Annotation Graphs, Annotation Servers
and Multi-Modal Resources

Infrastructure for Interdisciplinary
Education, Research and Development

Christopher Cieri and Steven Bird
University of Pennsylvania
Linguistic Data Consortium
{ccieri|sb}@ldc.upenn.edu
Motivation

• Distribute Language Resources
  – 190 data sets including speech, text, lexicons in 2 dozen languages
  – >11,000 copies of datasets distributed to >1000 organizations worldwide
  – more than twice that amount in archive

• Create Data Resources
  – Collect raw broadcast news, conversations, news text
  – Transcribe, time-stamp, annotate for topic, named entity, co-reference

• Publicly Funded Research
  – Metadata for linguistic databases
  – Infrastructure (formalism, tools) to support annotation
  – Standards and best-practices

• Needs: larger volumes, more languages, more sophisticated annotation and more communities

• Solutions: efficient collection & annotation processes, tools & best practices, re-annotation and reuse to accommodate common needs
Common Needs

Switchboard

- created for speaker ID and topic spotting
- transcribed (TI, Penn, MSU), tagged for POS, syntactic structure and disfluency (Penn), transcribed phonetically (UCLA, Berkeley), annotated for discourse function (Colorado)
- used in ASR, NLP, discourse analysis

TDT-2

- collected for story segmentation, topic detection and tracking
- transcribed manually and automatically, segmented at story boundaries, translated, annotated for topic relevance, entity detection and cross-reference
- used in Topic Detection and Tracking, Spoken Document Retrieval, Automatic Content Extraction and JHU workshops in Novel Information Detection, Mandarin English Information and Audio Visual Speech Recognition

Same raw data and annotation used for very different purposes

There are even more exotic examples of common needs.
Field Linguistics

Tools support format consistency.

Resulting data reused for MT, CL.

Open source available at Source Forge

We welcome participation

ACL - Sharing Tools & Resources Workshop  Toulouse, July 2001
New set of interesting problems in markets that are quite large.
### Data and Annotations for Sociolinguistics

<table>
<thead>
<tr>
<th>Independent Variable File</th>
<th>Token File</th>
<th>Annotation File</th>
<th>Page</th>
<th>Tokens/Page</th>
<th>Total Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>/Shared/TDdeletion.tag</td>
<td>/Shared/TDdeletion.tok</td>
<td>/cciери/TDdeletion.ann</td>
<td>1/83</td>
<td>25</td>
<td>2059</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>t/d:</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>C Untouched C Deleted C Retained C Unsure C NA</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Morphological:</th>
<th>Monomorpheme C Irregular_Past C Regular_Past</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Preceding:</th>
<th>C Stop C Lateral C Rhotic C Alveolar_Nasal C Other_Nasal C Alveolar_Fricative C Other_Fricative</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Following:</th>
<th>C Obstruent C Lateral C Rhotic C Clustering_Glide C Other_Glide C Vowel C Pause</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>comments:</th>
<th>vocalized 1</th>
</tr>
</thead>
</table>

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

train/dr1/fjsp0/sa1.wrd:

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

train/dr1/fjsp0/sa1.phn:

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
AG as Relational Tables

Time: | Arc:  | Label:  
----- | ----- | ------ 
N | T | A | X | Y | T | A | L 
0 | 0 | 1 | 0 | 1 | P | 1 | h# 
1 | 2360 | 2 | 1 | 2 | P | 2 | sh 
2 | 3270 | 3 | 2 | 3 | P | 3 | iy 
3 | 5200 | 4 | 3 | 4 | P | 4 | hv 
4 | 6160 | 5 | 4 | 5 | P | 5 | ae 
5 | 8720 | 6 | 5 | 6 | P | 6 | dcl 
6 | 9680 | 7 | 6 | 7 | P | 7 | y 
7 | 10173 | 8 | 7 | 8 | P | 8 | axr 
8 | 11077 | 9 | 8 | 9 | P | 9 | dcl 
9 | 12019 | 10 | 9 | 10 | P | 10 | d 
10 | 12257 | 19 | 3 | 6 | W | 18 | she 
14 | 16626 | 20 | 6 | 8 | W | 19 | had 
17 | 22179 | 21 | 8 | 14 | W | 20 | your 
22 | 14 | 17 | W | 21 | dark 
     | 22 | suit
Query Language, XML

http://BASE-URL/cgi-bin/query?
X[].Y<timit/word;
X.[:hv].[].*:[:ae].[]*.Y<-timit/ph

<?xml version="1.0"?>
<!DOCTYPE AGSet SYSTEM "ag.dtd">
<AGSet id="Timit" version="1.0" xmlns="http://www.ldc.upenn.edu/atlas/ag/
<Timeline id="T1"> <Signal id="S1" mimeClass="audio" mimeType="wav" encoding="wav"
    unit="16kHz" xlink:href="TIMIT/train/dr1/fjsp0/sal.wav"/>
 </Timeline>
<AG id="t1" type="transcription" timeline="T1">
<Anchor id="A3" offset="5200" unit="16kHz"/>
<Anchor id="A6" offset="9680" unit="16kHz"/>
<Annotation id="Ann10" type="W" start="A3" end="A6">
<Feature name="label">had</Feature>
</Annotation>
</AG></AGSet>
Annotation Server
P-S Interactions

SMART – paradigmatic data identifies syntagmatic data of interest

CALLHOME – paradigmatic data provides reference pronunciations for words in syntagmatic data

Comlex Syntax – paradigmatic data provides reference information for words in syntagmatic data and contains pointers to other texts
274.35 279.50 A.119 He carves out different figures in the shrubs

different  dɪfərənt

1139; Male; 50; Northern; 2
Conclusions

Demands for more data in more languages with more sophisticated annotation, in more communities lead naturally to data re-annotation and reuse.

Research communities joined by their need of similar data can exist in symbiosis. However, research into best practices across research communities is required.

Annotation Graph Technologies offer an efficient approach to key problems in resource development and sharing.

See [www.ldc.upenn.edu](http://www.ldc.upenn.edu), [www.talkbank.org](http://www.talkbank.org) for latest