

Progress Report from the Linguistic Data Consortium:

**recent activities in resource creation and distribution and
the development of tools and standards**

Christopher Cieri, Mark Liberman
[{ccieri,myl}@ldc.upenn.edu](mailto:ccieri,myl@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

- ***The Linguistic Data Consortium supports language-related education, research and technology development by creating and sharing linguistic resources: data, tools and standards.***
- **Activities**
 - **Distribute Data**
 - **Collect: news text, broadcast, conversation, meetings, read/[prompted speech](#) ...**
 - **Annotate: transcription, time-alignment, word segmentation, annotation for morphology, POS, gloss, syntactic structure, discourse structure & disfluency, annotation of topic relevance, entities, relations & events, summarization, translation**
 - **Lexicons: pronouncing, morphological, gloss**
 - **Infrastructure: OLAC, Annotation Graphs/AGTK, SPH_**
 - **Tools: Transcriber, MultiTrans, TableTrans, Buckwalter Arabic Morphological Analyzer, [Champollion](#)**
 - **Standards and Best Practices: [TDT v1.4](#), [Entity v2.5](#), [Relation v3.6](#), [Simple MDE v6.2](#)**

- Organizations join per year
- receive ongoing rights data released that year and
- online access to some corpora (LDC Online) and
- access to copies of data from closed membership years
- Some data available to non-members by sale or free distribution.

- **Benefits:**
 - broad data distribution across research communities
 - funding agencies avoid distribution costs
 - users receive vast amount of data; avoid enormous development cost\$

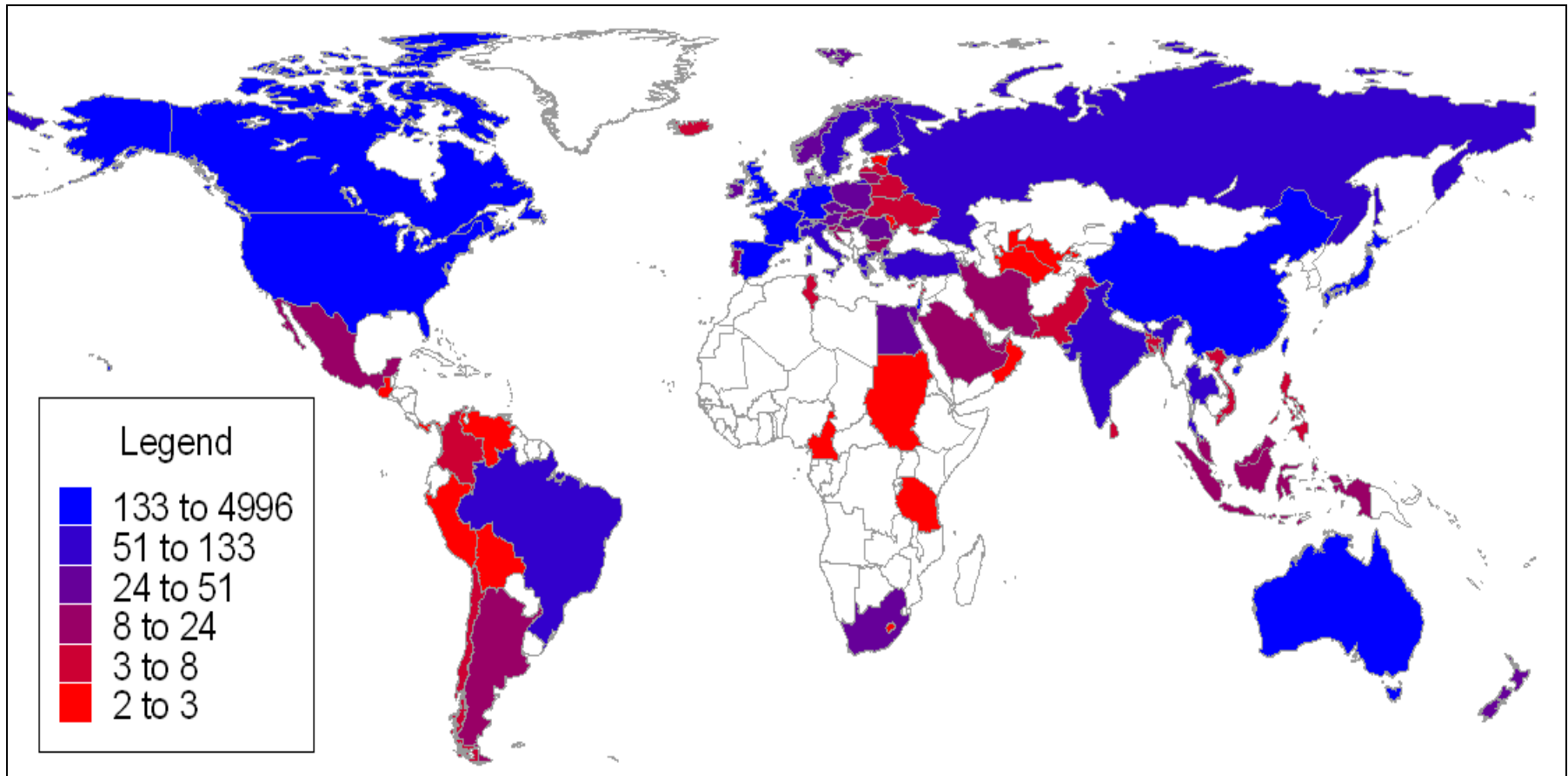
- Data comes from **donations**, funded projects at LDC or elsewhere, community initiatives, LDC initiatives
- Tools and specifications distributed without fee.

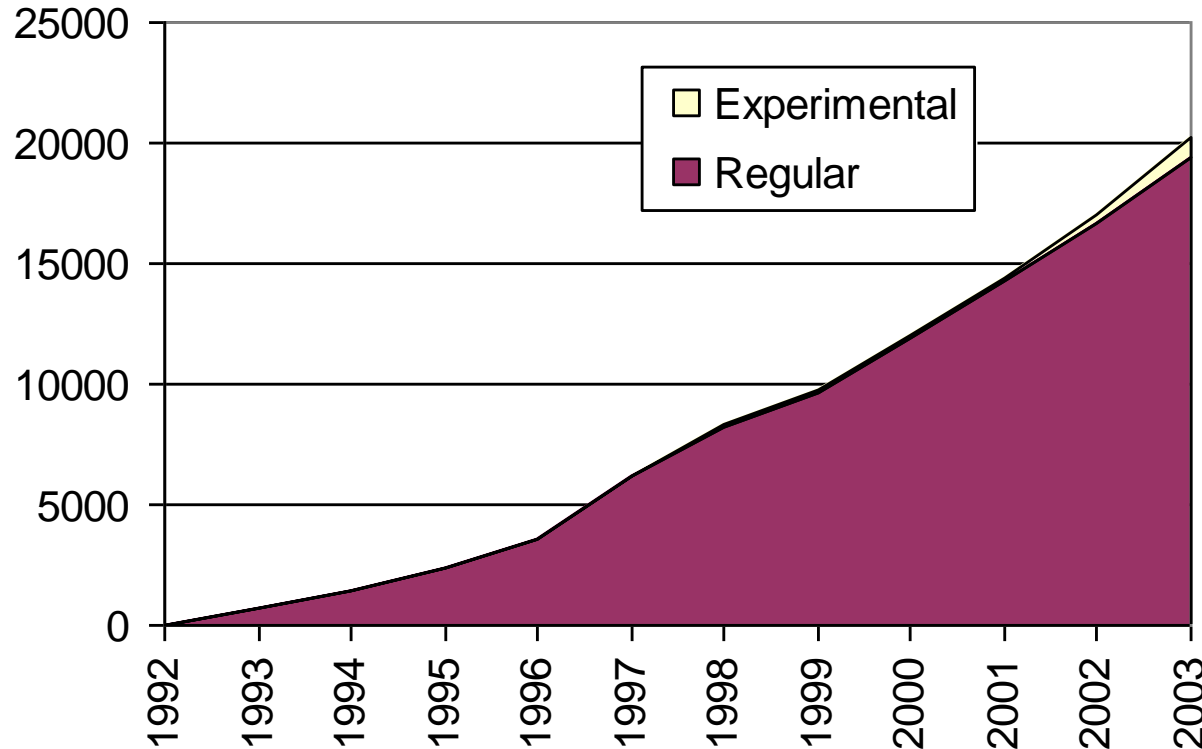
In operation 12 years

42 FTE staff of researchers, programmers, coordinators

288 Corpora + 2/month

>22,591 copies to 1720 organizations in 89 countries



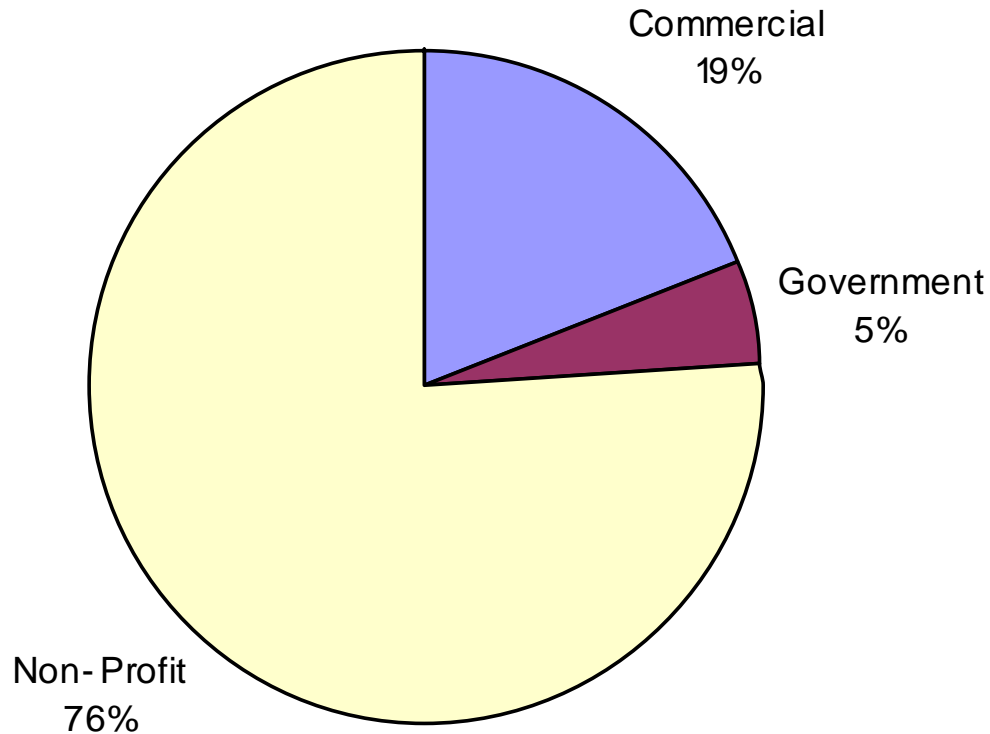


The core mission of any data center is to share data.

A central measure of effectiveness is the number and variety of organizations who benefit from data distribution.

“Experimental” corpora are collected and used initially for a specific purpose, a common task technology evaluation program or a commercial sponsor’s in-house R&D effort.

However, every corpus that LDC handles becomes generally available after its initial use.



Non-profits are still the biggest source of demand for LDC data.

Many government organizations outside the US use LDC data.

Commercial organizations may contract data creation through LDC provided that results are shared after a reasonable period of time.

A single distribution of a database to an organization may be shared throughout that organization.

- Language Modeling: **Gigaword** News text Corpora in Arabic, Chinese and English, AQUAINT Corpus of English News Text
- Tagging and Parsing: **Arabic Treebank** Parts 1 & 2, Korean-English Treebank, Morphologically Annotated Korean Text, Buckwalter Arabic Morphological Analyzer
- Machine Translation: updated Chinese-English Translation Lexicon and **Multiple-Translation** Corpora in Arabic and Chinese
- Speaker Recognition: Switchboard-2 PIII, 2001 NIST SRE
- ASR Prompted Speech: West Point Corpora in Arabic, Russian
- ASR Broadcast News: HUB4 English Speech and Transcripts
- ASR Meetings: **ICSI Meeting** Speech & Transcripts
- ASR Telephone: Voicemail Part II, HUB5 English, Egyptian Arabic, English, German, Mandarin, Spanish, CallHome style audio, transcripts and lexicon in Egyptian Arabic and Korean
- Dialog Systems: 2002 and 2001 **Communicator** Corpora
- Information Extraction, Summarization: MUC 6, ACE-2, TIDES Extraction (ACE) 2003 Multilingual, SummBank 1.0
- Gesture Recognition: FORM2 Kinematic Gesture
- Balanced Text: **American National Corpus**

- **Speech Recognition (LVCSR): CALLHOME**
 - 200 30 minute telephone calls among intimates
 - Japanese, Mandarin, English, Egyptian Arabic, German, Spanish
 - transcripts of 20 minutes of each call
 - pronouncing lexicon, POS, morphological analysis, frequency
- **Language Identification: CALLFRIEND**
 - 200 30-minute telephone conversations in 18 languages
- **Topic Detection and Tracking**
 - newswire and transcribed broadcast news with translations
 - story boundaries, topics and topic relevance judgments
 - Chinese, Arabic, English
- **Less Commonly Taught Languages**
 - survey of resource issues and resources in 320 languages
 - plain & parallel text, translation lexicons, topic relevance and entity tagging, POS taggers, encoding converters
 - Hindi, Bengali, Panjabi, Tamil, Tagalog, Cebuano, Tigrinya, Uzbek

- **EARS: Effective Affordable, Reusable Speech-to-Text**
 - Common task project to achieve 5 fold increase in ASR speech and accuracy and generate readable transcripts, adapted for downstream processing
 - LDC provides
 - BN: broadcast news, CTS: conversational telephone speech, meetings
 - Time aligned transcripts, MDE annotation
 - Training, development test and evaluation data
 - English, Mandarin and Arabic
 - Fisher: 16,454 ten-minute calls on 100 topics with gender, regional and age balance; 2742 hours of audio of which 2035 have been transcribed
- **TIDES: Translingual Information Detection, Extraction and Summarization**
 - News understanding system that, based on input language query performs retrieval and summarization of multilingual, multimodal news translated back into input language
 - LDC provides
 - newswire and broadcast news, captions, transcripts, ASR output
 - Annotation of topic relevance, entities, relations and events
 - Summaries, multiple translations and quality assessments
 - English, Mandarin and Arabic
 - Chinese and Arabic multiple translation corpora in which 4+ agencies translate the same input text at the sentence level; with human assessments of adequacy and fluency

EARS Data for 2004 - Netscape

EARS Data for 2004

Updated 5/21/04

Phase	Area	Language	Data Type	Source	Epoch	Amount	Annotation	Delivery	Notes
Evaluation	STT	English	BN	EARS 2003 collection	Dec 2003	180 min	careful transcription	9/1/2004	no IPR for some sources
		Chinese	BN	EARS 2003 collection	?	60 min	transcription	9/1/2004	no IPR for some sources
		Arabic	BN	EARS 2003 collection	Dec 2003	60 min	careful transcription	9/1/2004	no IPR for any sources
	STT	English	CTS	English Fisher collection	n/a	180 min	careful transcription	9/1/2004	
		Chinese	CTS	HKUST collection	n/a	60 min	transcription	9/1/2004	
		Arabic	CTS	Levantine Fisher collection	n/a	60 min	careful transcription	9/1/2004	must define selection criteria
	MDE	English	BN	same as STT	Dec 2003	180 min	MDE V6.2	9/1/2004	no IPR for some sources
		English	CTS	same as STT	n/a	180 min	MDE V6.2	9/1/2004	

LDC Projects: TIDES - Netscape	
	Arabic
Text	<ul style="list-style-type: none"> • Arabic Newswire Part 1 • Arabic Gigaword • Arabic Newswire Part 2: Ummah Corpus (100K), Date started: July 2002
Parallel Text	<ul style="list-style-type: none"> • UN Arabic English Parallel Text Created by Jinxi Xu et al at BBN. TIDES participants contact LDC for a pre-release copy (LDC catalog no.: LDC2002E15). • Ummah Arabic English Parallel News Text representing 3,039 news stories with Arabic-English parallel translation, 13K sentence pairs, 762K words two sides combined. contact LDC for a pre-release copy (LDC catalog no.: LDC2002E48).
Lexicons	<ul style="list-style-type: none"> • Buckwalter Arabic Morphological Analyzer Version 1.0
Morpho-Syntactic Tagged Text	<ul style="list-style-type: none"> • Arabic Treebank: Part 1 v2.0 POS & Treebank of text from AFP (LDC catalog no.: LDC2003T06) • Arabic Treebank: Part 2 v1.0 Part-of-speech of 144,199 tokens from the Ummah Arabic News Text (LDC catalog no.: LDC2003E17) • Arabic Treebank: Part 2 v1.1 Treebank of 168,297 tokens from the Ummah Arabic News Text (LDC catalog no.: LDC2003E24) • Arabic Treebank: Part 2 v2.0 Coming soon - POS & Treebank of text from the Ummah Arabic News Text (LDC catalog no.: LDC2003Txx)
Detection	<p>TDT 3 Arabic Text - TDT participants contact LDC (LDC catalog no.: LDC2002E32)</p> <p>TDT4 Multilanguage Text Version 1.1: contact LDC for a pre-release copy (LDC catalog no.: LDC2003E21)</p> <p>TDT4 Multilanguage Audio: contact LDC for a pre-release copy (LDC catalog no.: LDC2003E02)</p> <p>TDT4 Annotations: 2002 and 2003 topics completed</p> <p>TREC Cross-Language Topics - 2001 and 2002 topics completed</p>
Extraction	<p>Automatic Content Extraction (ACE) Corpora</p> <ul style="list-style-type: none"> • TIDES Extraction (ACE) 2003 Multilingual Training Corpus (LDC Catalog Number: LDC2004T09) <p>ACE/TIDES participants only (contact LDC):</p> <ul style="list-style-type: none"> • ACE 2004 Pilot Corpus V1.2 (LDC2004E03)
Summarization	<ul style="list-style-type: none"> • Multiple Translation Arabic Corpus Part 1 (2002 TIDES MT eval data) - 141 Arabic news articles, 10 sets of human translations, 2 sets of COTS outputs and human assessment of the two COTS outputs. TIDES MT participants, contact LDC for pre-release copy (LDC catalog no.: LDC2002E54) • Arabic News Translation Corpus Part 1 - 600 Arabic news articles, translated by 6 translation agencies (each

- **NSF funded project, CMU/Upenn/LDC develop new computational technologies to foster fundamental research in communication**
 - animal communication, child language, classroom discourse, conversation analysis, text and discourse, gesture, sociolinguistics
- **AGTK: Annotation Graph Toolkit**
 - builds upon Annotation Graphs (Bird, Liberman 2001), directed acyclic graphs where nodes are optionally anchored with offsets and arcs can be labeled with multi-field records; many linguistic annotations can be represented with AG
 - open-source implementation of the AG model plus software components for creating linguistic annotation tools (<http://agtk.sf.net>)
 - AG stored as XML-based or tabular, plug-ins exist for many file formats
- **New Data – more than 350 free copies distributed of these corpora:**
 - Korean Morphological Analyzer and Morphologically Annotated Text
 - SLx Corpus of Classic Sociolinguistic Interviews
 - Santa Barbara Corpus of Spoken American English Part 2
 - FORM Kinematic Gesture: video with gesture annotation
 - Grassfields Bantu Fieldwork (Dschang, Ngomba)

Arbitrary length audio files

AG-compliant XML

User defined tag set

Functions:

Listen to audio

Segment easily

Transcribe

Code

Output results in

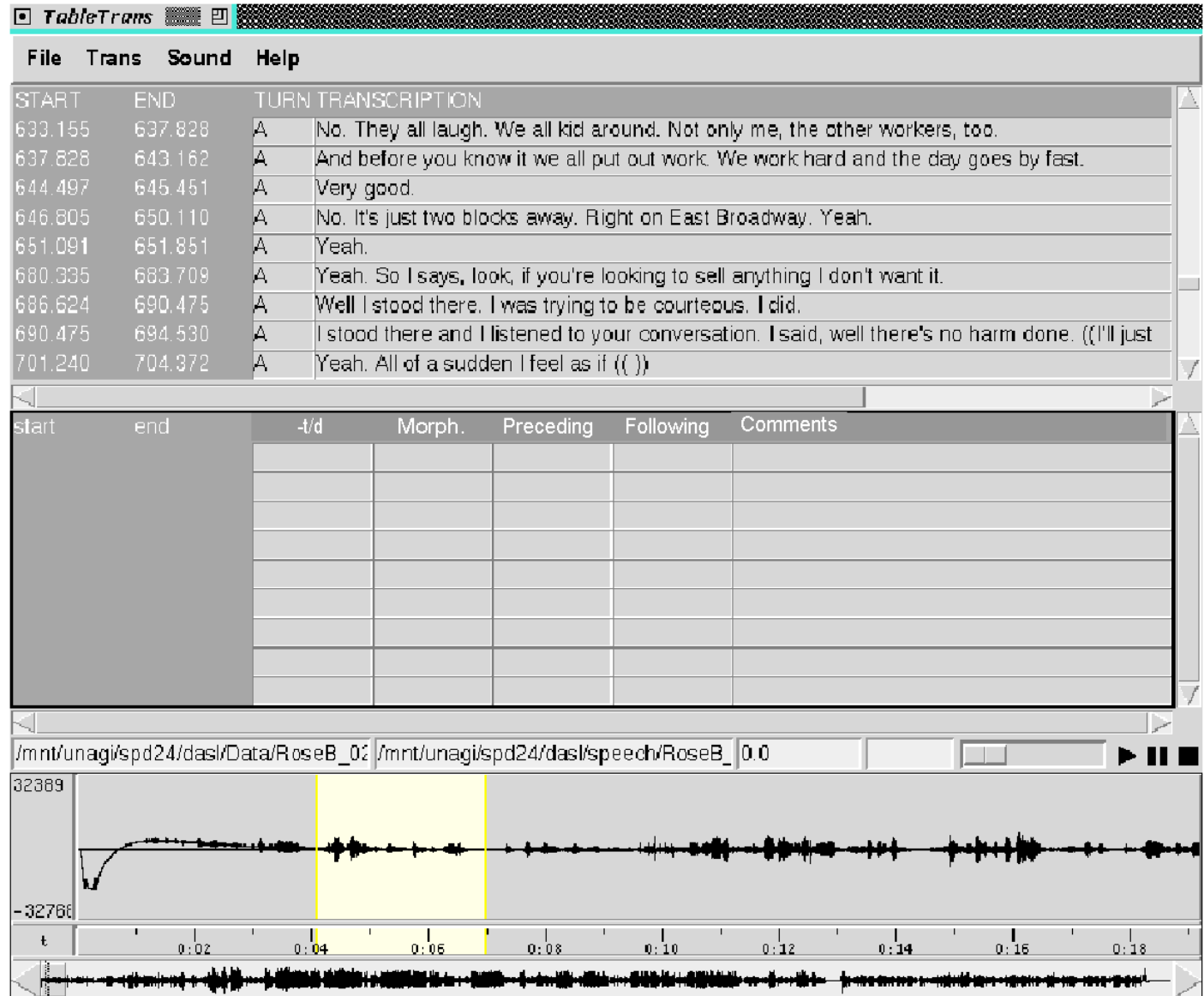
table format for

further analysis

Free and Extensible

via distributed source

code



The screenshot shows the TableTrans application window. The top part is a table with columns for START, END, TURN, and TRANSCRIPTION. Below the table is a detailed view of a transcription segment with columns for start, end, -t/d, Morph., Preceding, Following, and Comments. At the bottom, there is an audio waveform with a yellow highlight on a specific segment.

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

File path: /mnt/unagi/spd24/dasl/Data/RoseB_02 /mnt/unagi/spd24/dasl/speech/RoseB_0.0

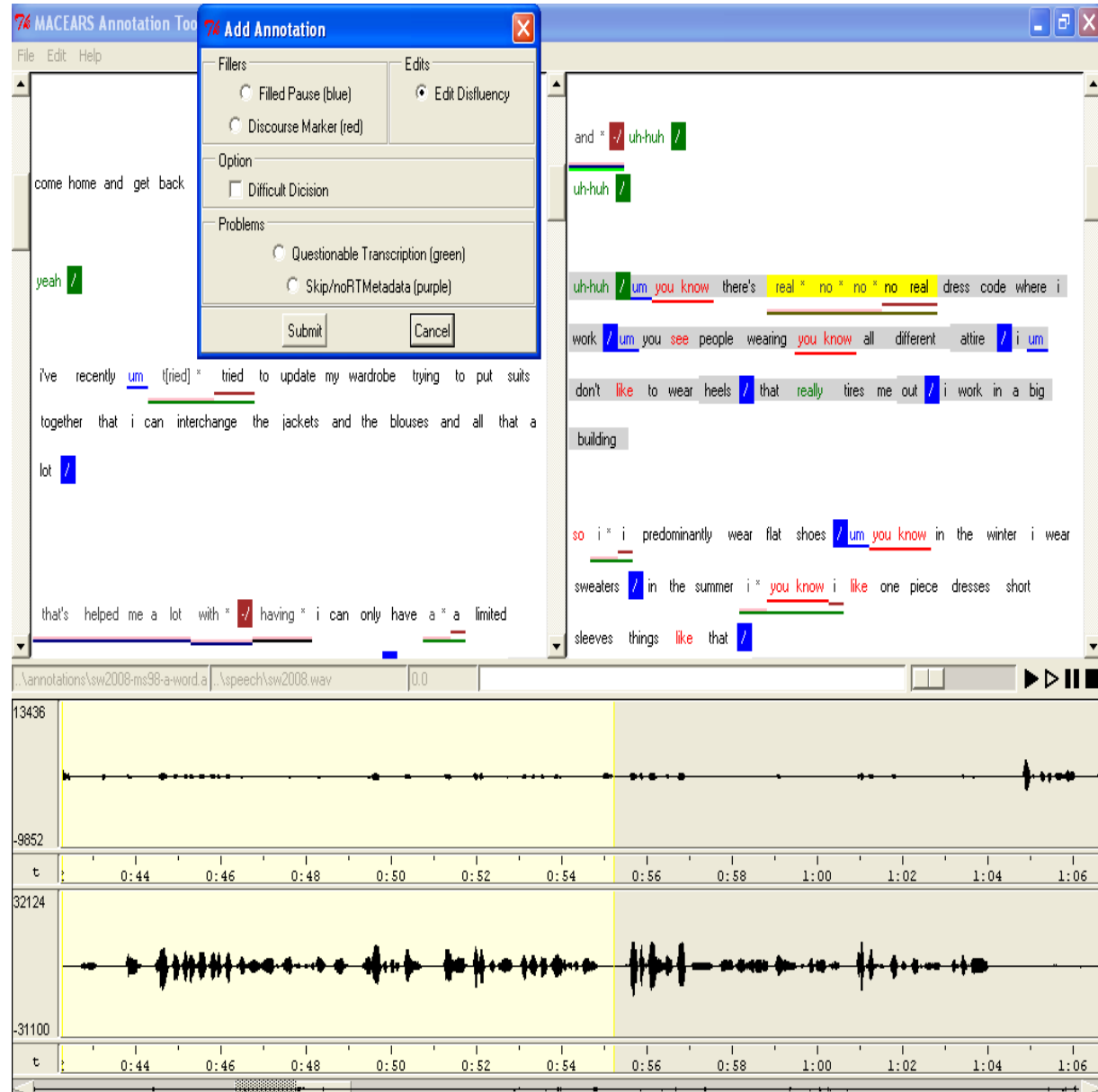
Waveform plot showing amplitude over time (t) from 0:00 to 0:18. A yellow vertical bar highlights a segment between approximately 0:04 and 0:06.

Conversational telephone speech and broadcast news data

Annotated for

- **Fillers:** filled pauses and discourse markers
 - » **Type:** repetition, revision, restart, complex
 - » **Structure:** original, interruption point, editing term, correction
- **SUs: semantic/syntactic units**
 - » **Sentence-level:** statement, question, backchannel, incomplete
 - » **Phrase-level**

English plus pilot studies in Chinese, Arabic



The screenshot displays the MACARS Annotation Tool interface. The main window shows a transcript of a speech sample with various annotations. A dialog box titled "Add Annotation" is open, allowing the user to select from different annotation types: Fillers (Filled Pause (blue), Discourse Marker (red)), Edits (Edit Disfluency), Option (Difficult Decision), and Problems (Questionable Transcription (green), Skip/noRTMetadata (purple)). The transcript text includes phrases like "come home and get back", "yeah", "I've recently um (tried) * tried to update my wardrobe trying to put suits together that i can interchange the jackets and the blouses and all that a lot", "that's helped me a lot with * having * i can only have a * a limited", and "uh-huh". The waveform below the transcript shows the audio signal with a yellow highlight indicating the current time position.

**News wire text
and transcribed
broadcast news**

Annotated for

Entities

PER, ORG, FAC

Relations

**ROLE.member-
of-group**

Events

**300K words each
of English,
Chinese, Arabic
for training data
in 2004**

ACETool: /mnt/talk/ACE/PHASE3/data/English/pilot-2003-fall/VOA20001121_0500.0060.sgm

File Mode Config Help

Text

This is VOA News. Tim Francis O'Leary, Florida's Supreme Court today continued to examine the issue of whether the results of ongoing manual vote recounts should be included in the state's final tally in the US presidential election. **Lawyers for Republican candidate George W Bush and his Democratic rival Al Gore**, presented their

Unattached Mentions

Mention - Head Assignment

Current Mention

Mention ID	Text	Entity ID	Type	Role	Metonymy
EM9	Lawyers for Republican candidate George W Bush and his Democratic rival Al Gore	E7	BAR	Difficult	yes

Entity Table

Entity ID	NAM	NOM	PRO	Other	Type	Subtype	Class	Gazet.
E5					Diff	Diff.	Diff	Yes
E6				US	GPE	Nation	ATR	Yes
E7			their	Lawyers for Republic	PER		SPC	Yes

- **A seminar series on issues in language data and database creation**
- **A selection of recent titles**
 - **Arabic Propbank, Mona Diab, Stanford University**
 - **The Contextualization of Linguistic Forms across Timescales, Stanton Wortham, Penn Graduate School of Education**
 - **Finite State Morphology using Xerox Software, Kenneth Beesley, XRCE**
 - **Interfaces for Parser and Dictionary Access, Malcolm D. Hyman, Harvard University**
 - **Mining the Bibliome: Information Extraction from Biomedical Text, Mark Liberman**
 - **The Pennsylvania Sumerian Dictionary Project, Stephen Tinney, Penn Museum**
 - **Project Santiago, Colonel Stephen LaRocca, Center for Technology Enhanced Language Learning**
 - **Searching the Prague Dependency Treebank, Jiri Mirovsky and Roman Ondruska, Charles University**
 - **Tongue-Tied in Singapore: A Language Policy for Tamil? Harold F. Schiffman, Penn Department of South Asia Studies**

- **LDC activities characterized by**
 - more, more, more (volume, languages, types of annotation)
 - better, faster, cheaper
- **LDC addressing needs by**
 - specific projects in data creation
 - actively publishing findings
 - **sharing tools and specifications**
 - networking where **fruitful**: OLAC, COCOSDA, ICWLR, ENABLER
 - open dialog in the LDC Institute
 - incorporating annotation, research and tool development: BITS, **EZ-Query**, AGTK, **QTr**, Champollion,
- **Data Centers need**
 - more intensive, bidirectional collaboration with researchers
 - more concrete collaboration amongst themselves
 - data **“donations”** from researchers
 - **most importantly continuing support from sponsors and researchers**