Sharable Resources for Sociolinguistic Research

Christopher Cieri

Stephanie Strassel

William Labov

{ccieri@ldc.upenn.edu,
strassel@ldc.upenn.edu,
labov@central.cis.upenn.edu}

Data and Annotations for Socio Linguistics

DASL: An investigation of best practices in the use of digital data & tools to support empirical linguistic inquiry and documentation. Funded by the National Science Foundation (BCS-998009, KDI, SBE) under the Talkbank project (www.talkbank.org) and by the Linguistic Data Consortium (www.ldc.upenn.edu).

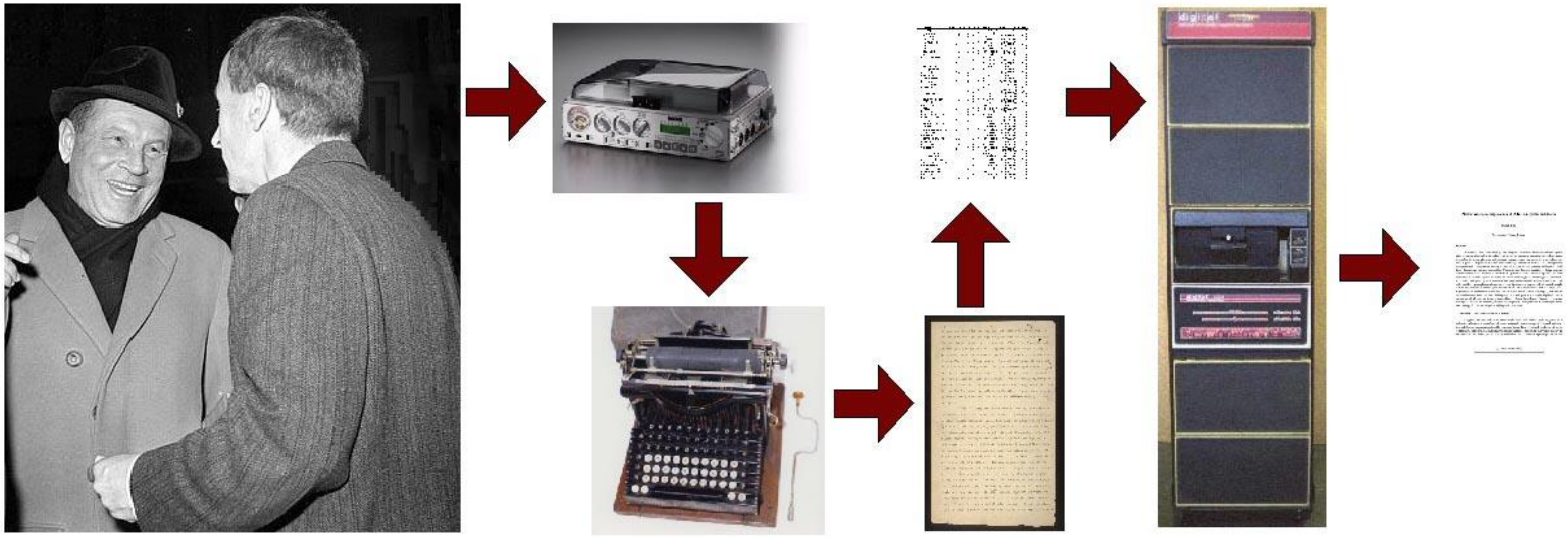
1. Vision for empirical, quantitative research that is robust, collaborative, and accountable.
2. Review of previous DASL (www.ldc.upenn.edu/Projects/DASL) analyses using corpora; discussion of resources and tools developed.
3. *SLX Corpus of Classic Sociolinguistic Interviews* conducted by William Labov and his students available for distribution 1/2003.

National Science Foundation
WHERE DISCOVERIES BEGIN
science

LDC

Methodologies

1970



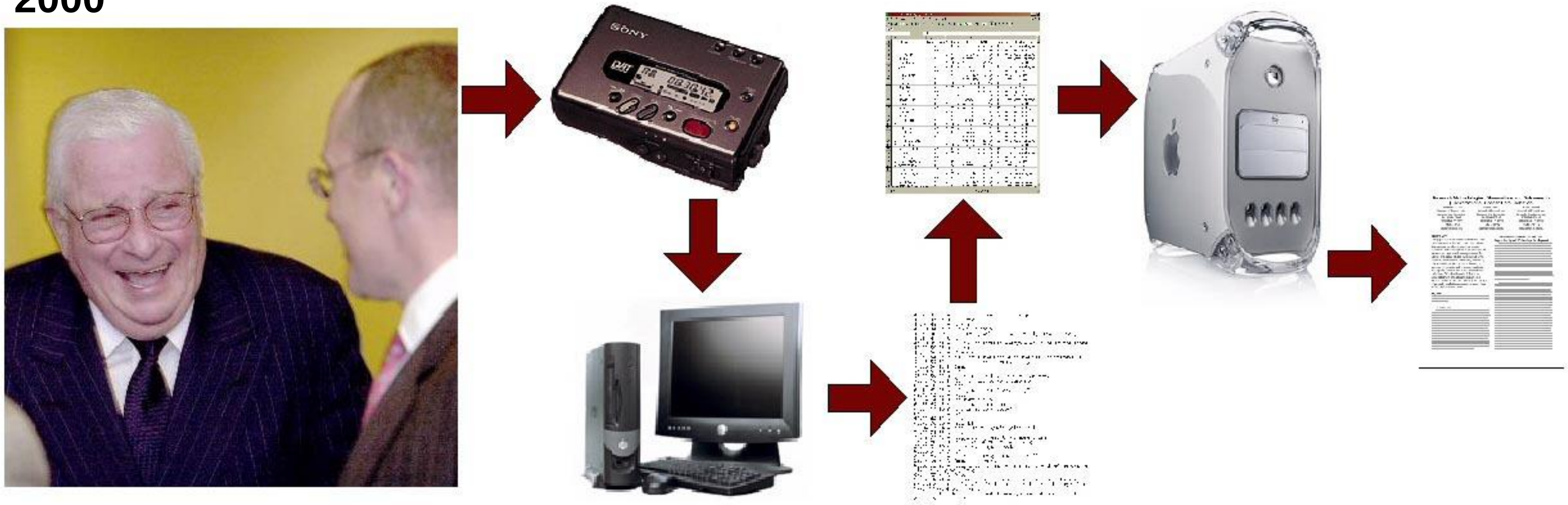
Interviews are recorded but not always transcribed; even then transcripts are often only partial.

Analytical tools are not integrated.

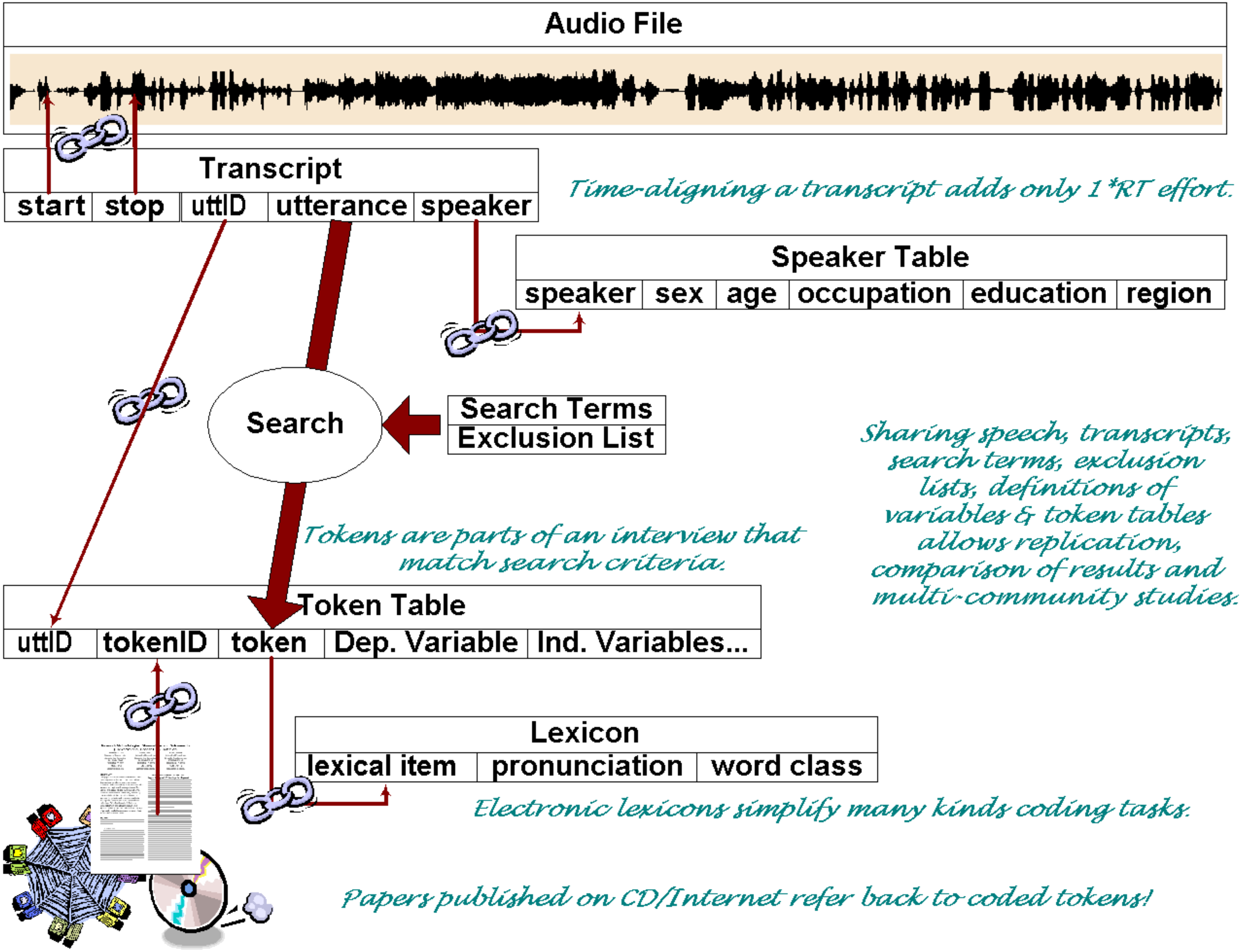
The presentation is an independent artifact.

After 30 years of technological advance, our use of data is largely unchanged; only the components differ.

2000



2002-



Audio File

Transcript				
start	stop	uttID	utterance	speaker

*Time-aligning a transcript adds only 1*RT effort.*

Speaker Table					
speaker	sex	age	occupation	education	region

Search

Search Terms
Exclusion List

Sharing speech, transcripts, search terms, exclusion lists, definitions of variables & token tables allows replication, comparison of results and multi-community studies.

Tokens are parts of an interview that match search criteria.

Token Table				
uttID	tokenID	token	Dep. Variable	Ind. Variables...

Lexicon		
lexical item	pronunciation	word class

Electronic lexicons simplify many kinds coding tasks.

Papers published on CD/Internet refer back to coded tokens!

Corpora

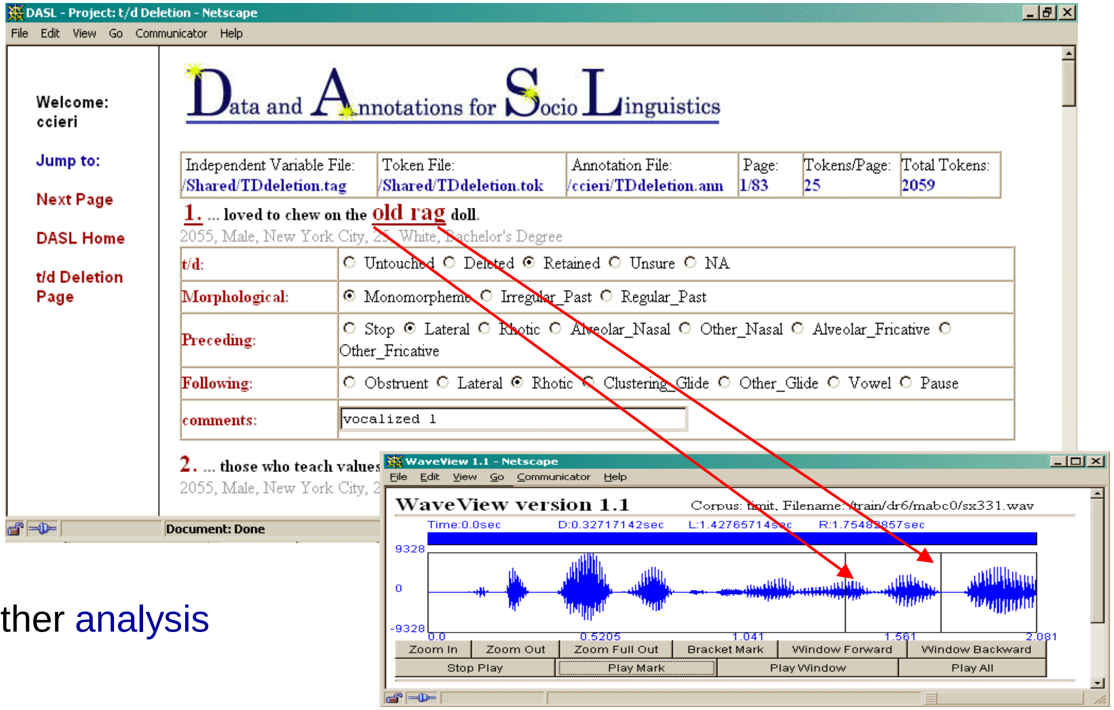
- Originally created for linguistic technology development
- Selected for range of styles, availability of time-aligned transcripts
- Contain basic speaker demographics

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*

Tools

- **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


DASL - Project: t/d Deletion - Netscape
File Edit View Go Communicator Help
Document: Done

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:	○ Untouched ○ Deleted ● Retained ○ Unsure ○ NA
Morphological:	● Monomorpheme ○ Irregular_Past ○ Regular_Past
Preceding:	○ Stop ● Lateral ○ Rhotic ○ Alveolar_Nasal ○ Other_Nasal ○ Alveolar_Fricative ○ Other_Fricative
Following:	○ Obstruent ○ Lateral ● Rhotic ○ Clustering_Glide ○ Other_Glide ○ Vowel ○ Pause
comments:	vocalized l

2. ... those who teach values

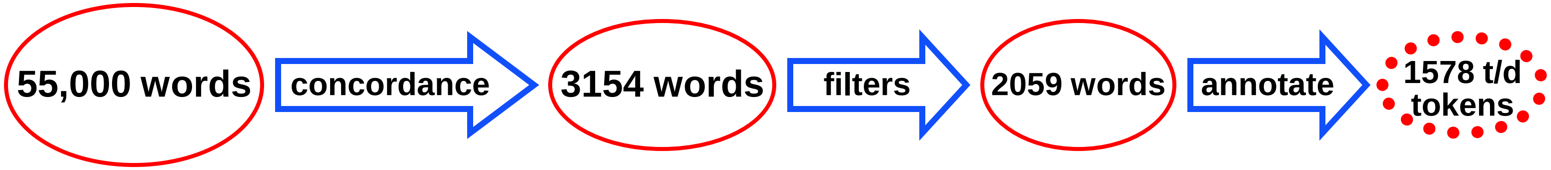
2055, Male, New York City, 2

WaveView version 1.1

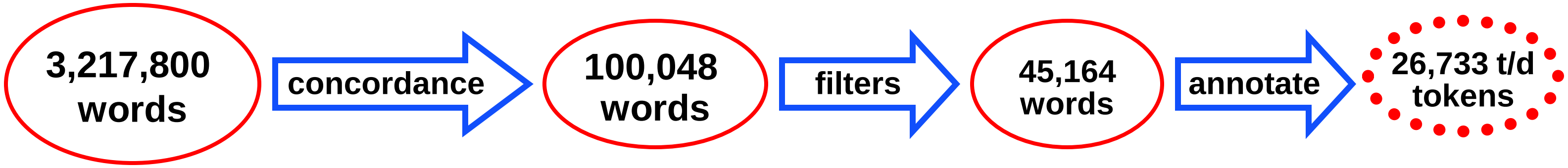
Corpus: timit, Filename: /train/dr6/mabc0/sx331.wav

Time:0.0sec D:0.32717142sec L:1.42765714sec R:1.75482857sec

TIMIT Corpus

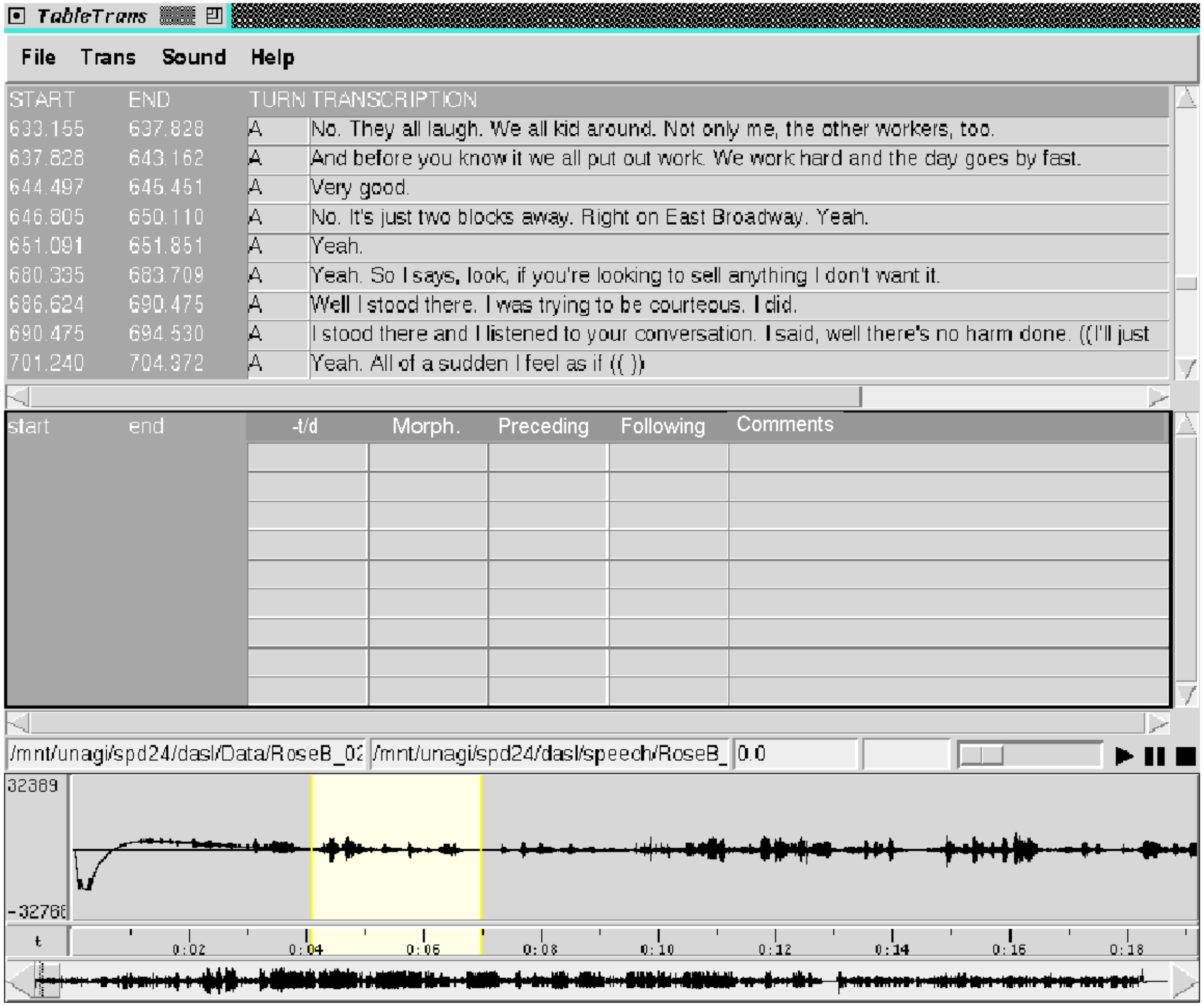


Switchboard Corpus



DASLTrans - transcription and coding tool

- **Audio:** handles arbitrary length audio files
- **Text:** AG compliant XML
- **Tag set:** user defined
- **Integration**
  - Listen to audio
  - Segment easily
  - Transcribe directly maintaining time-alignment
  - Code using same tool
  - Output results in table format for further analysis
- **Free**
- **Extensible** via distributed source code



START	END	TURN	TRANSCRIPTION
633.155	637.828	A	No. They all laugh. We all kid around. Not only me, the other workers, too.
637.828	643.162	A	And before you know it we all put out work. We work hard and the day goes by fast.
644.497	645.451	A	Very good.
646.805	650.110	A	No. It's just two blocks away. Right on East Broadway. Yeah.
651.091	651.851	A	Yeah.
680.335	683.709	A	Yeah. So I says, look, if you're looking to sell anything I don't want it.
686.624	690.475	A	Well I stood there. I was trying to be courteous. I did.
690.475	694.530	A	I stood there and I listened to your conversation. I said, well there's no harm done. ((I'll just
701.240	704.372	A	Yeah. All of a sudden I feel as if (( ))

start	end	-t/d	Morph.	Preceding	Following	Comments

/mnt/unagi/spd24/dasl/Data/RoseB_02 /mnt/unagi/spd24/dasl/speech/RoseB_ 0.0

SLX Corpus

- **SELECTION: Sessions selected by William Labov where observation effects are minimized, style more closely approximates vernacular and sound quality is high. Interviews conducted between 1963 and 1973 primarily by Labov.**

- **Sessions 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.**

Speaker	Age	Speech Community	Occupation	Tapes	Others	Minutes	Words	Types
Adolphus H.	82	Hillsboro, NC	Farmer	2	3	85	9660	1494
Bobbie A.	32	Ayr, Scotland	Saw Doctor	1	1	44	8990	1769
Henry G.	61	E.Atlanta, GA	Railroad Mechanic	3	5	112	20012	2372
Jerry T.	20	Leakey, TX	Gas Attendent	2	1	66	11264	1700
Joe D.	21	Liverpool, ENG	Docker	2	0	100	19798	2515
Eddie M.	20	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.	36	New York, NY (LES)	Seamstress	3	3	60	12184	1938

Segmentation

-one audio file per speaker

-distinguish target speaker from other speakers, silence, non-speaker noise

First pass segmentation

- ID basic utterance boundaries
- Play audio
- Hit <enter> to create boundary
- Close to 1X real time

Second pass segmentation

- Finer-grained boundaries
- Additional breakpoints at
  - Sentence/phrase boundaries
  - Noticeable pauses (>500ms)
  - Breath groups

Transcription

First pass segmentation

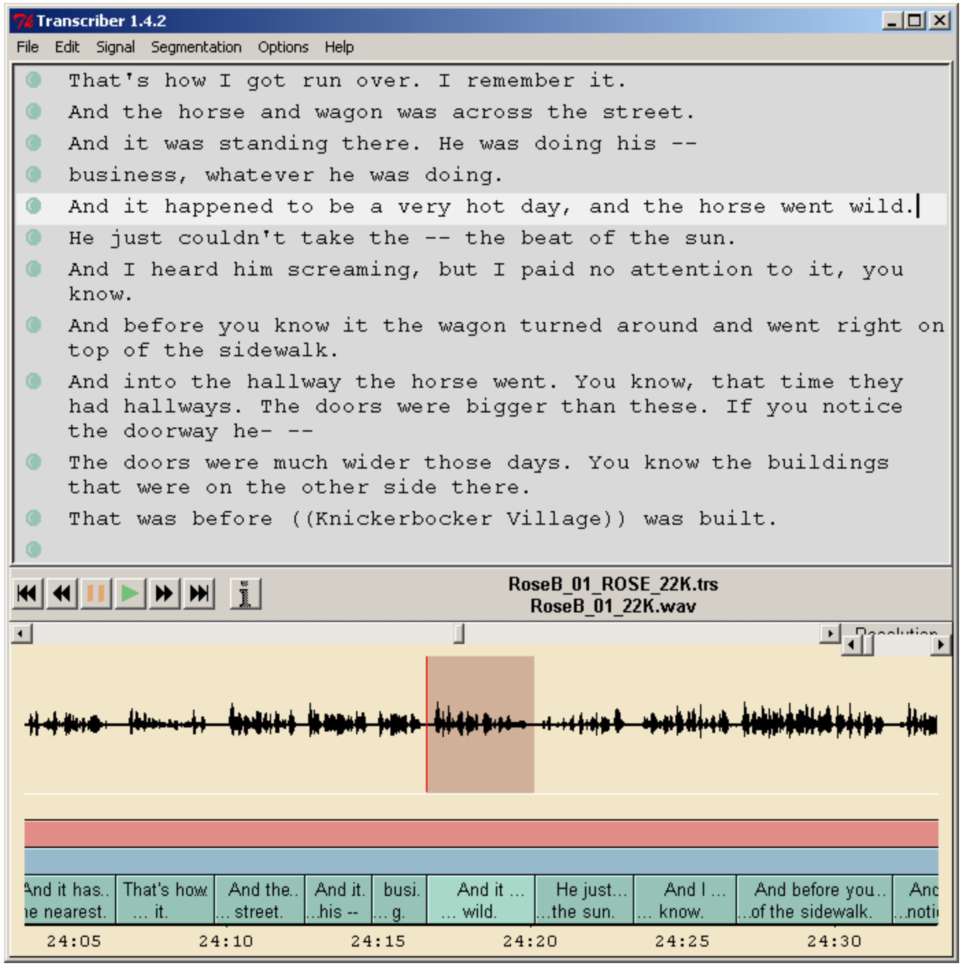
- Verbatim transcript
- No “correction” of speakers’ grammar, pronunciation
- Standard orthography, punctuation
- Special conventions for
  - -unintelligible speech
  - -non-standard lexical items
  - -speaker restarts

Second pass segmentation

- Verify existing transcript
- Revisit ((unintelligible)) sections

Third pass segmentation - specialized

- Dialect-specific review



That's how I got run over. I remember it.

And the horse and wagon was across the street.

And it was standing there. He was doing his --

business, whatever he was doing.

And it happened to be a very hot day, and the horse went wild.

He just couldn't take the -- the beat of the sun.

And I heard him screaming, but I paid no attention to it, you know.

And before you know it the wagon turned around and went right on top of the sidewalk.

And into the hallway the horse went. You know, that time they had hallways. The doors were bigger than these. If you notice the doorway he- --

The doors were much wider those days. You know the buildings that were on the other side there.

That was before ((Knickerbocker Village)) was built.

Is that **((Hugh Potty))?**	→	Is that **how you put it?**
She done **her lovely**!	→	She done **a wobbler**!
Bloody **(( )) uh**.	→	Bloody **nutters, youse are**.
All **((amber)) heads**.	→	All **them birds**.

What Variables Appear in the Corpus?

First pass – variable survey to encourage future research

- Examine general and dialect-specific variables
- Determine presence/absence of each variable for all speakers

Variable Type	Categories	Subcategory Examples
Phonological, Phonetic, Prosodic: ***90 variables***	**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: ***60 variables***	**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

Variable profile, examples of each variable for each speaker

Legal, Ethical Issues

- Collection now requires informed consent of subjects.
- Shared data must protect subjects’ anonymity.
- Distribution requires permission from copyright holder.
- SLX Corpus of Classic Sociolinguistic Interviews may be used only for linguistic education, research and technology development.
- Researcher is responsible to make best effort to ensure that uses of shared data respect the dignity of subjects.
- Need community specific code of ethics for shared data.

Publication Standards

- Development, productions 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.

- Shared data resources and tools
  - enables comparison of results across studies and over time
  - serves as 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
- **SLX Corpus of Classic Sociolinguistic Interviews**
  - classic interviews, cited in literature
  - demonstrate best practice in conducting sociolinguistic interviews
  - represent variety of speech communities
  - effect of observation minimized
  - sound quality high
  - demonstrate best practice in digitization, segmentation, transcription
  - teaching tool for sociolinguists
  - stable benchmark for training/comparing transcription and coding
  - an example of a multi-community sociolinguistic corpus