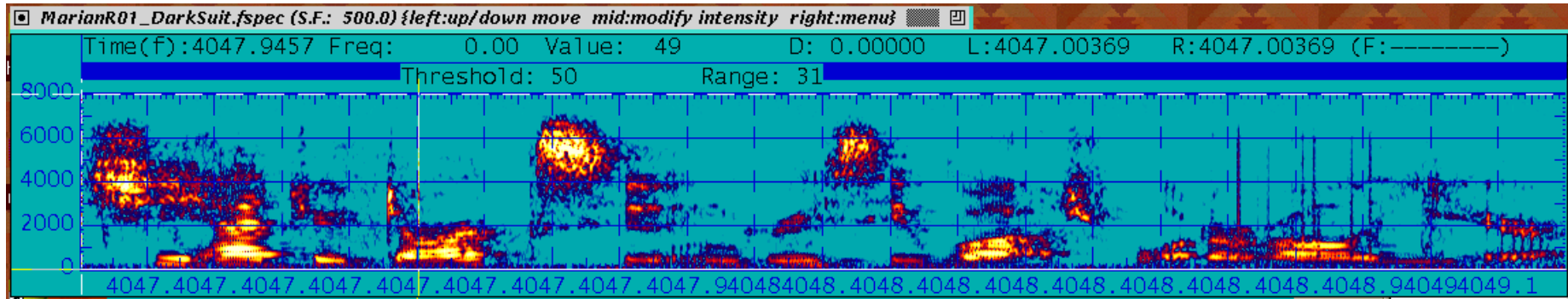
- Lavalier microphone and minidisk



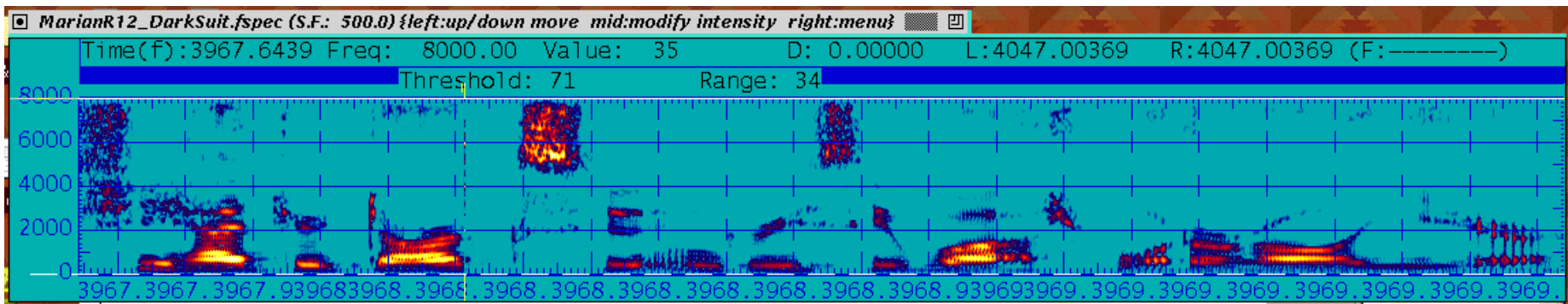
- Lavalier microphone and computer sound board



- PZM



- Lavalier and Walkman DAT

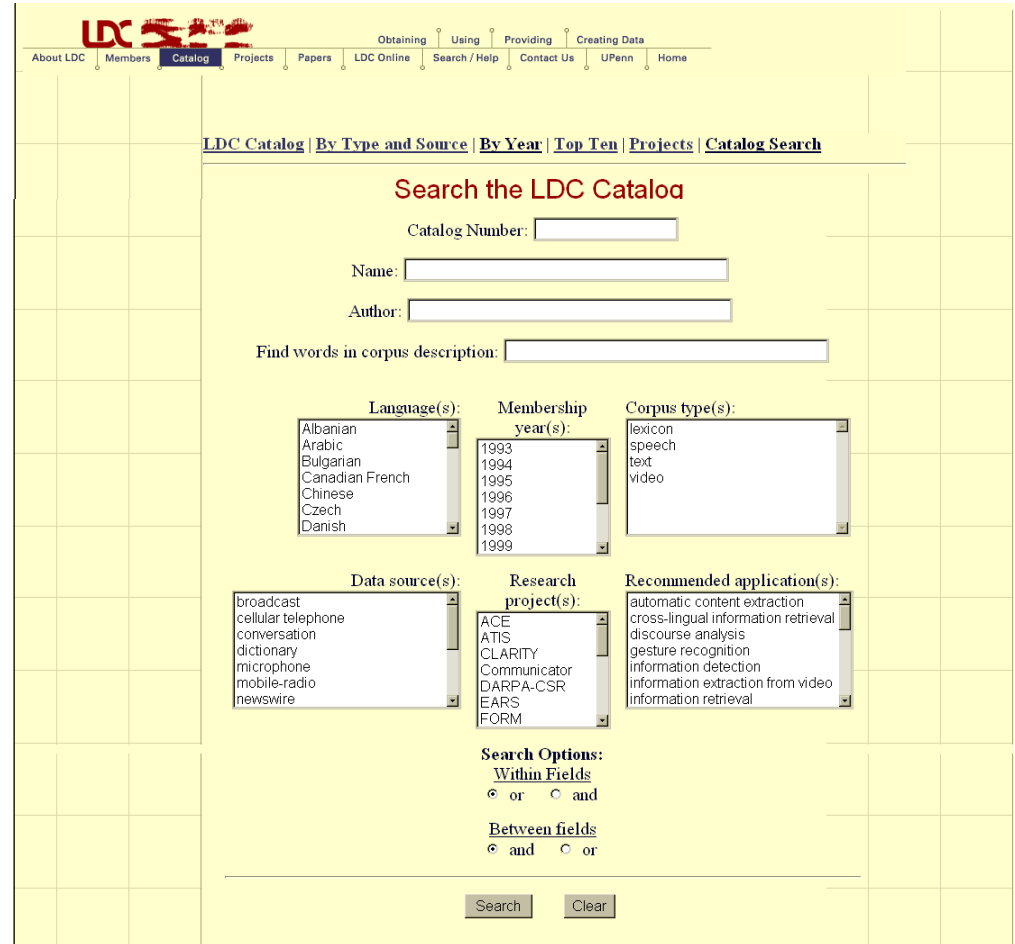


Best Practices for Digital Methodology: Published Data

- **Linguistic Corpus: a body of records of linguistic behavior collected and annotated for a specific purpose**
- **Why should a sociolinguist want to use someone else's data?**
 - Exploratory study before doing individual data collection
 - Broaden scope
 - Locate 'rare' constructions
 - Supplement individual data collection
 - Lots more data, possibly greater range of data
 - Low- or no-cost access to data
 - Often highly searchable - get lots done quickly
 - New perspective

- LDC: <http://ldc.upenn.edu/Catalog>

- Free text search in catalog number, corpus name, author, corpus description, and or select one or more search terms in language, membership year, corpus type, data source, sponsoring project or recommended application menus



The screenshot shows the LDC Catalog search interface. At the top, there is a navigation bar with links: About LDC, Members, Catalog, Projects, Papers, LDC Online, Search / Help, Contact Us, UPenn, and Home. Below this, there is a sub-navigation bar with links: LDC Catalog, By Type and Source, By Year, Top Ten, Projects, and Catalog Search. The main heading is "Search the LDC Catalog". Below this, there are several search fields: "Catalog Number:", "Name:", "Author:", and "Find words in corpus description:". There are also six dropdown menus for selecting search criteria: "Language(s)", "Membership year(s)", "Corpus type(s)", "Data source(s)", "Research project(s)", and "Recommended application(s)". At the bottom, there are "Search Options" for "Within Fields" (radio buttons for "or" and "and") and "Between fields" (radio buttons for "and" and "or"). Finally, there are "Search" and "Clear" buttons.

- ELRA: <http://www.elra.info/>
- Select: “Fast track to ELRA’s Catalogue”
- Search for words anywhere in catalog entry

The screenshot shows the ELRA website interface. On the left is a purple sidebar with the ELRA logo and a list of links: 'Fast track to ELRA's catalogue', 'ELRA', 'Language Res', 'Services arou', 'Newsletter', 'Market Studie', 'Members' Only', 'LREC Confere', 'Links', and 'Press Release'. Below these links is a 'Bug Report Service' button. The main content area has a header 'European Language Resources Association' and a 'LRs' tab. Navigation links include 'Definition', 'Use and Benefits', 'Applications', 'Catalogue', 'Site Map', and 'Home'. The title 'Catalogue of Language Resources' is centered. The text describes the ELDA catalogue and provides a link to www.elda.fr. A search box with a 'Help' link, a search input field, a 'GO!' button, and a dropdown menu set to 'all words (AND)' is present. The footer contains the copyright notice 'Copyright © 2001-2003 ELRA - Webmaster'.

European Language Resources Association

LRs

[Definition](#) | [Use and Benefits](#) | [Applications](#) | [Catalogue](#) | [Site Map](#) | [Home](#)

Catalogue of Language Resources

An increasing number of language resources in the various fields of HLT (namely spoken, written and terminological resources) are made available via the Evaluations & Language resources Distribution Agency (ELDA), in its catalogue which you can consult on-line @ www.elda.fr ([catalogue section](#))

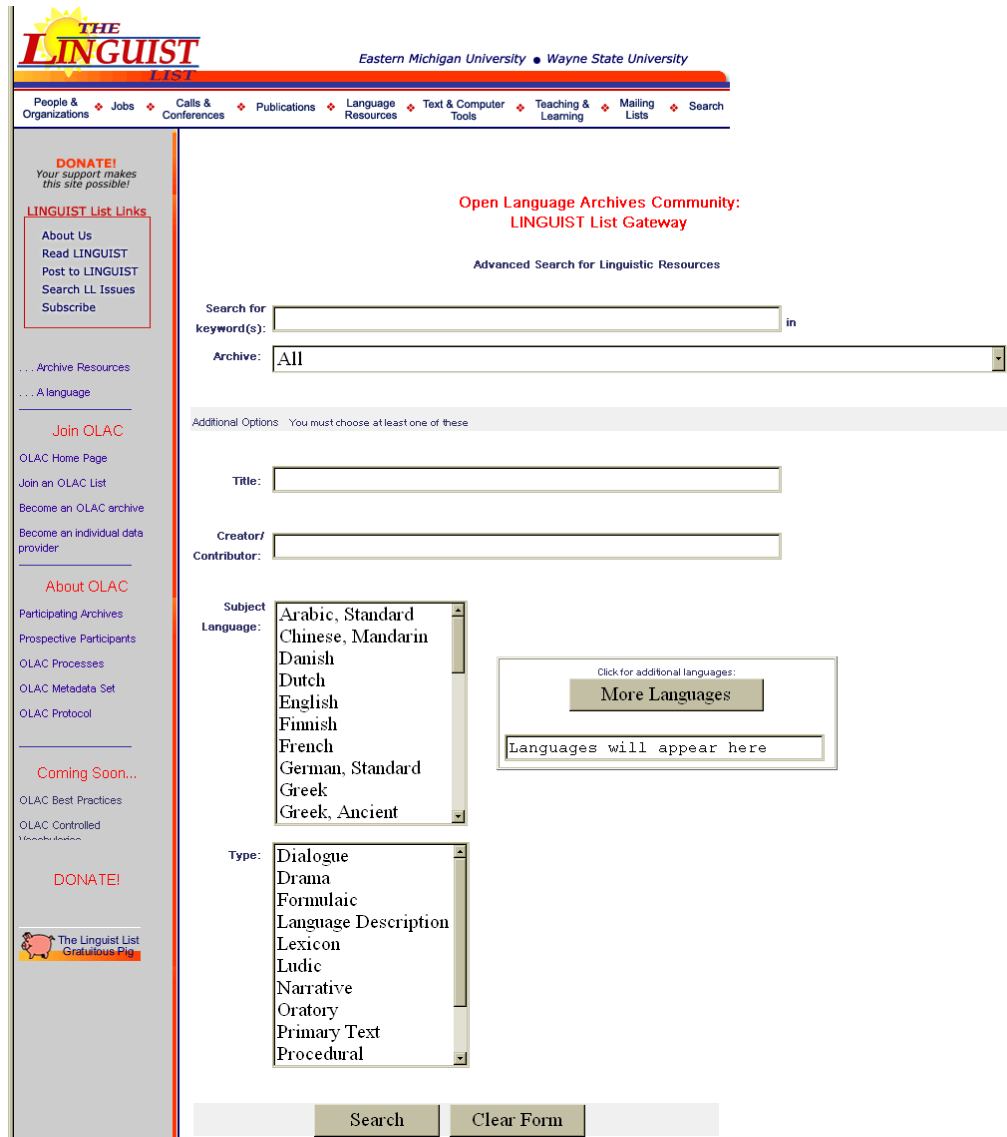
If you have any suggestion or comments, or need any further details about ELDA and its catalogue of language resources, please consult the [ELDA web site](#) or [contact us](#).

[Help](#)

Search catalogue all words (AND)

Copyright © 2001-2003 ELRA - [Webmaster](#)

- OLAC: <http://www.language-archives.org/>
- Union catalog of 28 other providers of linguistic resources
- Free text search in title, contributor and corpus description, and/or select one or more search terms in archive, language, corpus type menus



THE LINGUIST LIST Eastern Michigan University • Wayne State University

People & Organizations ♦ Jobs ♦ Calls & Conferences ♦ Publications ♦ Language Resources ♦ Text & Computer Tools ♦ Teaching & Learning ♦ Mailing Lists ♦ Search

DONATE!
Your support makes this site possible!

LINGUIST List Links

- About Us
- Read LINGUIST
- Post to LINGUIST
- Search LL Issues
- Subscribe

... Archive Resources
... A language

Join OLAC

- OLAC Home Page
- Join an OLAC List
- Become an OLAC archive
- Become an individual data provider


About OLAC

- Participating Archives
- Prospective Participants
- OLAC Processes
- OLAC Metadata Set
- OLAC Protocol

Coming Soon...

- OLAC Best Practices
- OLAC Controlled Vocabulary

DONATE!

 The Linguist List Grateful Pig

Open Language Archives Community: LINGUIST List Gateway

Advanced Search for Linguistic Resources

Search for keyword(s): in

Archive:

Additional Options You must choose at least one of these

Title:

Creator/Contributor:

Subject Language:
Chinese, Mandarin
Danish
Dutch
English
Finnish
French
German, Standard
Greek
Greek, Ancient

Type:
Drama
Formulaic
Language Description
Lexicon
Ludic
Narrative
Oratory
Primary Text
Procedural

Click for additional languages:
More Languages

Languages will appear here

- **Original fieldwork will always be necessary, providing**
 - In-depth knowledge of the speech community
 - New communities and language varieties
 - Valuable researcher training and experience
 - New methodological perspectives
 - Potential new contributions of data to public archive
- **Corpus-based approaches can complement firsthand fieldwork**
 - Permits comparison of results across studies and over time
 - Provides a stable benchmark for competing theories
 - Allows re-annotation and reuse of existing data
 - Supports measurement of inter-annotator consistency
 - Reduces impediments facing new researchers
 - Allows established scholars to tackle broader issues
 - Demonstrates best practice in corpus creation
 - Serves as a teaching tool
 - Allows for multi-site collaboration

- **(De)Compressing Audio**
 - Tony Robinson's *Shorten*
 - Lossless (2:1) and (3-5:1) lossy modes
 - Windows: <http://www.softsound.com/Shorten.html>
 - Macintosh and Linux: <http://www.hornig.net/shorten/>
- **Converting from NIST Sphere audio to .wav, .aiff, .au**
 - Dave Graff's *sph_convert*
 - Win32: ftp://ftp ldc.upenn.edu/pub/ldc/misc_sw/sph_convert_v2_1.zip
 - Mac: ftp://ftp ldc.upenn.edu/pub/ldc/misc_sw/sph_convert_v2_0.sit
- **Other Conversions**
 - Chris Bagwell's SoX
 - <http://sox.sourceforge.net/>
 - Does audio type, sample rate and byte order conversions
- **Viewing text**
 - Internet Explorer 5 and later handle Unicode (<http://www.microsoft.com/>)
 - Gaspar Sinai's *Yudit* (<http://www.yudit.org/>)
- **Citing the corpus as you would any publication**
 - But who is the author?

Best Practices for Digital Methodology: Code of Ethics

- **Assure that data users respect rights of participants, contributors**
- **Participants sign Informed Consent release approved by local IRB**
- **Data collected before IRB system, from non-funded work, from speakers of indigenous, endangered languages may be exempted. Such data collected is still subject to the same ethical concerns.**
- **Respect for Participants who make an important, generous contribution to scientific research by permitting scholars to access and analyze their linguistic behavior**
 - avoid open public criticism of these individuals
 - avoid comparisons in terms of intelligence, verbal facility, social skills, or physical appearance
- **Confidentiality by avoiding any identifying information apart from video and audio records and demographic information**
- **On discovering personal acquaintance with a participant,**
 - refrain from using the data
 - acquire explicit permission from participant
- **This requirement does not extend to use of depersonalized data or in which participants' identity is not examined.**

- **Respect for Groups who may be justifiably sensitive to criticism from the wider society.**
 - avoid making between-group comparisons that impact core features of social identity and worth.
- **Seek of professional review in cases where data publication may compromise the principles of respect for participants or groups.**
- **Share Data so that others can benefit as you have.**
- **Sanctions: It is the responsibility of the entire community to counter misuse in public forums and through personal contact.**
- **For more information, see:**
<http://www.talkbank.org/share/ethics.html>

Annotation: Adding value to the data

- **Divides the corpus into manageable units**
 - To indicate structural boundaries in audio file
 - To make subsequent transcription easier
 - To provide time-alignment for transcripts and other annotations
- **Preserve integrity of original signal**
 - Virtual, not actual, chopping of digital signal
- **Segmentation for a specific purpose**
 - Speaker turn level, utterance level, breath/pause group
 - Word level
 - Phone level
 - Finer-grained segmentation best handled as additional, specialized pass over data



Audio Segmentation

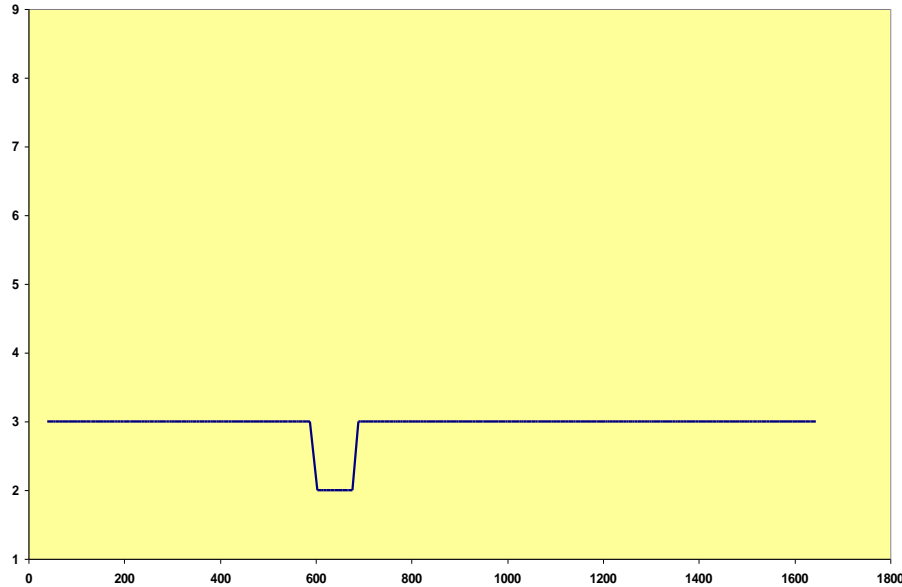
- **Requirements for any segmentation specification**
 - Specify level of granularity
 - Treatment of multiple speakers on one channel
 - Overlapping speech
 - Pauses
- **Additional features**
 - Background or other non-speaker noise
 - Speaker ID, speaker changes
 - Fidelity
- **Cost**
 - Turn-level segmentation can proceed at close to 1 x Real Time
 - Utterance, pause, breath group segments at 5+ x Real Time
 - Word, phone level segmentation
 - » Requires initial segmentation at broader granularity
 - » Much more difficult (and therefore costly)
 - » Imparts additional level of analysis
 - And requires specialists
 - Manual verification of automatic process can save time

- **Why a full transcription?**
 - Index to speech
 - Searchable
 - Provides stable basis for subsequent annotations
- **Requirements for any transcription specification**
 - Conventions for capitalization, punctuation, spelling
 - Description of any special markup
 - Treatment of variation
 - » Distinguish production error from non-standard usage
 - » Use standard orthography with markup
 - Need to find all occurrences of same word
 - Disfluencies
 - » Filled pauses, repetitions, restarts, etc.
 - Overlapping speech on same channel
 - Non-lexemes, interjections and other speaker noise
 - Sections of transcriber uncertainty

- **Quick Orthographic Transcription**
 - Speed over accuracy; close to verbatim; limited markup
 - Adequate for some purposes; 5 x Real Time
- **Verbatim Orthographic Transcription**
 - Word-for-word accurate
 - Limited additional markup
 - Hesitations, disfluencies, overlaps not carefully handled
 - Requires 2 passes minimum; 35+ x Real Time per channel
- **Careful Orthographic Transcription**
 - Verbatim, plus
 - Special treatment for range of features
 - » E.g., proper names, disfluencies, non-standard variants
 - » Background noise conditions, speaker ID, careful treatment of difficult sections
 - Requires multiple passes; 50+ x Real Time per channel
- **Phonetic Transcription**
 - Based on careful orthographic transcription
 - Automatic transcription with human verification/correction
 - Inter-annotator agreement rates at 70-90%
 - Cost much higher (estimates?)

- **What parameters drive token selection?**
 - phonological, morphological, syntactic
 - balance across extra-linguistic features
 - Are there hidden parameters?
 - » Convenience
 - » Time
 - » Fatigue
- **Incomplete coverage, lack of balance affects the study itself**
- **Variation across studies affects the ability to compare results**
- **Pronouncing dictionaries can mediate token selection**
- **What do we know about time as independent variable?**

Time as Variable



Time is on the horizontal axis.
Conversational situation (style) is on the vertical.

Larger numbers mean greater formality.

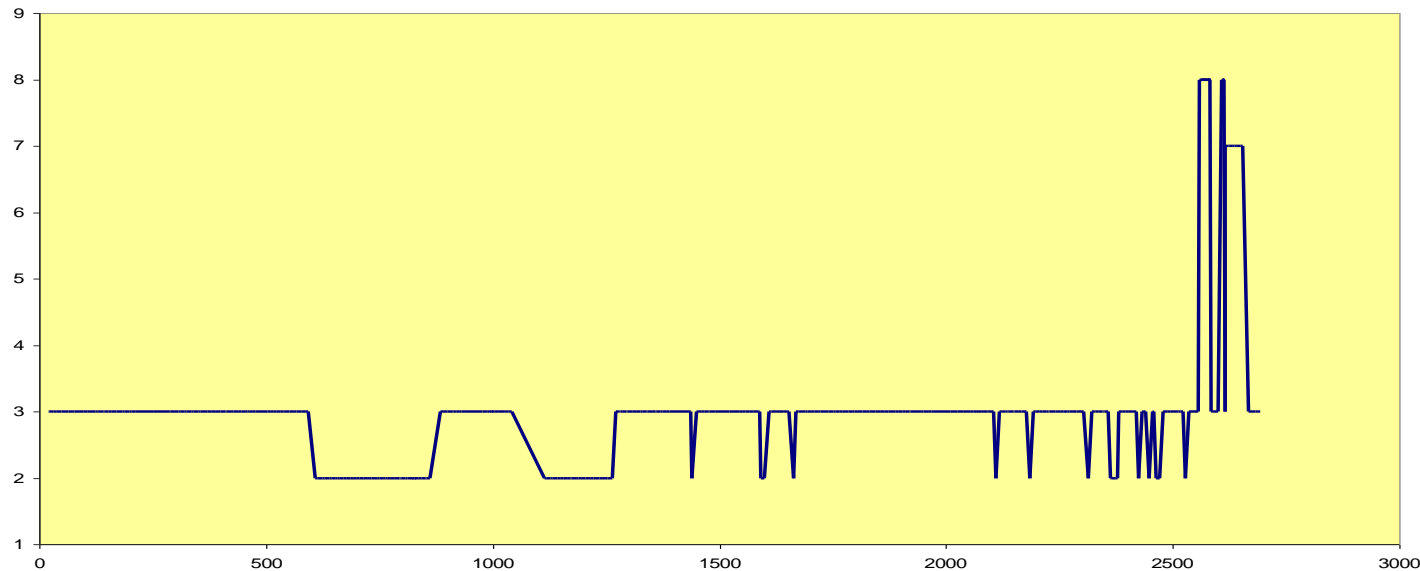
4+ are elicited styles

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!



- **Coding Specification**

- Difficulty of achieving fully explicit guidelines
- Coding of **independent** variables also a source of error
- E.g., DASL t/d deletion study
 - » Published studies vary in terms of detail in guidelines
 - » Complex factor groups, e.g. Morphology
 - » Passives, e.g. ‘I was frightened’
 - » But also seemingly simple factor groups
 - What to do with nasal flaps?
 - Glottalized segments?
 - How to measure pause?

- **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

Annotation Tools Overview

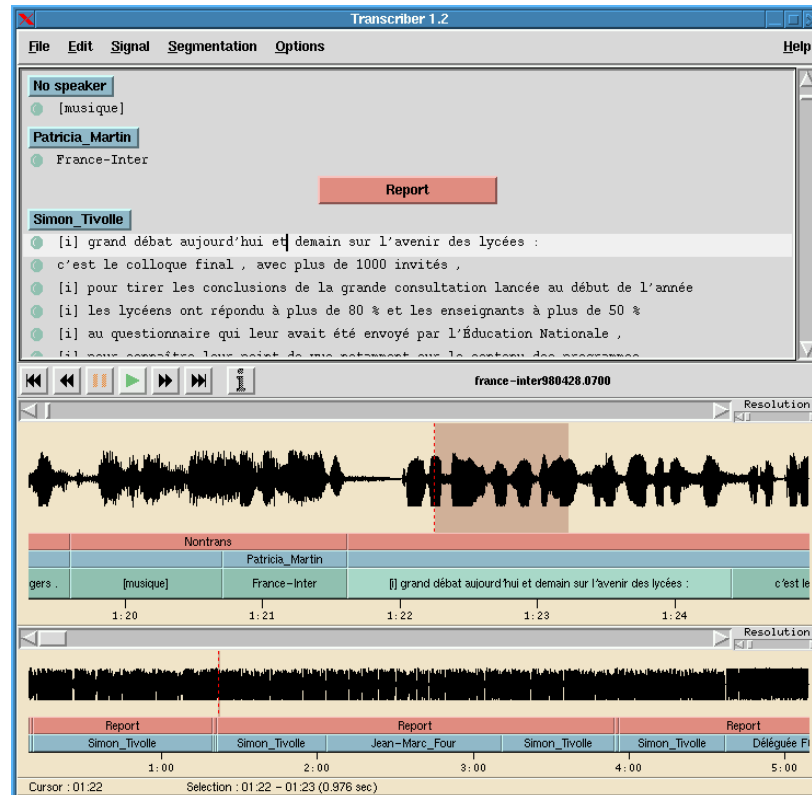
- <http://www ldc.upenn.edu/annotation/>



Linguistic Resources

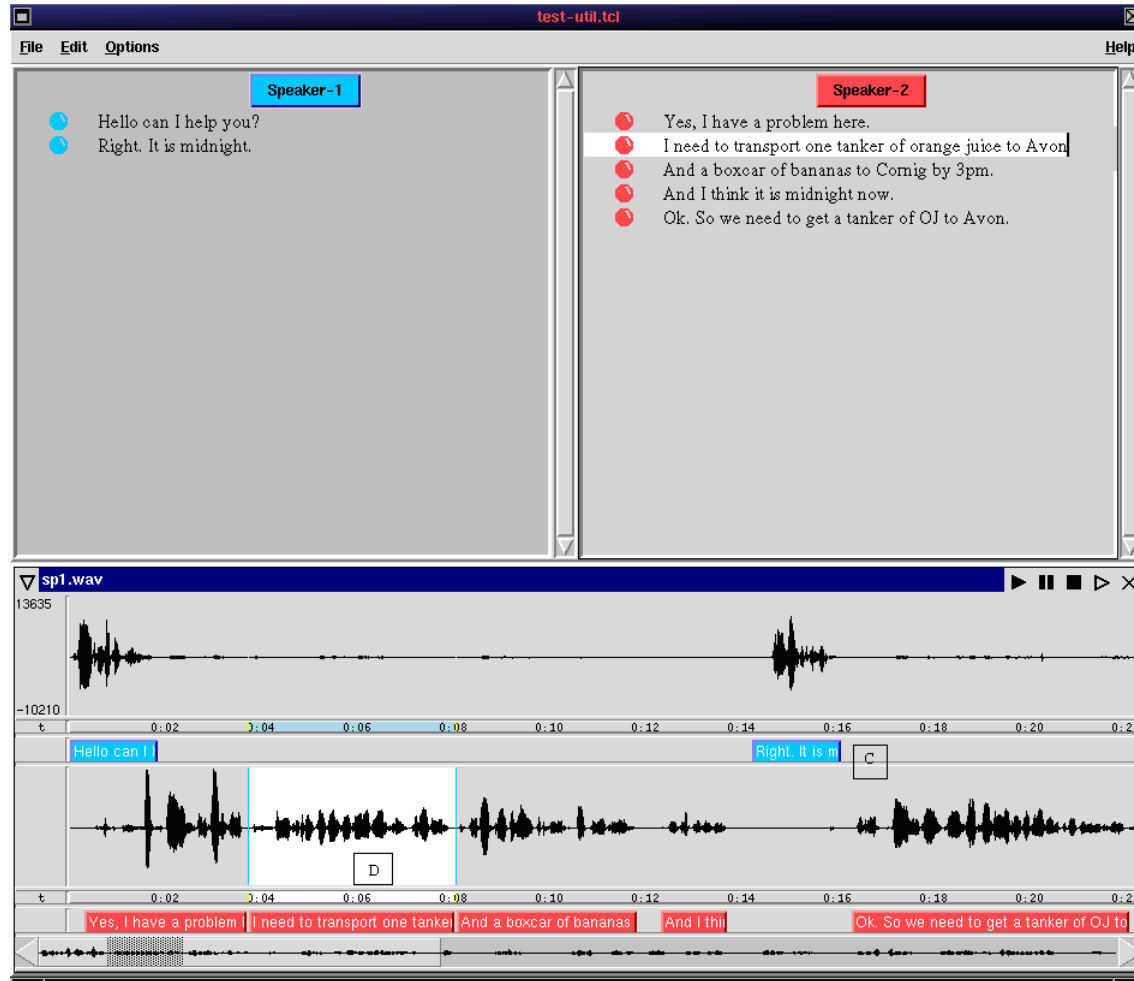
DT [UW]	<p>Alembic Workbench (David Day)</p> <p>Alembic Workbench is an SGML-based annotation system. Apart from the usual kinds of textual annotations, the workbench enables various kinds of specialized annotations including co-reference annotations (cf. the Message Understanding Conference markup rules), various kinds of user-defined inter-tag pointers, and (shortly) general template annotation (aka relations, frames, or events). The Alembic multi-lingual NLP system provides access to taggers for a wide variety of extraction levels, and applications have now been built for several languages. The software has a sophisticated visualisation component. It runs on Sun workstations and is freely distributed.</p>
FTD	<p>LACTO Linguistic Data Archiving Project (Boyd Michailovsky, John B. Lowe, Michel Jacobson)</p> <p>Projet Archivage, based at LACTO in Paris, aims to provide tools and formats for linguistic and anthropological field data. An interesting feature is the use of XML markup, with a DTD that supports transcriptions, phrasal and word-by-word interlinear translations, and audio references. Some XSL style sheets are provided that illustrate the potential power of XML markup to support web browsing for material of this type, giving access to text and sound.</p>
FP	<p>ATLAS - Architecture and Tools for Linguistic Analysis Systems (Steven Bird, David Day, John Garofolo)</p> <p>ATLAS is a joint initiative of NIST, MITRE and the LDC to build a general purpose annotation architecture and a data interchange format. The starting point is the annotation graph model, with some significant generalizations. An LREC paper describes the model.</p>
P	<p>CA - Conversational Analysis</p> <p>This page of transcriptions by Emanuel Schegloff exemplifies the style of transcription traditional among researchers working on conversational analysis.</p>
FC	<p>CES (Nancy Ide, Greg Priest-Dorman, Patrice Bonhomme)</p> <p>The Corpus Encoding Standard (CES) is a part of the EAGLES Guidelines developed for language engineering research and applications. CES is an SGML-based, TEI-conformant specification of a minimal encoding level that corpora must achieve to be considered standardized in terms of descriptive representation (marking of structural and typographic information) as well as general architecture (so as to be maximally suited for use in a text database). It also provides encoding specifications for linguistic annotation, together with a data architecture for linguistic corpora. A section of CES on speech annotation (part 6) is under construction. Projects using CES are listed here. An XML version of CES called XCES is under development.</p>
FTDPRC [WM]	<p>CHILDES (Brian MacWhinney, Steven Gillis)</p> <p>The CHILDES project provides a large database of first and second language acquisition data from over 30 languages in a constant format, called</p>

- User-friendly GUI for segmentation, transcription and transcript labeling
- Open-source; handles variety of audio, text formats; multi-platform
- Limitations
 - Requires full segmentation of audio
 - Customized for single-channel broadcast news recordings
 - Inelegant handling of overlapping speech

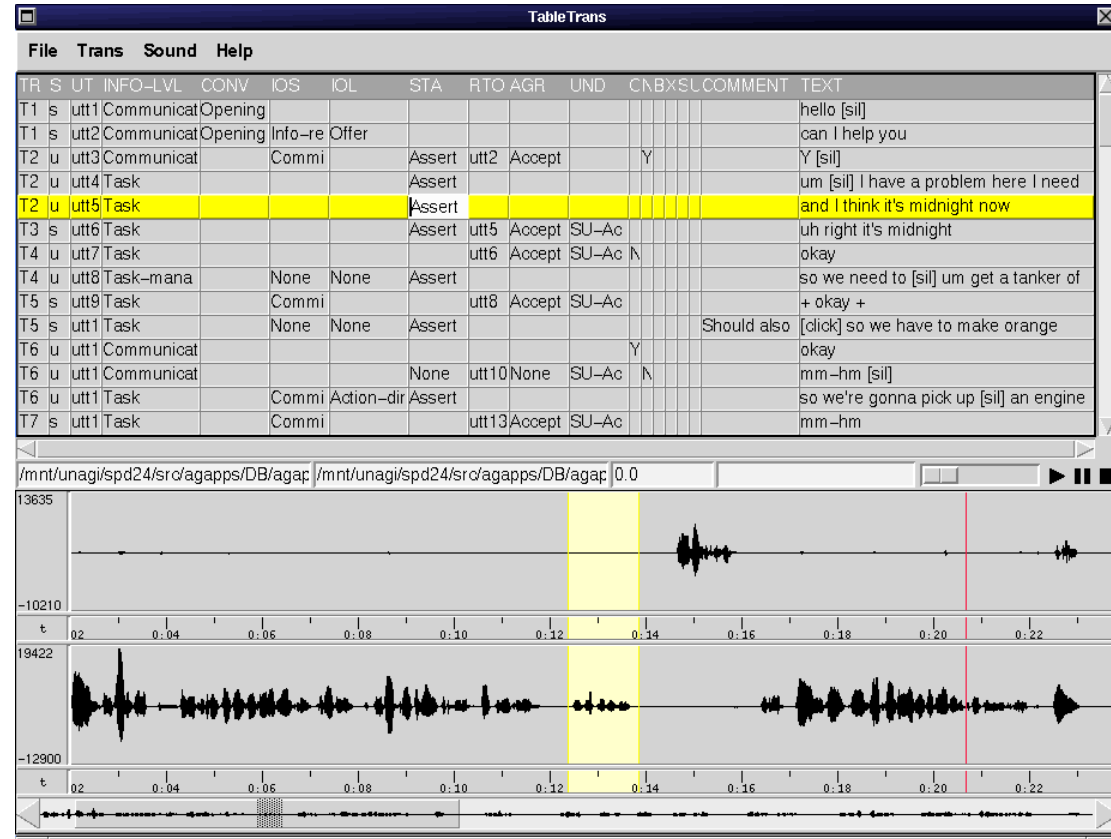


- **Annotation Graph Toolkit: agtk.sourceforge.net**
- **Suite of tools for various types of annotation**
- **Developed by LDC**
- **Open-source**
- **Handles variety of audio, text formats**
- **Multi-platform**
- **SLX Corpus Tools utilize AGTK**
 - **MultiTrans for transcription**
 - **DASLTrans (version of TableTrans) for coding**

- Transcription tool for transcribing multiparty conversations
- Similar to Transcriber but MultiTrans has one transcription panel for each channel in the signal



- Spreadsheet-style linguistic annotation tool
- User-defined features (column headings)
- Spreadsheet, audio are time-aligned
- Each row corresponds to region of audio signal
- Import existing annotation files in XML, table (csv) and LDC format
- Export annotation files in table format for further analysis



- Tools read most standard audio formats (via Snack library)
- **Transcriber**
 - Default format is .trs,
 - Accepts .typ format
 - Default segment boundary format
 - » `<Sync time="48.428"/>`
- **MultiTrans**
 - Default is LDC-style format (.lcf)
 - Segment boundary format
 - » 213.33 234.15 A:
- **TableTrans/DASLTrans**
 - Accepts MultiTrans .lcf files as input
 - » Start Time, End Time, Channel/Speaker, Transcription as first four columns
 - Accepts table format as input
 - » Tab or comma delineated spreadsheet
 - » Exclude column headers
 - Accepts ag-xml input (.aif)
 - » Native AGTK format
 - Outputs table or ag-xml format
 - » Can import table to Excel or stats packages

- **Development, production methods fully documented**
- **Complete audio available in standard format (AIFF, RIFF, SPH) 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**

DASL Project

- **Motivation**
 - 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

- Data originally created for linguistic technology development
- Selected for range of styles, availability of time-aligned transcripts
- Basic speaker demographics available

Corpus	ISBN	Minutes	Type of Data
TIMIT	1-58563-019-5	6300	Phonetically Rich Sentences
Switchboard-1	1-58563-121-3	12000	Short Conversations with Constrained Topics among Strangers

- t/d deletion case study
- Well-documented and well understood, stable indicator
- Are corpus data results comparable to traditional studies?
- Linguistic and social factors
 - morphological, preceding & following phonological environments, stress, cluster complexity
 - age, gender, education, region, race
- Results are substantially similar to previous t/d studies
 - See Strassel - NWAV2001 for discussion

DASL - Project: t/d Deletion - Netscape

File Edit View Go Communicator Help

Welcome:
ccieri

Jump to:
Next Page
DASL Home
t/d Deletion Page

Data and Annotations for Socio Linguistics

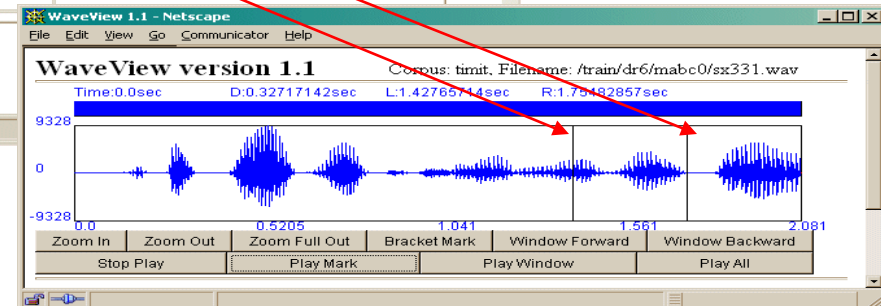
Independent Variable File:	Token File:	Annotation File:	Page:	Tokens/Page:	Total Tokens:
/Shared/TDdeletion.tag	/Shared/TDdeletion.tok	/ccieri/TDdeletion.ann	1/83	25	2059

1. ... loved to chew on the **old rag** doll.
2055, Male, New York City, 25, White, Bachelor's Degree

t/d:	<input type="radio"/> Untouched <input checked="" type="radio"/> Deleted <input type="radio"/> Retained <input type="radio"/> Unsure <input type="radio"/> NA
Morphological:	<input checked="" type="radio"/> Monomorpheme <input type="radio"/> Irregular Past <input type="radio"/> Regular Past
Preceding:	<input type="radio"/> Stop <input checked="" type="radio"/> Lateral <input type="radio"/> Rhotic <input type="radio"/> Alveolar_Nasal <input type="radio"/> Other_Nasal <input type="radio"/> Alveolar_Fricative <input type="radio"/> Other_Fricative
Following:	<input type="radio"/> Obstruent <input type="radio"/> Lateral <input checked="" type="radio"/> Rhotic <input type="radio"/> Clustering_Glide <input type="radio"/> Other_Glide <input type="radio"/> Vowel <input type="radio"/> Pause
comments:	vocalized l

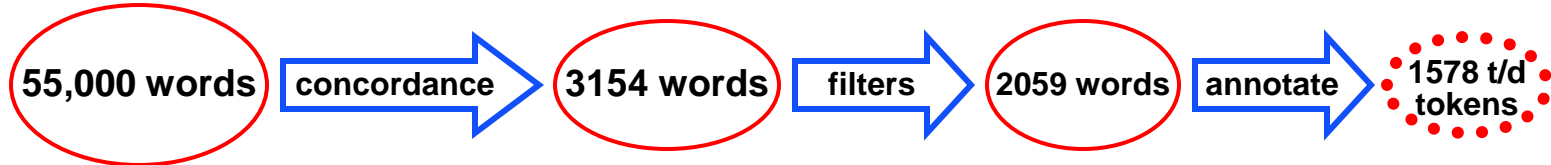
2. ... those who teach values **first abolish** cheating ...
2055, Male, New York City, 25, White, Bachelor's Degree

Document: Done

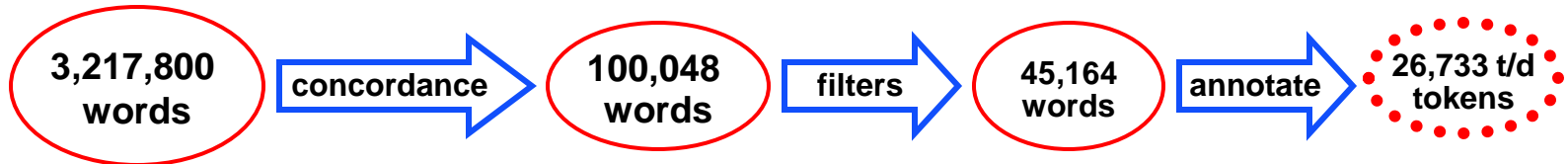


- Concordance identifies tokens of interest through regular expression query
- Filters remove additional non-tokens
- Tag set specifies factors to code
- Web browser displays annotation file
 - Listen to audio
 - Code tokens quickly
 - View demographic information
- Save results and output to text file for further analysis

TIMIT Corpus



Switchboard Corpus



- Substantially reduces overall effort
- Ensures that all tokens satisfying selection criteria are analyzed
 - More robust than manual selection, which might miss or overlook tokens

- **Value of public data**
- **Need for rigorous specifications**
 - Of collection methodology
 - Fully specified coding guidelines
- **Collaborative data development is feasible**
- **Need for end-to-end digital methodology**
 - With supporting tools and best practices
- **New data contributions from sociolinguists**
- **New collections guided by insights from DASL**

SLX Corpus

- Interviews conducted in 60s-70s primarily by Labov
- Exemplify a wide variety of regional and social dialects
- Broad spectrum of speaking styles, including spontaneous speech, narratives, responses and formal linguistic tasks
- Sessions selected by Labov where
 - Observation effects are minimized
 - Style more closely approximates vernacular
 - Sound quality is high

Speaker	Age	Speech Community	Occupation	Tapes	Others	Minutes	Words	Types
Adolphus H.	81	Hillsboro, NC	Farmer	2	3	85	9660	1494
Bobbie A.	22	Ayr, Scotland	Saw Doctor	1	1	44	8990	1769
Henry G.	60	E.Atlanta, GA	Railroad Mechanic	3	5	112	20012	2372
Jerry T.	19	Leakey, TX	Gas Attendent	2	1	66	11264	1700
Joe D.	21	Liverpool, ENG	Docker	2	0	100	19798	2515
Eddie M.	19	Liverpool, ENG	Docker	2	0	100	19798	2515
Kathy D.	15	Rochester, NY	Student	2	2	64	29001	1938
Louise A.	53	Knoxville, TN	Mother/Domestic	3	0	76	11348	1521
Rose B.	43	New York, NY (LES)	Seamstress	3	3	60	12184	1938

- **Original recordings on Nagra III or IVS with Sennheiser dynamic microphones**
- **Digitized from open reel tapes onto DAT/disk at 16bit, 44KHz sampling**
- **Monaural signal passed through 2 channels at levels differing by 20% to capture best digital copy in single pass**
- **Technician monitored recording, adjusted for sustained changes in speech levels.**
 - **Digital files show no significant clipping in the digital domain**

- **Using Transcriber tool, create**
- **One audio file for each speaker in interview**
 - Including non-target speakers (interviewer, etc) – to provide context
 - Distinguish target speaker from others, silence, non-speaker noise
 - Limitations of Transcriber in dealing with overlapping speech
- **First pass**
 - ID basic utterance boundaries
 - Process
 - » Play audio, hit <enter> at boundaries
 - » Close to 1 x Real Time
- **Second pass**
 - Finer-grained boundaries
 - Additional breakpoints at
 - » Sentence/phrase boundaries
 - » Noticeable pauses (>500ms)
 - » Breath groups

- **First pass**
 - Verbatim transcript
 - No “correction” of speakers’ grammar, pronunciation
 - Standard orthography, punctuation
 - Special conventions for
 - » Unintelligible speech
 - » Non-standard variants
 - » Speaker restarts, disfluencies, hesitations
- **Second pass**
 - Verify existing transcript
 - Revisit ((unintelligible)) sections

- **Third pass**

- Dialect-specific review

Original

Is that ((Hugh Potty))?

She done her lovely.

Bloody (()) uh.

All ((amber)) heads.

Revised

Is that how you put it?

She done a wobbler.

Bloody nutters, youse are.

All them birds.

- **Fourth pass**

- “Bleeping” of proper names

- **Segmentation, transcription process and guidelines fully documented**

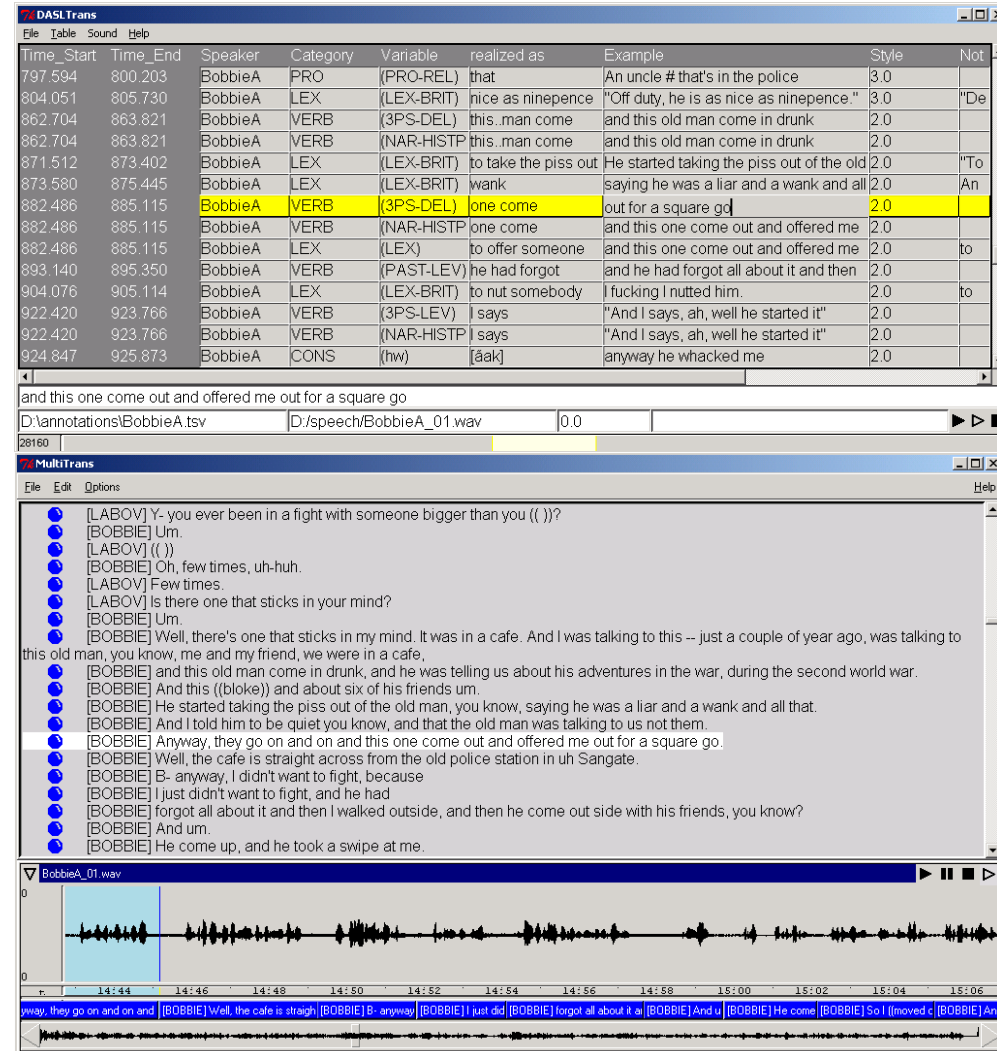
- **Identify sociolinguistic variables of interest**
 - Cross-dialectal as well as dialect-specific variables
 - » -ing, t/d deletion, negative concord
 - » habitual 'be' in AAVE; stop frication in Liverpool speech
- **Determine presence/absence of variable for each speaker**
 - Not all speakers were coded for all variables
 - Nor were speakers coded exhaustively for any variable
- **Code each variant for stylistic context**
 - Seven basic categories plus additional subtypes
 - Ranging from casual speech to formal linguistic tasks
- **Survey is experimental, non-systematic and principally descriptive**
 - Not an exhaustive account of variation in this data
 - Provides snapshot of range of intra- and inter-speaker variation in the corpus

- **Original coding done with Excel and Transcriber**
 - Code speaker, file, timestamp for each token
 - Unique token ID
 - “Realized_as” field provides IPA transcript
- **Over 150 variables surveyed**
 - Broken down by category and subtype

Variable Type	Categories	Subcategory Examples
Phonological, Phonetic, Prosodic: <i>90 variables</i>	Consonants	(DH) - voiced interdental fricative
	Front Vowels	(ae-NAS) - tensing of short-a before nasals
	Back Vowels	(ahr) - realization of /ahr/ sequence
	General Vowels	(SCHWA) - realization of schwa
	Diphthongs	(aw) - realization of /aw/
	Prosody	(RISE) - rising final intonation
Grammatical, Lexical: <i>60 variables</i>	Prepositions	(PREP-DEL) - preposition deletion
	Adjectives	(ADJ-WO) - non-standard ADJ word order
	Determiners	(DET-DEL) - determiner deletion
	Negation	(NEG-AINT) - use of ain't in neg. constructions
	Word Order	(WO-LEFTDIS) - left dislocation of initial NP
	Pronouns	(POS-LEV) - leveling of possessives to mine paradigm
	Verbs	(COP-DEL) - copula deletion
	Quantifiers	(Q-BUT) - but as quantifier
	Agreement	(PLURAL) - singular ending on plural noun

- **Optimized for exploration of SLX Corpus**
- **SLX Corpus Browser**
 - interactive assistant to step through corpus documentation, transcript and speech files and sociolinguistic variable survey
- **MultiTrans**
 - provides merged or individual-speaker view SLX transcripts and audio
- **DASLTrans**
 - interactive view of the sociolinguistic variable survey
- **Several additional components**
 - Transcriber
 - Fonts
 - Audio packages

- Unite functions of MultiTrans and DASLTrans to allow segmentation, transcription, coding within single tool
- Handle multi- or single-channel audio, including multi-speaker on one channel
- All annotations synchronized to single audio file
- Multiple audio, text formats supported
- Output results in table format for further analysis
- Extensible via distributed source code
- Multi-platform
- Freely available



The image shows two software windows. The top window, titled 'DASLTrans', displays a table of linguistic annotations. The bottom window, titled 'MultiTrans', shows a transcription of a speech sample with speaker labels and a corresponding audio waveform.

Time_Start	Time_End	Speaker	Category	Variable	realized as	Example	Style	Not
797.594	800.203	BobbieA	PRO	(PRO-REL)	that	An uncle # that's in the police	3.0	
804.051	805.730	BobbieA	LEX	(LEX-BRIT)	nice as ninepence	"Off duty, he is as nice as ninepence."	3.0	"De
862.704	863.821	BobbieA	VERB	(3PS-DEL)	this..man come	and this old man come in drunk	2.0	
862.704	863.821	BobbieA	VERB	(NAR-HISTP)	this..man come	and this old man come in drunk	2.0	
871.512	873.402	BobbieA	LEX	(LEX-BRIT)	to take the piss out	He started taking the piss out of the old	2.0	"To
873.580	875.445	BobbieA	LEX	(LEX-BRIT)	wank	saying he was a liar and a wank and all	2.0	An
882.486	885.115	BobbieA	VERB	(3PS-DEL)	one come	out for a square go	2.0	
882.486	885.115	BobbieA	VERB	(NAR-HISTP)	one come	and this one come out and offered me	2.0	
882.486	885.115	BobbieA	LEX	(LEX)	to offer someone	and this one come out and offered me	2.0	to
893.140	895.350	BobbieA	VERB	(PAST-LEV)	he had forgot	and he had forgot all about it and then	2.0	
904.076	905.114	BobbieA	LEX	(LEX-BRIT)	to nut somebody	I fucking I nutted him.	2.0	to
922.420	923.766	BobbieA	VERB	(3PS-LEV)	I says	"And I says, ah, well he started it"	2.0	
922.420	923.766	BobbieA	VERB	(NAR-HISTP)	I says	"And I says, ah, well he started it"	2.0	
924.847	925.873	BobbieA	CONS	(hw)	[aak]	anyway he whacked me	2.0	

and this one come out and offered me out for a square go

D:\annotations\BobbieA.tsv | D:\speech\BobbieA_01.wav | 0.0

28160

MultiTrans

File Edit Options Help

[LABOV] Y- you ever been in a fight with someone bigger than you (())?

[BOBBIE] Um.

[LABOV] (())

[BOBBIE] Oh, few times, uh-huh.

[LABOV] Few times.

[LABOV] Is there one that sticks in your mind?

[BOBBIE] Um.

[BOBBIE] Well, there's one that sticks in my mind. It was in a cafe. And I was talking to this -- just a couple of year ago, was talking to this old man, you know, me and my friend, we were in a cafe,

[BOBBIE] and this old man come in drunk, and he was telling us about his adventures in the war, during the second world war.

[BOBBIE] And this ((bloke)) and about six of his friends um

[BOBBIE] He started taking the piss out of the old man, you know, saying he was a liar and a wank and all that.

[BOBBIE] And I told him to be quiet you know, and that the old man was talking to us not them.

[BOBBIE] Anyway, they go on and on and this one come out and offered me out for a square go.

[BOBBIE] Well, the cafe is straight across from the old police station in uh Sangate.

[BOBBIE] B- anyway, I didn't want to fight, because

[BOBBIE] I just didn't want to fight, and he had

[BOBBIE] forgot all about it and then I walked outside, and then he come out side with his friends, you know?

[BOBBIE] And um.

[BOBBIE] He come up, and he took a swipe at me.

BobbieA_01.wav

0

14:44 14:46 14:48 14:50 14:52 14:54 14:56 14:58 15:00 15:02 15:04 15:06

way, they go on and on and [BOBBIE] Well, the cafe is straight [BOBBIE] B- anyway [BOBBIE] I just did [BOBBIE] forgot all about it [BOBBIE] And u [BOBBIE] He come [BOBBIE] So I (moved c [BOBBIE] An

DEMO