

Towards Best Practices in Sociophonetics:

Robust, Digital, Empirical, Reproducible Sociolinguistic Methodology

Christopher Cieri, Stephanie Strassel

Linguistic Data Consortium

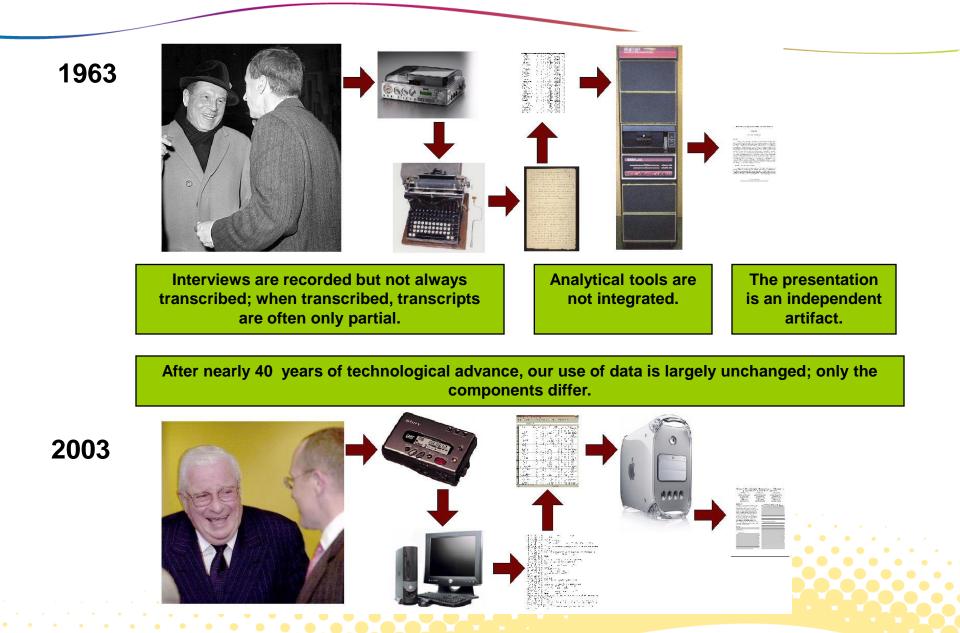


History

- 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?





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



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.



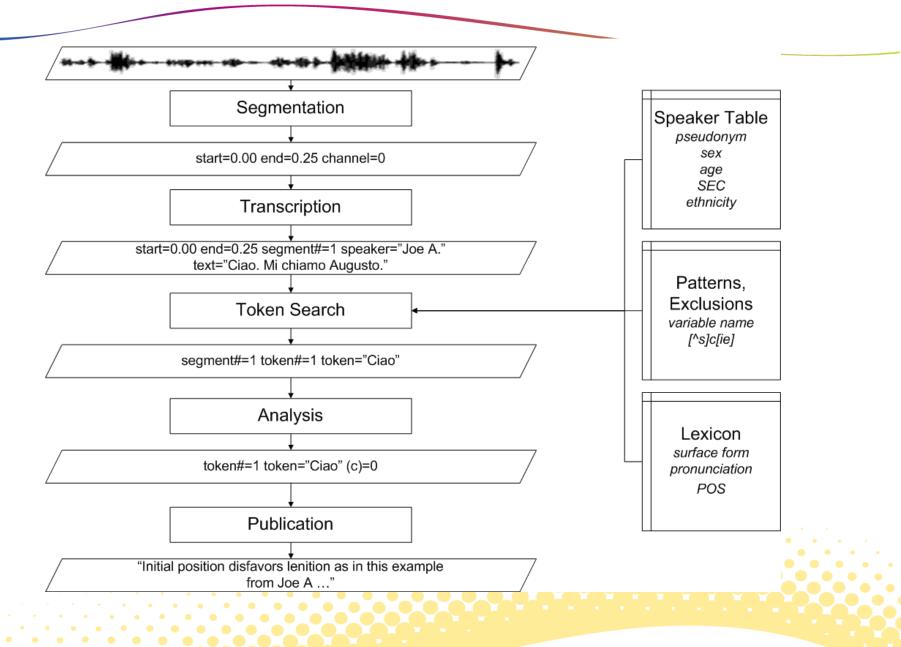
Optimal Methods

- 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







- 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



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

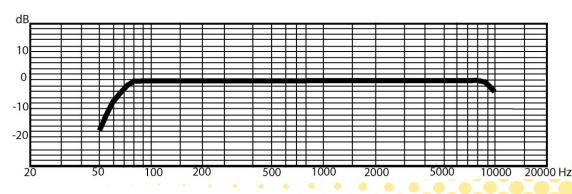
- temperature, medium treated as fixed
- speakers not selected for ability to sit still and speak clearly
- Sometimes Controllable
 - external noise
 - reflection
 - distance
 - subject to microphone
 - subject to interviewer

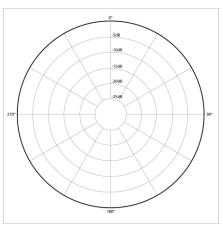


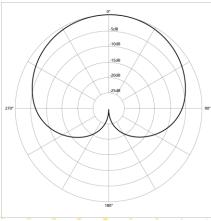
Collection

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)

Cieri, Strassel: Robust, Digital, Empirical, Reproducible Sociolinguistic Methodology, NWAV 39 November 4-6, 2010 San Antonio, Texas



Recorders

15

- 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

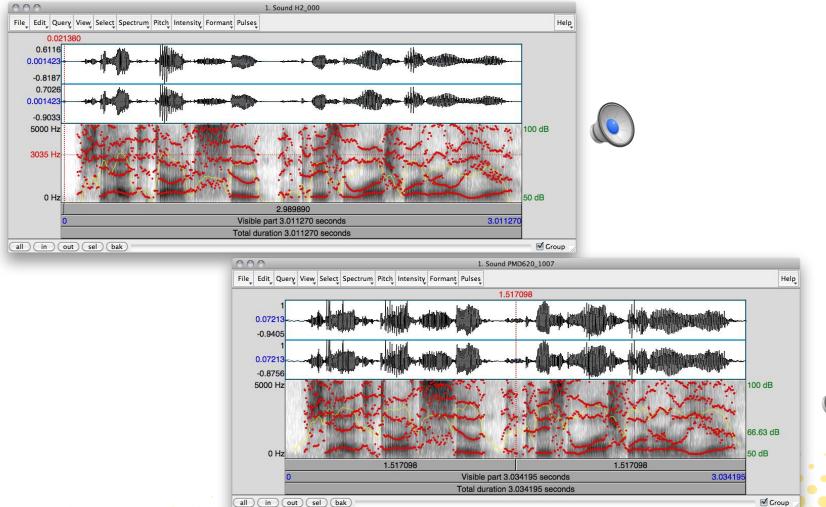




- quality generally very good
- factory settings slightly too sensitive for test case
 - some clipping



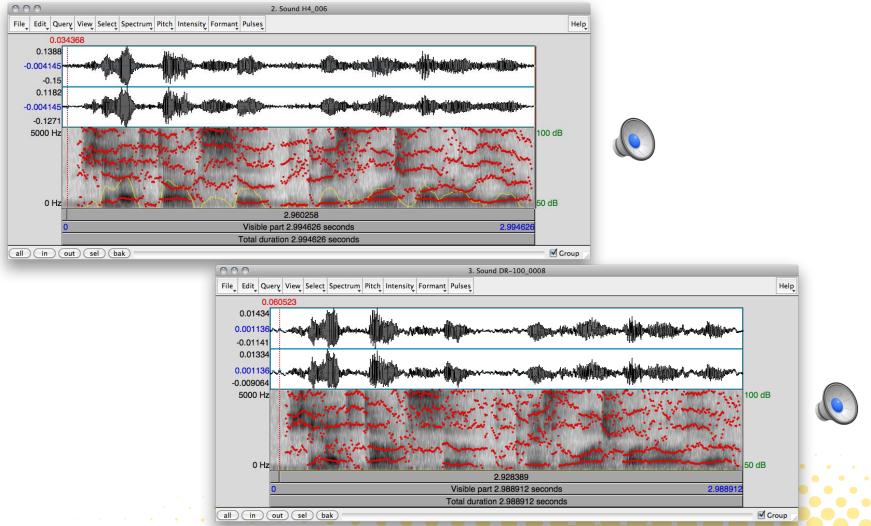
inexpensive recorders, well placed produce good results





20

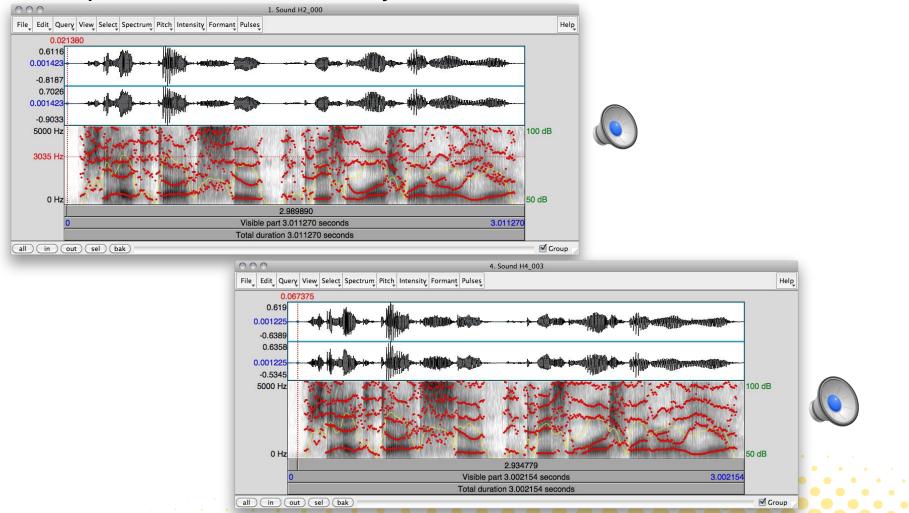
expensive recorders poorly placed produce poor results





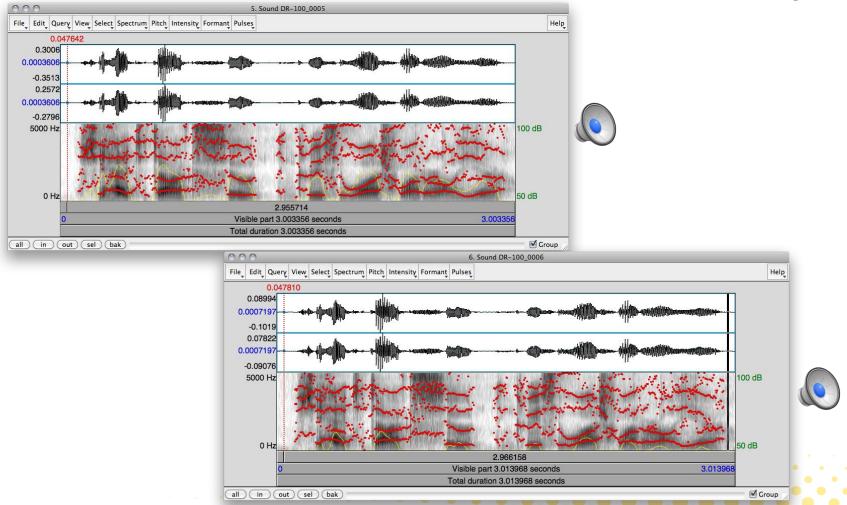
21

expensive recorders may not warrant extra cost





difference between unidirectional and omnidrectional 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)

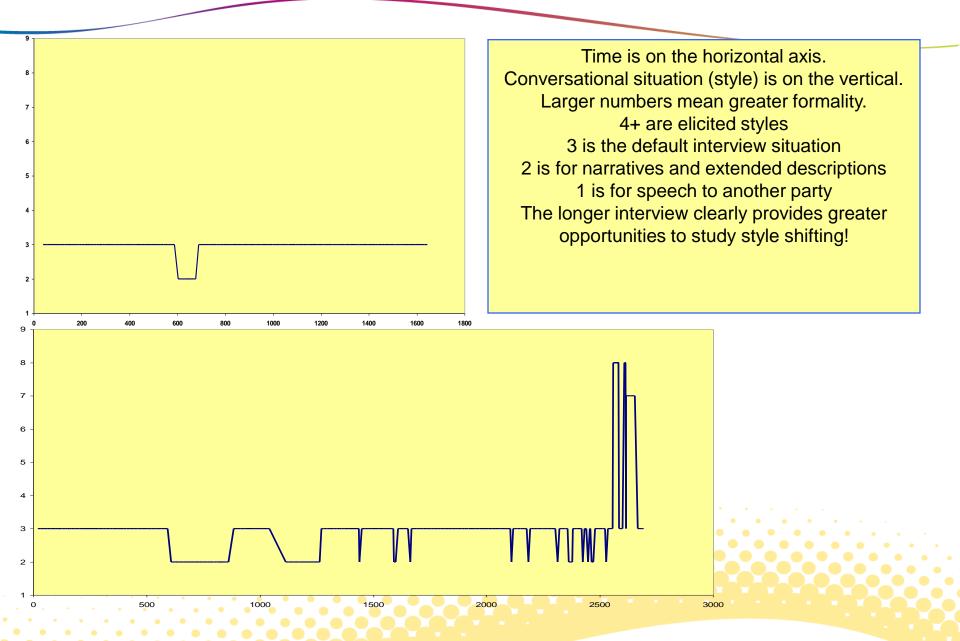


Segmentation

- 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







Transcription

 Stoker '97 provides early justification for transcription in related field





 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



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%)



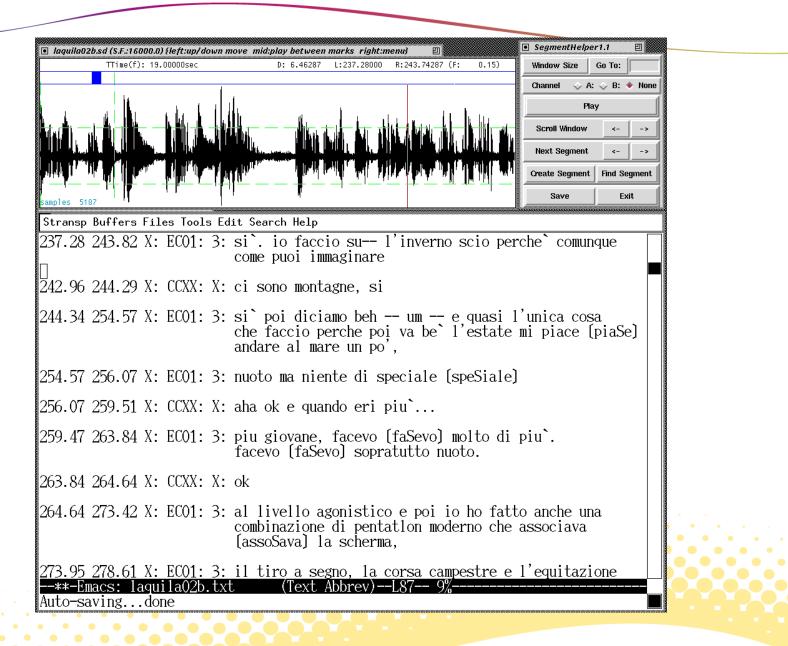
- ASR Mediated Transcription experiment
 - native speaker trained Dragon Naturally Speaking Italian
 - listened to tapes via foot-pedal controlled device
 - repeated each utterance to Naturally Speaking & corrected its mistakes

	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
 - <u>http://cmuspice.org/</u>, Prof. Tanja Schutz, CMU



Strans +





Transcriber

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

	B			<u> </u>	•		•
X				Transcriber 1.2			
<u>F</u> ile	<u>E</u> dit <u>S</u> ignal	<u>S</u> egmentation	<u>O</u> ptions				<u>H</u> elp
No	speaker						
	[musique]						-
Pat	tricia Martin						
	France-Inter						
				Devent			
				Report			
Sin	non_Tivolle						
			•	sur l'avenir des	lycées :		
				1000 invités ,			
	-		-	rande consultatio			
	-	-	-	80 % et les ensei			
	• • •	•		é envoyé par l'Éd			
	and and a fi				rance - inter980428.070		
					rance-inter900420.070		Resolution
ЫI							
•		ladaana Analana Analana			***	++++	
		Nontrans	inin bilandin				
	fraueigu		ricia_Martin ance-Inter	51 arand dábat avi	ourd'hui et demain sur l'a		o/oot lo
gers	. (musiqu		srice-initer	[i] grand debat aut		wernir des lycees .	c'est le
	1:20	1:3	21	1:22	1:23	1:24	Resolution
M_							
in the second	المانكين البنائي ودائل والتنا	hanna a Mille Milling and an Ling by a				فالترجيل المفتحة المحمد المحمد	
1	1 1 1 1			ب بي الماري الم الم الم الم			
	Report			Report			Report
	Simon_Tivolle	Simor	_Tivolle	Jean-Marc_Four	Simon_Tivolle	Simon_Tivolle	Déléguée Fi
	1:	00	2:00	3:0	0	4:00	5:00
-		0.1					

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/

Xtrans: 571959_a.tdf	_ ∂ ×
File Edit View Tools	
Respondeni when I was in high school that integrated Respondeni the the schools I was taking algebra and there's a lot of stories about how I was mistreated and stuff. Respondeni But I was having difficulty with with algebra. Respondeni And I was sitting at the kitchen table trying to do my homework. Respondeni And I said I got frustrated and said I just can't figure this out, I'm just Respondeni And I said I got frustrated and said U just can't figure this out, I'm just Respondeni And I said (tit's this)) algebra, and he said well let me look at it. I said Dad this is algebra. Respondeni They didn't even have algebra in your day. Respondeni And u h and I went to sleep, I went to bed. Respondeni And around four o'clock that morning he woke me up, he said Respondeni And I said what what's wrong. He said let's talk about this algebra. Respondeni And I said what what's wrong. He said let's talk about this algebra. Respondeni And He we went over algebra. He taught me algebra. Respondeni And then he explained the problems to me, Respondeni And then he explained them, and understand them. Respondeni And to this day I live my life trying to be half the man my father was.	Interviewer (F) Respondent (M) 70S VAS VOS VAS VSI ESIQ MRG VIRGQ RS LAS
2:38.1585 2:20.6893 2:25.3747 4.6855	
2:05.0 2:10.0 2:15.0 2:20.0 2:25.0 2:30.0 2:35.0 2:45.0 2:45.0 2:50.0 2:55	.0 3:00.0
eady	•

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

http://	www.lat-mpi.eu/tools/elan
000	Elan – RoseB_01_ROSE.eaf
File Edit Annotation Tier T	ype Search View Options Window Help
Gri	d Text Subtitles Audio Recognizer Metadata Controls
Recognizer: Silence Recognizer MP	I−PL 🗘 File(s): 🗹 RoseB_01.wav
Selections	Parameters
Add Remove	Minimal Silence Duration: 400 Milliseconds
Add Tier	
Add Hel	
Segmentations	Progress
Create Tier(s)	Start Report
	Selection: 00:07:24.432 - 00:07:28.766 4334 Image: Selection in the image: Selection Mode Loop Mode Image: Selection in the image: Selection Mode Loop Mode
00:07:25.000	00:07:26.000 00:07:27.000 00:07:28.000 00:07:29.000 00:07:30.000 00:07:31.000 00:07
00:07:25.000	00:07:26.000 00:07:27.000 00:07:28.000 00:07:29.000 00:07:30.000 00:07:31.000 00:07
Sections [1]	
Unspecified [576]	why is it that they have a certain kind of an accent. But an accent that
•	

- 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?



- 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 nontokens
 - 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

DASL - Project: t/d D			_ 8 ×							
Welcome: ccieri	Data and A	Innotations for Socio Linguistics	ĺ							
Jump to:	Independent Variable ////////////////////////////////////		ns:							
Next Page DASL Home		n the <mark>Old Fag</mark> doll. : City, 25, White, Bachelor's Degree								
t/d Deletion	t/d: Q Untouched O Deleted © Retained O Unsure O NA									
Page	Morphological: © Monomorpheme & Irregular_Past © Regular_Past									
	Preceding: O Stop O Lateral O Rhotis O Alveolar_Nasal O Other_Nasal O Alveolar_Fricative O Other_Fricative									
	Following: O Obstruent O Lateral Rhotic O Chustering_Glide O Other_Glide O Vowel O Pause									
	comments: vocalized 1									
	2055, Male, New 1	WaveView 1.1 - Netscape e Edit View Go Communicator Help VaveView version 1.1 Corpus: timit, Filename: /transdr6/mabc0/sx331.wav								
§ =0=	Document: Done	Time:0.0sec D:0.32717142sec L:1.42765714sec R:1.75482857.sec								
	0	328 328 328 328 328 328 328 0.0 Com Out Zoom Full Out Bracket Mark Window Forward Window Backward Stop Play Play Mark Play Window Play All	IB1							

Cieri, Strassel: Robust, Digital, Empirical, Reproducible Sociolinguistic Methodology, NWAV 39 November 4-6, 2010 San Antonio, Texas

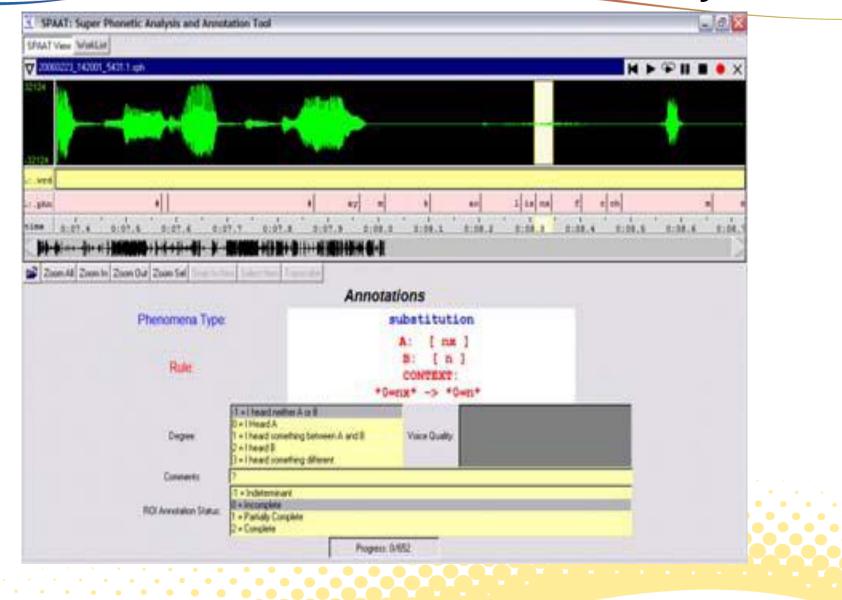


TableTrans

🗉 TableT	rans 🎆 🗉								
File Tra	ans Sound	Help							
START	END	TURN TRAN	ISCRIPTION						
533.155	637.828					ily me, the oth			
37.828	643.162		~	ow it we all pi	ut out work. V	Ve work hard	and the da	ly goes by fas	ət.
44.497	645.451	A Very							
46.805	650.110			ocks away. Rij	ght on East B	Iroadway. Yea	uh.		
51.091	651.851	A Yeah.							
80.335	683.709					anything I do	n't want it.		
86.624	690.475			I was trying t					
90.475	694.530					tion. Isaid, we	ell there's n	io harm done	. ((I'll just
01.240	704.372	A Yeah.	All of a sude	len I feel as if	(())				S
1									
	end	-t/d	Morph.	Preceding	Following	Comments			4
									N
nnt/unag	i/spd24/dasl/D	ata/RoseB_02	//mnt/unagi/	spd24/dasl/sp	eech/RoseB	_]0.0			
389									
-				-+ ++ ++				********* ****	
32768									
t	0:02	0:04	' 0:06	0:08	' 0:10	0:12	 0:14	' 0:16	' ' 0:18
الم معرفة الم	لى بى يى	L	ی میں مار اور اور اور اور اور اور اور اور اور ا		A A A A A A A A A A A A A A A A A A A	ىدىرى مەركىيە ئۇرۇشىيە يەركى		ير ان ريسين جرمر الأيريا	
							-12 - 12 - 12 - 14		



SPAAT (Super Phonetic Annotation & Analysis Tool)



Formant Analysis

Pattern e/BB

F3: 2611.825193

375 io stavo ??? que

450 ci vorrebbe anda

355 perché la legge

374 io stavo ??? que 490 non ti da questa

492

428

531

108

562

607

778

fa la legge che

è diverso -- è d

(SE01: eccesso v

poi è successo u

qua sorpassa des

(SB01:ecco .) %b

poi è bello legg

429 piano piano stia

699 come se tu doves

506 e che siamo , be

488 non ti da questa

232 colazione io ???

347 tutti prati bell

106 che qua non avev

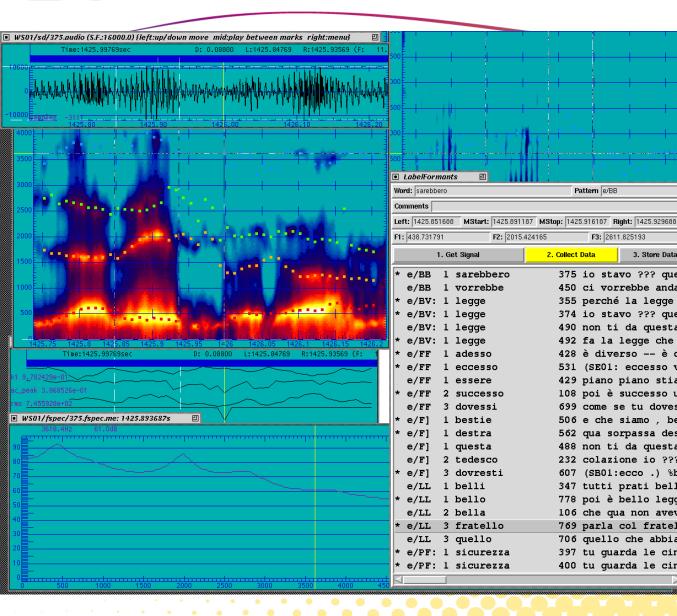
769 parla col fratel

706 quello che abbia

397 tu guarda le cin

400 tu guarda le cin

3. Store Data



Linguistic Data Consortium

Token Selection

Vowel Segmentation

Identification of central tendency of word stressed vowel

Hand checking of formant tracker values for F1 and F2



Impressionistic Coding

_Code_cie								
en Start: 704.2061 Token End	d: 704.5877 Cod	le: 1 Comments:						
Play	Segment	[c] Code Standard	Code Intermediate	[S1 Code Changed	Code Other	Code Uncertain	Code Bad Token	iu a
	-							
) DONE 807 AE		÷				-		. mi piace cucinare [c
) DONE 808 AE						-		. mi piace cucinare [c
) DONE 815 AE		- /		-		-		. mi piace cucinare [c
) DONE 816 AE				-				. mi piace cucinare [c
) DONE 822 AE		-						. mi piace cucinare [c
) DONE 854 AE						-	-	7a cucinato . (CCXX: s
) DONE 858 AE			-	-	-	-	-	7a cucinato . (CCXX: s
) DONE 867 AE			di nei momenti [mu	-	-			re [iar'].
) DONE 880 AE			io questo :	-		-	-	
C) DONE 889 AE	301	place <= {cc.	Tandus} T.uot-dod	non mi piac	e [piac.] (BIUI:	a me piace tant	lissimo l'not-dog	g .) (CCXX: X: davvero
. LongSound AB01		10000000	1000	10.0	20000	00000		
Edit Query View Select	t Spectrum Pitch	h Intensity Formant Pulses						
704.2	206100 0.381600	704.587700						
-1	TT T		1. J. J.		1 1	1 1 1 1 1 1		1 ITT 4
8000 Hz		Jung Marian		and the second secon				100 dB 500 Hz (76 67 db ,µE)
0 Hz 0.816100 3.390000 703.390000	0.381600			Visible part 8.1	6.972300 70000 seconds			188 Hz 75 Hz 711.560000 1988 4
					0.00000 accorde			711.500000 1500.44
in out sel				Total duration 270	0.000000 seconds			► 🗹 Grou



Annotations



U1	U2	U3						U6	U7
			U4:	una	donna	bella	U5		
					H1:	bella			
					S1:	Е			
						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 Vowe	l	
Speaker	Speaker	Actual Pron	Preceding Env		
Age	Situation		Following Env.		
Sex					
Ed Level					
Profession					
Region					
Location					



hitnum=1267 word=gabbia utterance=gabbia uttnum=376

pattern=a/BB

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 (<u>http://ncslaap.lib.ncsu.edu/tools/</u>)
 - 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 (<u>http://ncslaap.lib.ncsu.edu/tools/</u>)
 - Icf2txt.pl: Xtrans .lcf
 Text (for forced aligner)
 - Icf2TextGrid.pl: Xtrans .lcf Praat TextGrid
- Penn Phonetics Lab Forced Aligner (http://www.ling.upenn.edu/phonetics/p2fa/)
 Praat TextGrid
 Cieri, Strassel: Robust, Digital, Empirical, Reproducible Sociolinguistic Methodology, NWAV 39 November 4-6, 2010 San Antonio, Texas



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



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 platformindependent 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



Fine

Cieri, Strassel: Robust, Digital, Empirical, Reproducible Sociolinguistic Methodology, NWAV 39 November 4-6, 2010 San Antonio, Texas

49



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

