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

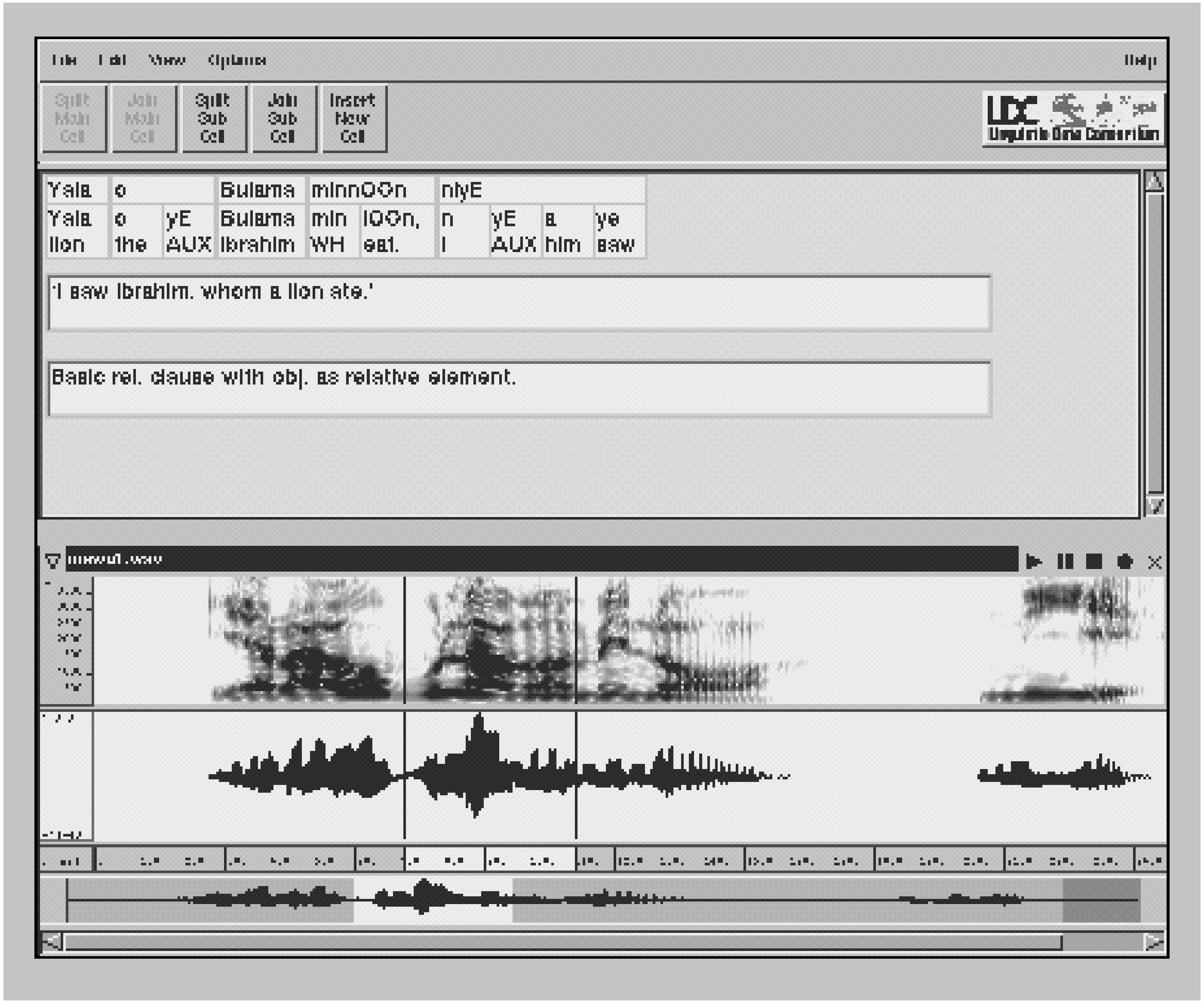
TDT-2

- 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



Yala	o		Bulama	minnOOn		niyE			
Yala	o	yE	Bulama	min	lOOn,	n	yE	a	ye
lion	the	AUX	Ibrahim	WH	eat.	I	AUX	him	saw

'I saw Ibrahim. whom a lion ate.'

Basic rel. clause with obj. as relative element.

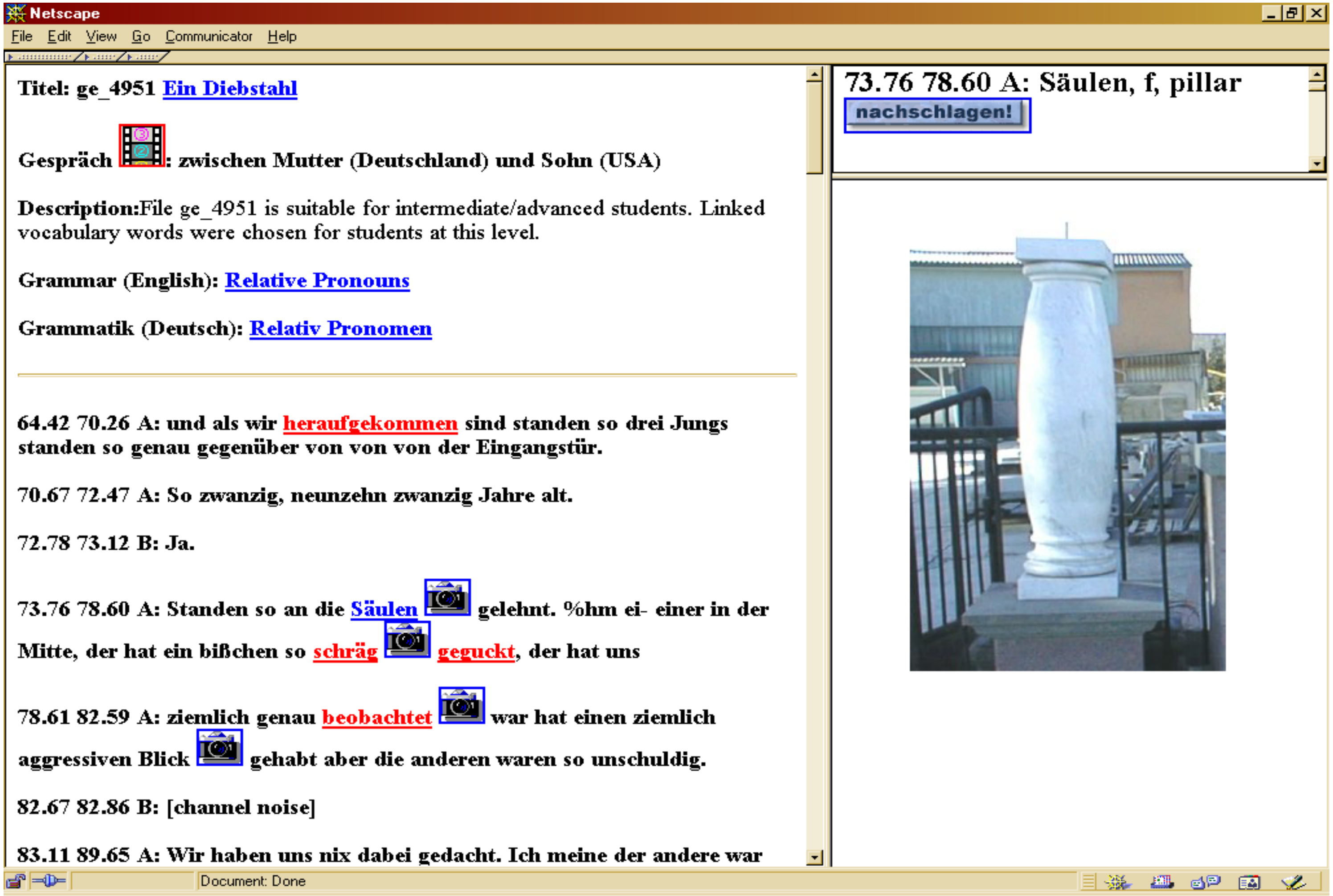
Tools support format consistency.

Resulting data reused for MT, CL.

Open source available at Source Forge

We welcome participation

Source Media Authoring



Titel: ge_4951 Ein Diebstahl

Gespräch : zwischen Mutter (Deutschland) und Sohn (USA)

Description:File ge_4951 is suitable for intermediate/advanced students. Linked vocabulary words were chosen for students at this level.

Grammar (English): Relative Pronouns

Grammatik (Deutsch): Relativ Pronomen

64.42 70.26 A: und als wir heraufgekommen sind standen so drei Jungs standen so genau gegenüber von von von der Eingangstür.

70.67 72.47 A: So zwanzig, neunzehn zwanzig Jahre alt.

72.78 73.12 B: Ja.

73.76 78.60 A: Standen so an die Säulen gelehnt. %hm ei- einer in der Mitte, der hat ein bißchen so schräg geguckt, der hat uns

78.61 82.59 A: ziemlich genau beobachtet war hat einen ziemlich aggressiven Blick gehabt aber die anderen waren so unschuldig.

82.67 82.86 B: [channel noise]

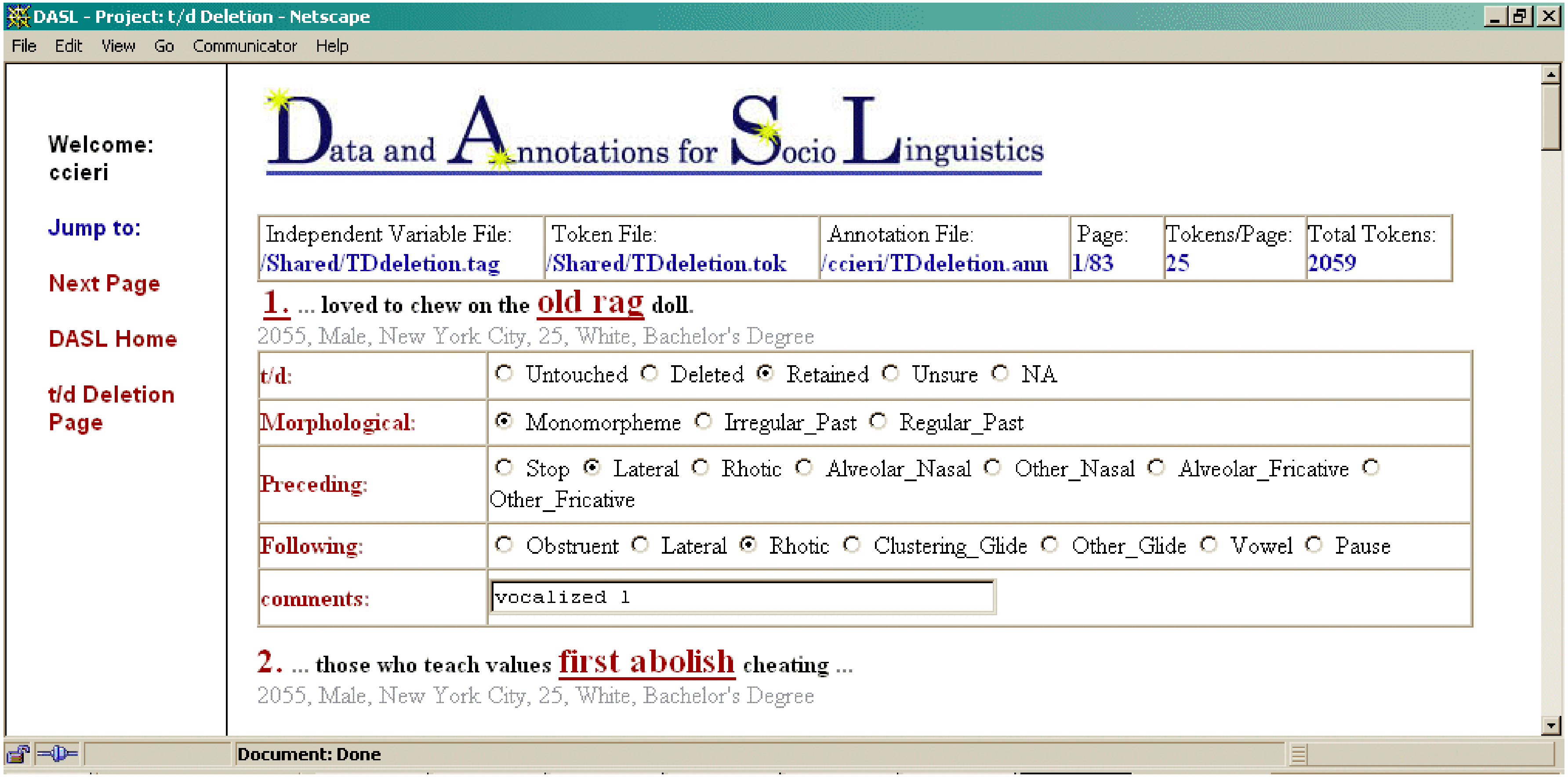
83.11 89.65 A: Wir haben uns nix dabei gedacht. Ich meine der andere war

73.76 78.60 A: Säulen, f, pillar

nachschlagen!

New set of interesting problems in markets that are quite large.

Sociolinguistics



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

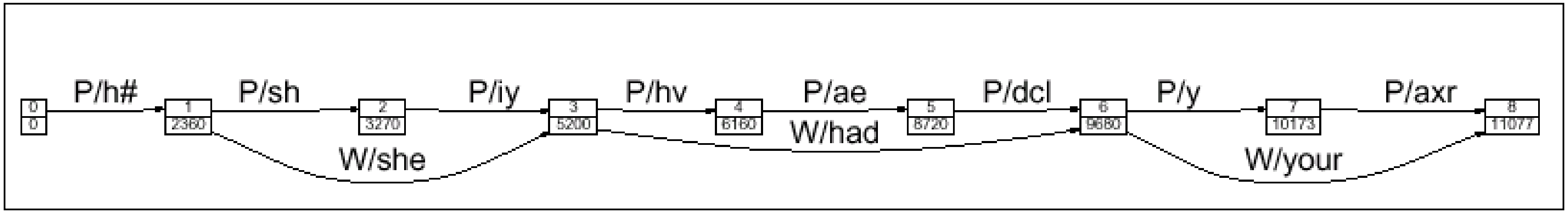
2. ... those who teach values **first abolish** cheating ...

2055, Male, New York City, 25, White, Bachelor's Degree

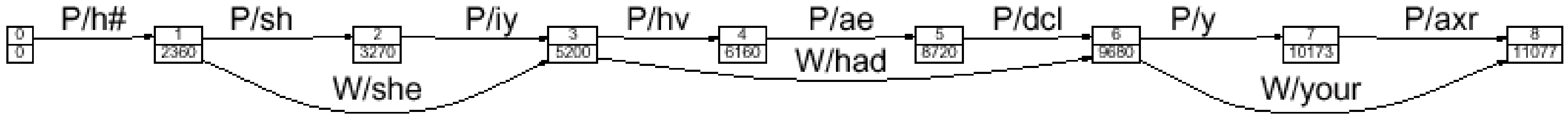
Annotation Graphs

```
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:

N	T
0	0
1	2360
2	3270
3	5200
4	6160
5	8720
6	9680
7	10173
8	11077
9	12019
10	12257
14	16626
17	22179

Arc:

A	X	Y	T
1	0	1	P
2	1	2	P
3	2	3	P
4	3	4	P
5	4	5	P
6	5	6	P
7	6	7	P
8	7	8	P
9	8	9	P
10	9	10	P
19	3	6	W
20	6	8	W
21	8	14	W
22	14	17	W

Label:

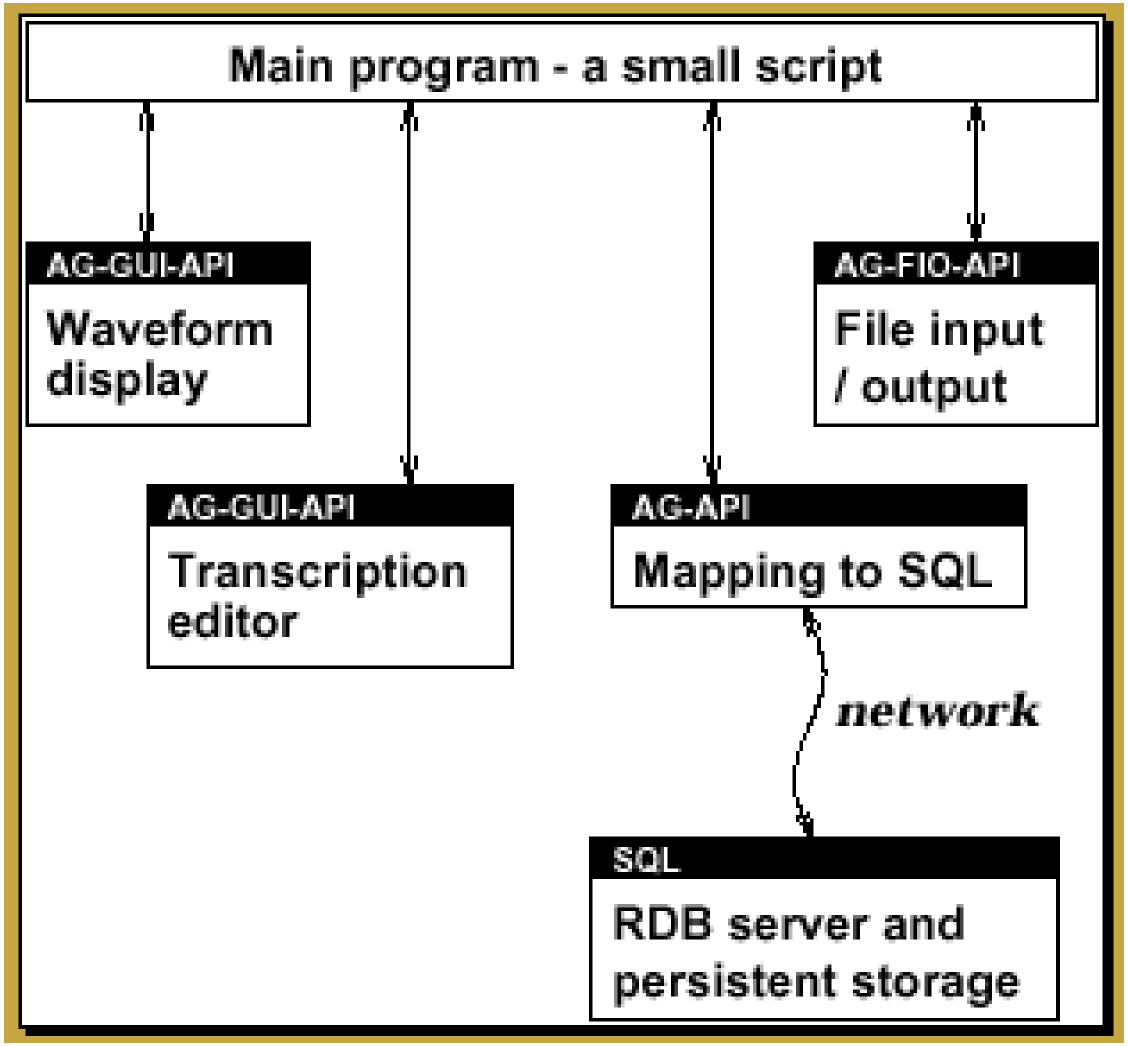
A	L
1	h#
2	sh
3	iy
4	hv
5	ae
6	dcl
7	y
8	axr
9	dcl
10	d
18	she
19	had
20	your
21	dark
22	suit

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/"
    xmlns:xlink="http://www.w3.org/1999/xlink" xmlns:dc="http://purl.org/DC/documents/rec-
    dces-19990702.htm">
<Timeline id="T1"> <Signal id="S1" mimeClass="audio" mimeType="wav" encoding="wav"
    unit="16kHz" xlink:href="TIMIT/train/dr1/fjsp0/sa1.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>
```

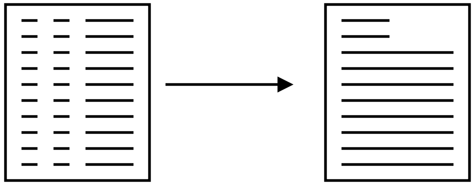
Annotation Server



P-S Interactions

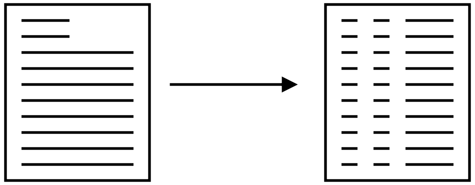
SMART – paradigmatic data identifies syntagmatic data of interest

PS



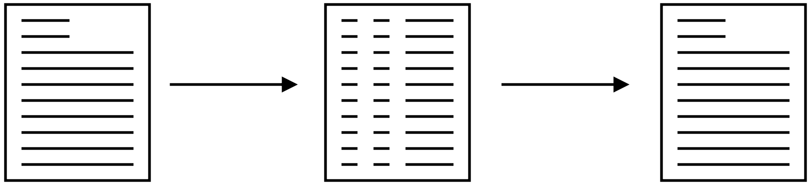
CALLHOME – paradigmatic data provides reference pronunciations for words in syntagmatic data

SP

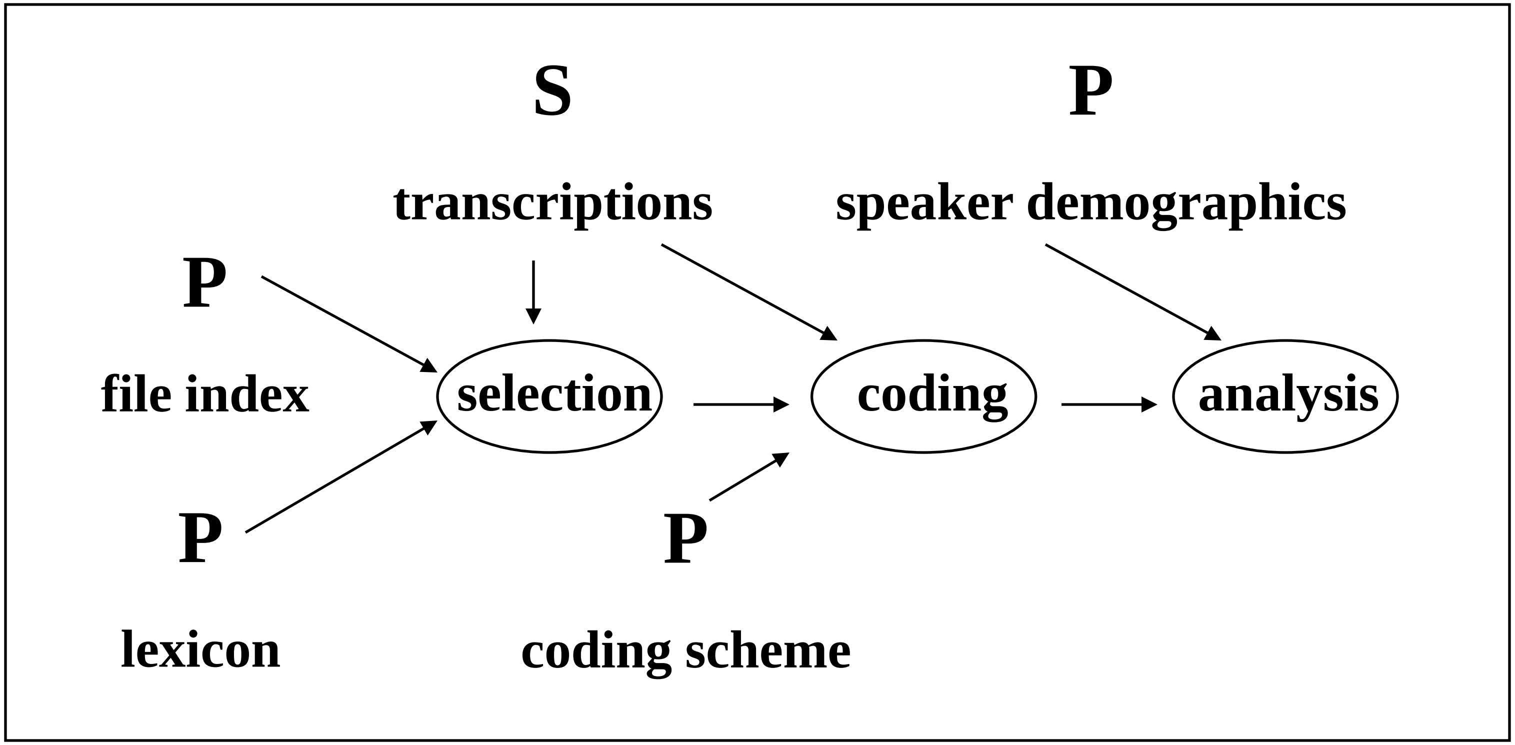


Comlex Syntax – paradigmatic data provides reference information for words in syntagmatic data and contains pointers to other texts

SPS



DASL Challenge



S

transcriptions

P

speaker demographics

P

file index

selection

coding

analysis

P

lexicon

P

coding scheme

`274.35 279.50 A.119 He carves out different figures in the shrubs`

`different   dɪfrɛnt`

`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, www.talkbank.org for latest