New Methods for Constructing Annotated Speech Corpora

Steven Bird

Linguistic Signals

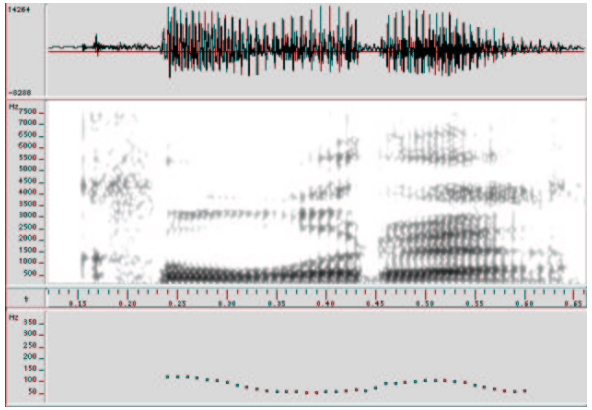
- **Broadcast news:** ***USC Marketplace***
  NIST CSR EVAL
- **Helicopter:** ***Out of Fuel***
  AFRL/DUKE
- **Telephone:** ***Callhome***
  LVCSR

Dialect variation:
1. New England
2. Northern
3. North Midland
4. South Midland
5. Southern
6. New York City
7. Western
8. Army Brat

TIMIT

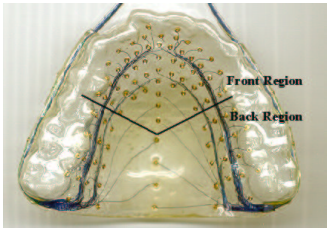
Time-series record of a linguistic "performance"

Audio Signals



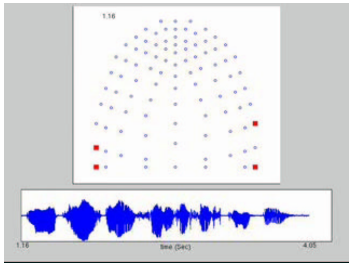
Physiological Signals

EPG: Electropalatograph



Artificial palate: 96 electrodes measure tongue contact

Video of EPG data



Other Kinds of Signal

- Video
  - e.g. studies of classroom interaction, gesture, sign
- Microphone arrays
  - e.g. for recording meetings
- Hydrophone arrays
  - e.g. studies of whale communication
- fMRI
  - e.g. studies of neural activity during linguistic performace
- Combinations

Language Science...

DATA INTENSIVE:

- >6 billion language speakers
- hundreds of utterances per day
- in ~6800+ languages
- with 10-100,000 word vocabularies

LINGUISTIC DATABASES:

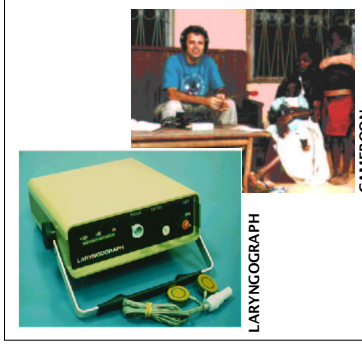
A digital repository of structured information intended to document natural language and natural communicative interaction

- bilingual dictionary
- collection of audio recordings with transcription and demographic data
- linguistic field notes

Data Collection: Lab & Field

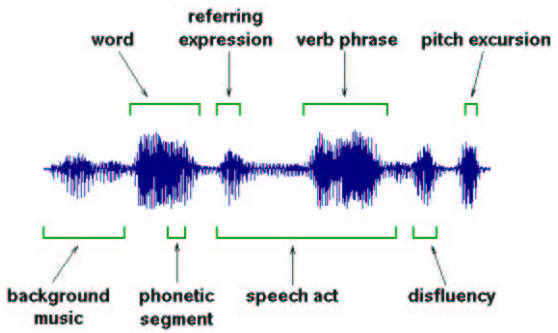
LABORATORY

FIELD



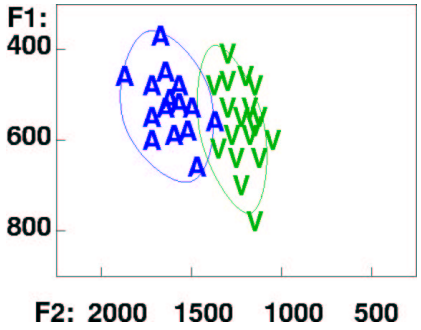
Linguistic Annotation

Associates structured symbolic information with a region of a linguistic signal



Example: Vowel Analysis

With a set of annotations we can analyze the corresponding regions of signal.



Comparing two vowels in the F1-F2 vowel space.

How do discrete linguistic categories relate to continuous acoustic parameters?

Example: Animal Communication

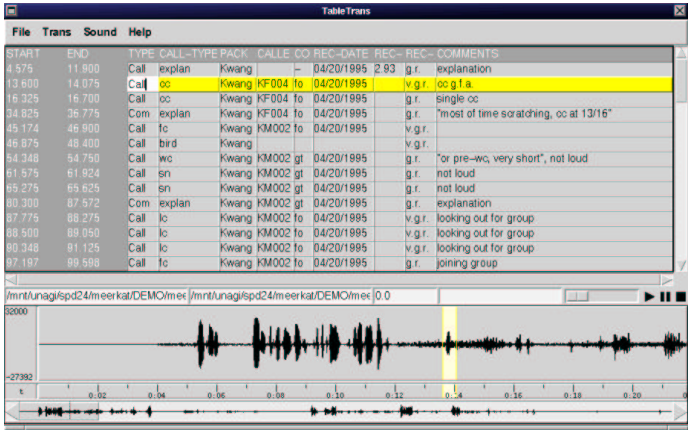
Meerkat recording by Marta Manser, South Africa

- **Field trip to South Africa in 2000**
- **80 hours of digital audio**
- **>20,000 annotations**

Research Questions / Methodology

- **How do animals use sound to communicate?**
- **What is the relationship between vocal communication and:**
  - **ecology?**
  - **social behavior?**
- **Methodology:**
  - **Record known individuals**
  - **Add detailed commentary on social events**
  - **Formulate hypotheses about how calls affect behavior**
  - **Test hypotheses using playback experiments**

Ethology Annotation Tool



Linguistic Databases in Language Technology R&D

- automatic speech recognition (ASR)
- machine translation
- text retrieval
- message understanding
- language teaching

"The evolution of ASR systems has been strictly related to the availability of large corpora of speech and the current systems achieve optimal performances only if proper databases are used."

- Becchetti & Ricotti 1999

Example of a Linguistic DB: TIMIT

Name: TIMIT = TI + MIT

- the first annotated speech database

Research questions and methodologies:

- *What acoustic properties of speech are invariant across speakers of different dialects?*
  - Build ASR systems and evaluate performance
- *How is the same phoneme realized differently in different contexts, by different speakers?*
  - Build parametric models of timing and co-articulation to account for the variation

Contents:

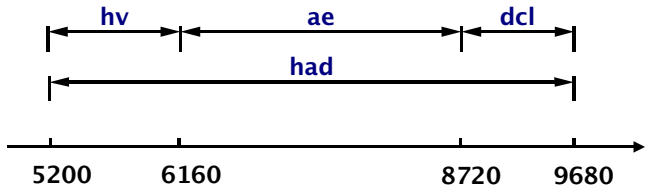
- 6300 phonetically transcribed recordings
- 630 speakers, 8 US dialects/regions

TIMIT Annotation

```
WORDS
2360 5200 she
5200 9680 had
9680 11077 your
11077 16626 dark
16626 22179 suit
22179 24400 in
24400 30161 greasy
30161 36150 wash
36720 41839 water
41839 44680 all
44680 49066 year
```

```
PHONEMES
0 2360 h#
2360 3720 sh
3720 5200 iy
5200 6160 hv
6160 8720 ae
8720 9680 dcl
9680 10173 y
10173 11077 axr
11077 12019 dcl
12019 12257 d
...
```


TIMIT: Structure




```
5200 9680 had
```

```
5200 6160 hv
6160 8720 ae
8720 9680 dcl
```


Phonetic Queries

- Find all instances of the phonetic segment "a"
- Find words whose phonetic transcription contains a "d" and ends with a "k"
- Find phonetic segments which immediately precede a vowel that overlaps a high tone

Another Example: Switchboard

Corpus of 2400 telephone conversations

Originally transcribed on three levels:

- conversation, speaker turn, word

Subsequently annotated for:

- syntactic structure
- breath groups and disfluencies
- speech acts
- phonetic segments

Features:

- proliferation of layers with different tokenizations

SWB: Example

```
B.22:   Yeah, / no one seems to be adopting it. /
  Metric system, [ no one's very, + {F uh, } no one
  wants ] it at all seems like. /
```

```
((S
  (NP-TPC Metric system) ,
  (S-TPC-1 (EDITED (RM [)
                   (S (NP-SBJ no one)
                      (VP 's (ADJP-PRD-UNF very))) ,
                   (IP +))
           (INTJ uh) ,
           (NP-SBJ no one)
           (VP wants (RS ]) (NP it) (ADVP at all)))
  (NP-SBJ *)
  (VP seems
           (SBAR like (S *T*-1))) . E_S))
```


Switchboard: Example

```
B 21.86 0.26 Metric
B 22.12 0.26 system,
B 22.38 0.18 no
B 22.56 0.06 one's
B 22.86 0.32 very,
B 23.88 0.14 uh,
B 24.02 0.16 no
B 24.18 0.32 one
B 24.52 0.28 wants
B 24.80 0.06 it
B 24.86 0.12 at
B 24.98 0.22 all
B 25.66 0.22 seems
B 25.88 0.22 like.
```

```
[ Metric/JJ system/NN ]
,/,
[ no/DT one/NN ]
's/BES
very/RB ,/,
[ uh/UH ] ,/,
[ no/DT one/NN ]
wants/VBZ
[ it/PRP ]
at/IN
[ all/DT ]
seems/VBZ
like/IN ./.
```


Learn more...

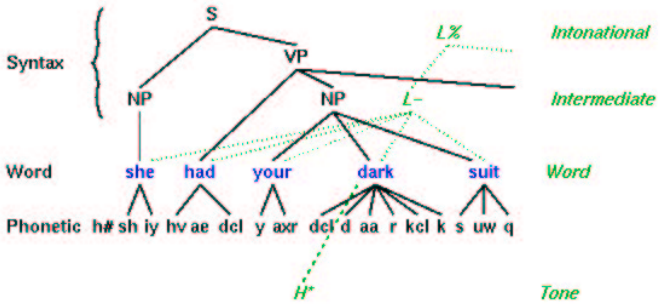
- Graff & Bird
  Many uses, many annotations for large speech corpora: Switchboard and TDT as case studies
  LREC 2000
  http://arxiv.org/abs/cs/0007024

SWB: queries

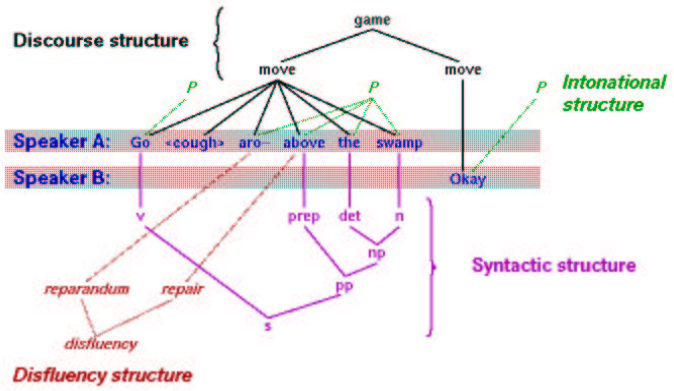
- To what extent do disfluencies and repairs respect syntactic structure?
- To what extent can prosodic phrasing be predicted by syntactic structure?
- ...

Aside: Intersecting Hierarchies

Syntactic and prosodic hierarchies intersect at the word level

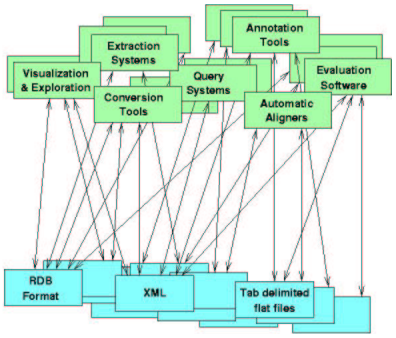


Aside: Intersecting Hierarchies



The tool problem

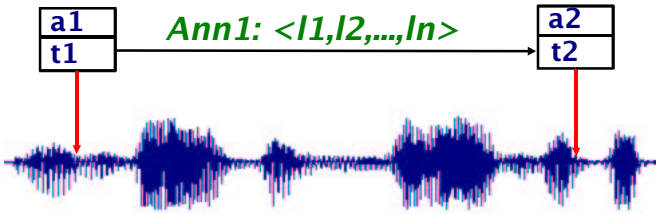
- formats, user interfaces, coding specs
- in-house tools
  - distribution?
  - facilitation?
- two-level model



Data Modeling Questions

- What is the model?
- Shopping list:
  - intervals and instants
  - sequential and parallel organization
  - hierarchy:
    - cross-cutting hierarchies
    - partial hierarchies

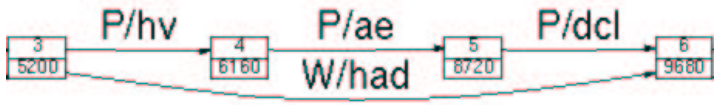
AG: Annotation



Relational representation in 3 tables:

- anchor, annotation (=arc), feature (=label)

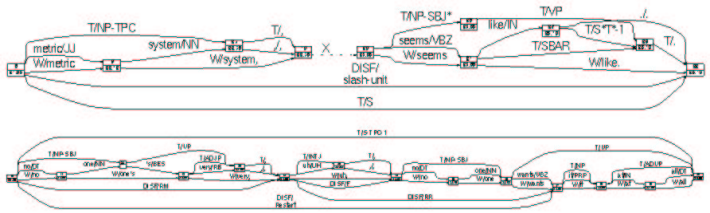
TIMIT: Annotation Graph



W = word level
5200 9680 had

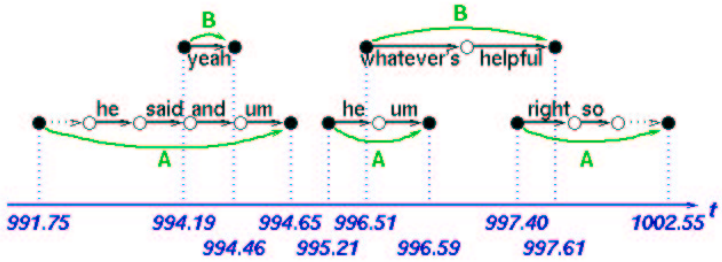
P = phoneme level
5200 6160 hv
6160 8720 ae
8720 9680 dcl

SWB: Annotation Graph



Disjoint Components

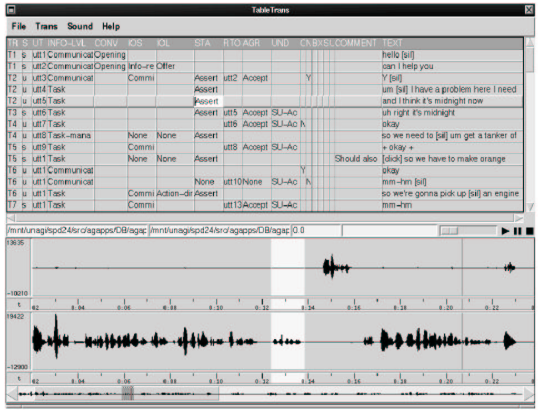
Logically independent events are graphwise disjoint; logically related events are connected.



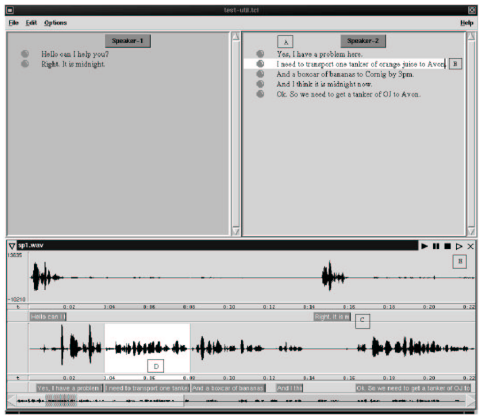
Learn more...

- **Bird & Liberman**
  A formal framework for linguistic annotation
  Speech Communication
  http://arxiv.org/abs/cs/0010033

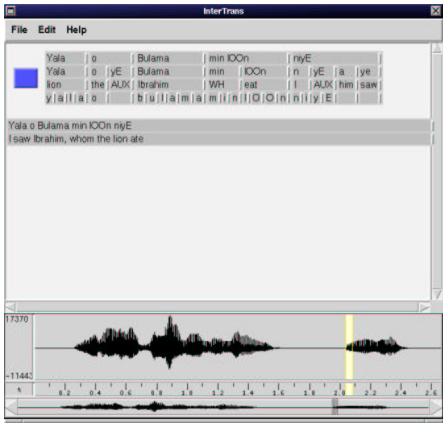
TableTrans



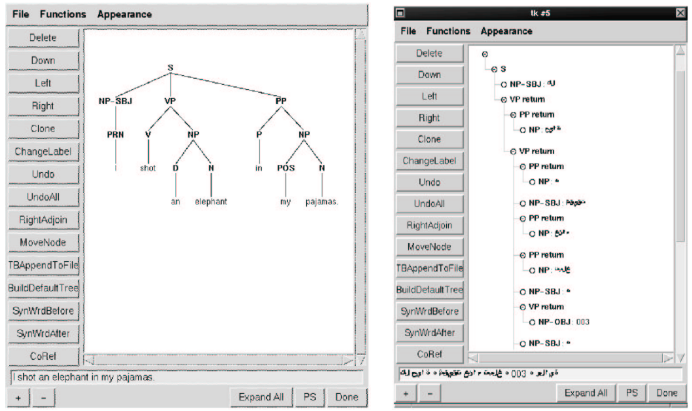
MultiTrans



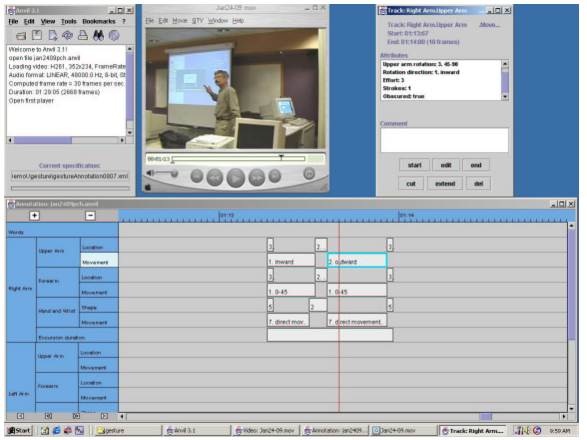
InterTrans



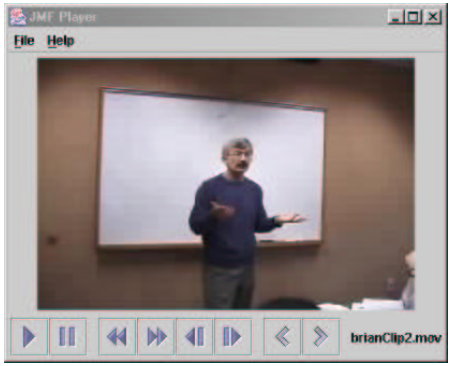
TreeTrans



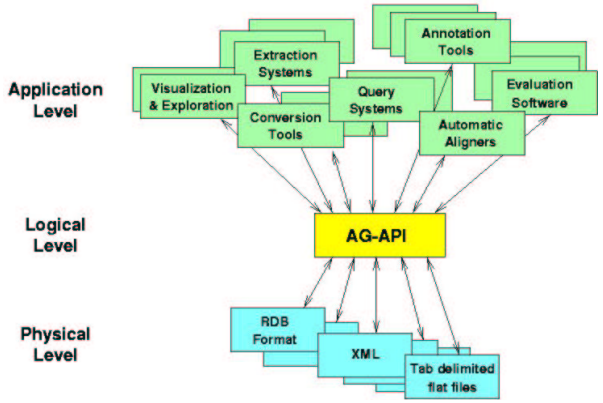
Anvil...



Video widget



Integration



Learn more...

- Bird, Maeda, Ma, Lee, Randall, and Zayat,
  TableTrans, MultiTrans, InterTrans and TreeTrans: Diverse Tools Built on the Annotation Graph Toolkit
  LREC 2002
  http://arxiv.org/abs/cs/0204006
- Cotton & Bird,
  An Integrated Framework for Treebanks and Multilayer Annotations
  LREC 2002
  http://arxiv.org/abs/cs/0204007
- Maeda & Bird,
  A formal framework for interlinear text
- agtk.sf.net

Research Questions

- Data models and APIs
  - new tasks, e.g. CA, gesture, treebanking, ...
  - what is the structure of the data?
  - what are the well-formed operations?
- Query languages
  - efficient storage, indexing
  - expressive and tractable languages
- Finite state processing
  - alternative to RDBMS, SQL
  - map AGs to FSMs and queries to FSTs
- Reconciling expressiveness with tractability...

Research: further reading...

- Data Models and APIs
  - Maeda & Bird - interlinear text
  - Cotton & Bird - treebank
- Query Languages
  - Bird, Buneman & Tan (LREC 2000)
  - Cassidy & Bird (ADC 2000)
  - Cieri & Bird (ACL 2001)
  - Ma, Lee, Bird & Maeda (LREC 2002)
- Finite State Processing
  - Bird (AAAS 2002)

OLAC: Open Language Archives Community

- resource discovery problem
- metadata
- Dublin Core
- Open Archives Initiative
- Demonstration

The future?

- adopting AG tools in-house
  - learning curve for developers
  - R&D not just D!
  - now is the time to switch...
- publishing corpora with tools
  - documentation
  - roles in an open source initiative
  - on-site training workshops

Future of Talkbank

- 5 year NSF project: 1999-2003
- www.talkbank.org
- phase 1: tools (years 1-3)
- phase 2: data (years 4-5)
  - domains: linguistic exploration, sociolinguistics, gesture, ethology
  - infrastructure: tools, metadata

A vision...

- camera-ready corpora?
  - evolution of print publishers
  - XML, Unicode
  - provide the tools others will use to give us data
- facilitating others
  - training
  - establishing best practices
- research
  - understanding, connections, head-start
  - resource allocation: invention vs re-invention