Data Centers: Sharing, Standards, and Linkage for the New Infrastructure

Brian MacWhinney

CMU - Psychology, Modern Languages, LTI,

http://talkbank.org

Coming of Age in Philadelphia

Whither Data Centers?

Advances in web data, data-sharing, and interoperability will lead to a new understanding of Data Centers.

Data will be distributed *in some uniform format*.

This requires accessibility and linkage methods.

There must be extensive sampling across languages, genres, situations, etc.

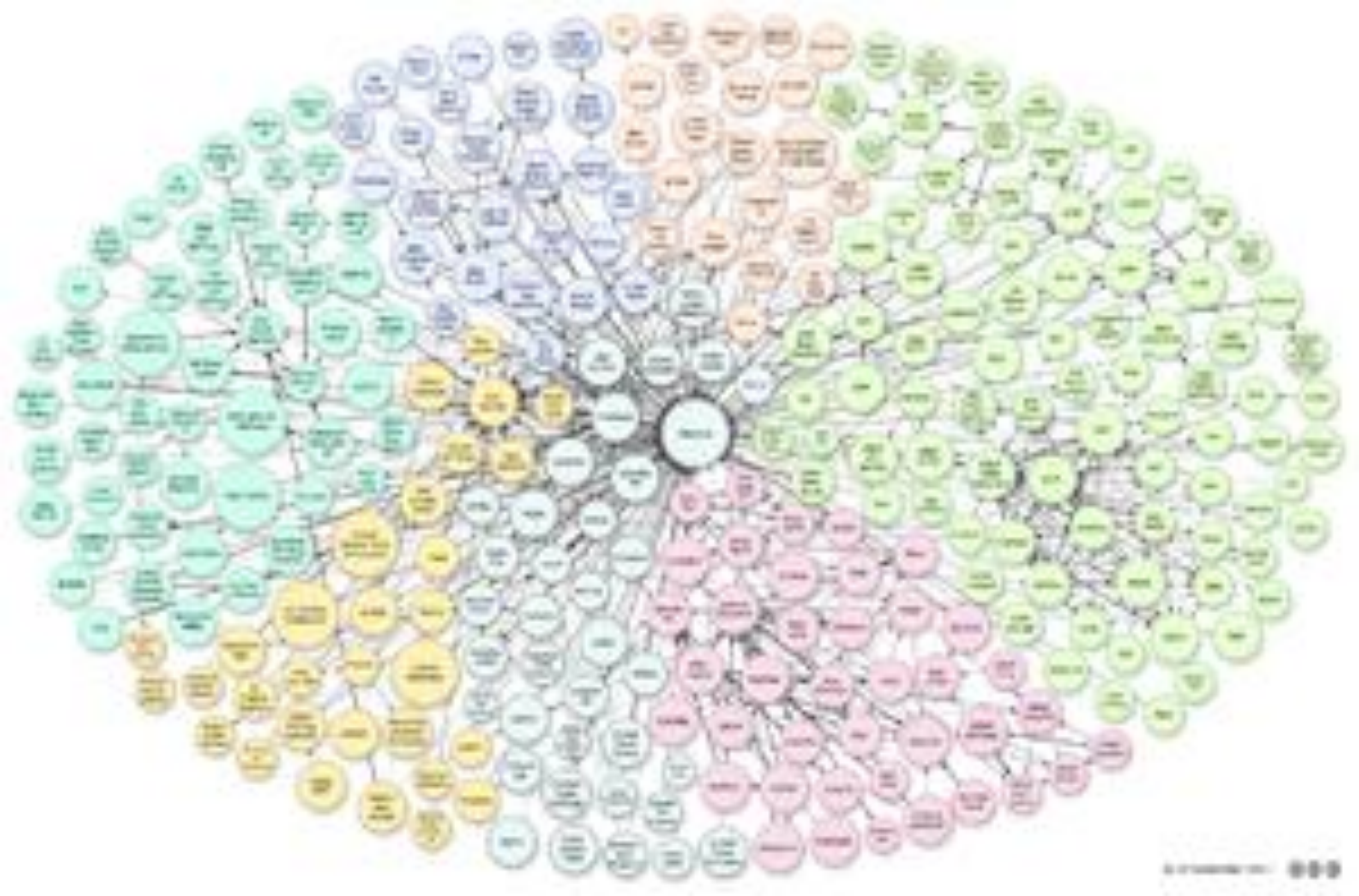
Linguistic Data Centers

- Corpora: spoken, written
- Diverse formats, standards
- Media: mostly audio, some video
- Tools: Lexicons, Taggers, Parsers
- Availability
  - declared through Metadata
  - licensing, purchase, memberships

The New Infrastructure

- The Web: Facebook, Twitter, YouTube, Blogs, Tutors, Support Groups
- Lexicon: WN, FN, VN, PropBank
- Ontologies: RDF, OWL
- Knowledge: Wikipedia, Google, Maps, extracted Knowledge Bases
- NLP: Parsers, Taggers, Segmenters
- Speech Technology, Informedia
- Multimedia Corpora, Google Maps
- Translation, Subtitles

Open Data Cloud



Toward a Shared Infrastructure

Users want to

- Locate appropriate data
- Relevant to the requirements of their group
- Retrieve or browse the data
- Specify analyses using a single interface
- Get automated results
- Download multimedia samples for further reflection and analysis
- Actually, user requirements are infinite

Users do not want

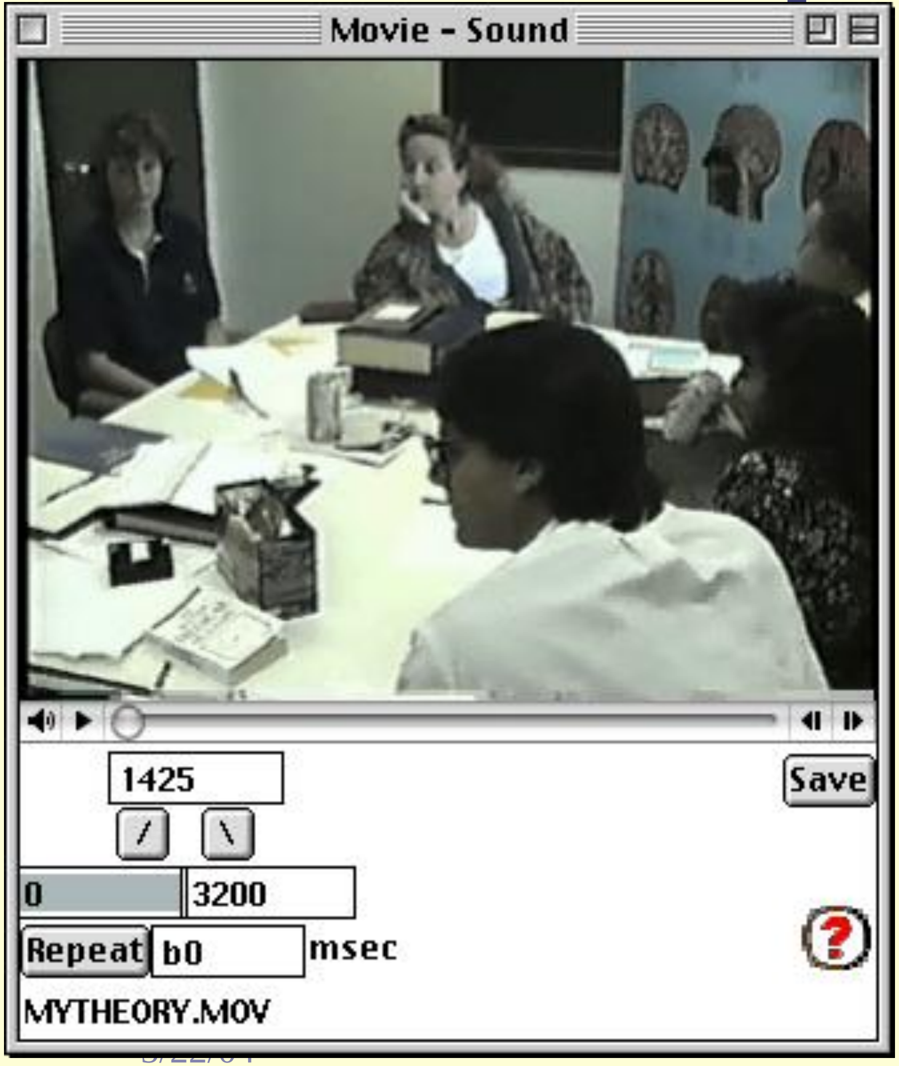
- Access restrictions
- Unclear documentation
- Poor transcription, poor media
- Gaps in data types
- Diverse transcription formats
- Diverse and complex analysis programs

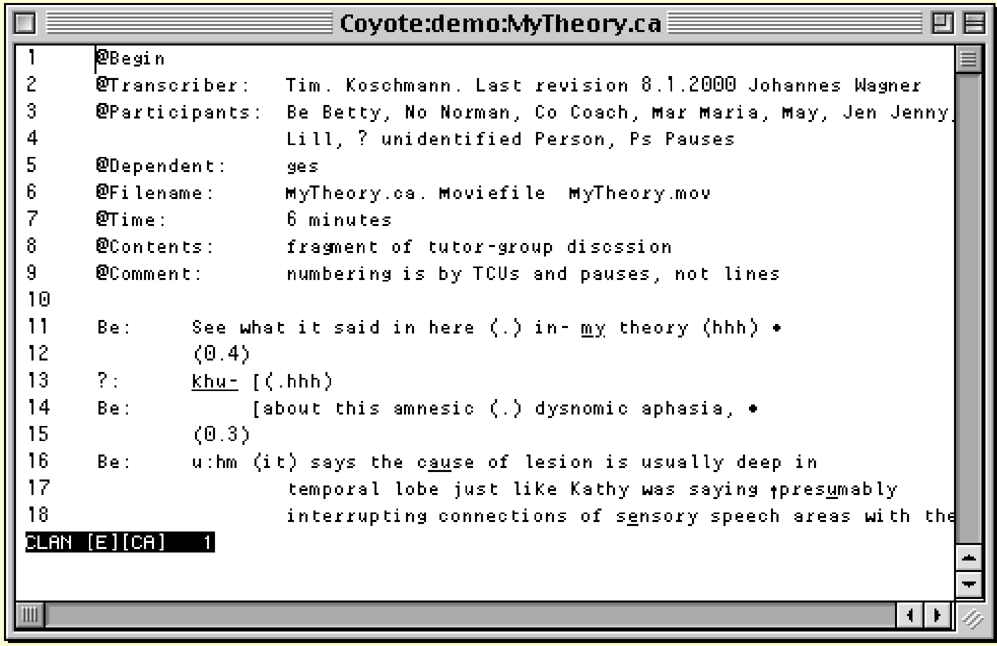
Example Challenges/Opportunities

- The TalkBank challenge
- The Speechome challenge
- The METAL challenge
- The SLA challenge

1. The TalkBank Challenge

Transcript linked to video





```
1       @Begin
2       @Transcriber:   Tim. Koschmann. Last revision 8.1.2000 Johannes Wagner
3       @Participants:  Be Betty, No Norman, Co Coach, Mar Maria, May, Jen Jenny,
4                       Lill, ? unidentified Person, Ps Pauses
5       @Dependent:     ges
6       @Filename:      MyTheory.ca. Moviefile  MyTheory.mov
7       @Time:          6 minutes
8       @Contents:      fragment of tutor-group discssion
9       @Comment:       numbering is by TCUs and pauses, not lines
10
11      Be:     See what it said in here (.) in- my theory (hhh) •
12              (0.4)
13      ?:      khu- [(.hhh)
14      Be:          [about this amnesic (.) dysnomic aphasia, •
15              (0.3)
16      Be:     u:hm (it) says the cause of lesion is usually deep in
17                      temporal lobe just like Kathy was saying ↑presumably
18                      interrupting connections of sensory speech areas with the
CLAN [E][CA]   1
```

CHILDES and TalkBank

	CHILDES	TalkBank
Age	26 years	10 years
Words	52 million	8 + 55 million
Media	2 TB	1.2 TB
Languages	33	18
Publications	3500+	180
Users	3200	600
Hits	1.5 million	.7 million

Database vs Data Center

- TalkBank data are in a single common format (CHAT)
- CHAT is adapted for use of research communities: CABank, PhonBank, BilingBank, AphasiaBank
- Programs are designed to operate on that format
- New data are reformatted to the standard
- This approach is important to make full use of the New Infrastructure

Principles

- Multilingual, multimodal
- Data-sharing
- Interoperability
- Mesh to CLARIN, LDC, Databrary, OLAC, IMDI, Open Data Cloud
- Consistent data format with checking
- Programs for data analysis

Gestural Views

Segment	N1
Action	rests chin on hand, elbow on table, right shoulder back
Gaze	front to Deedee
Classification	Attention
Meaning	Attention

*D: [så er det snart] [torturtid→]

%ges: |-------D1------| |----D2----|

|-------------N1-------------|

%com: assimilating the pronounciation of a danish actor in a then tv show

pic * pic *

Interoperability to

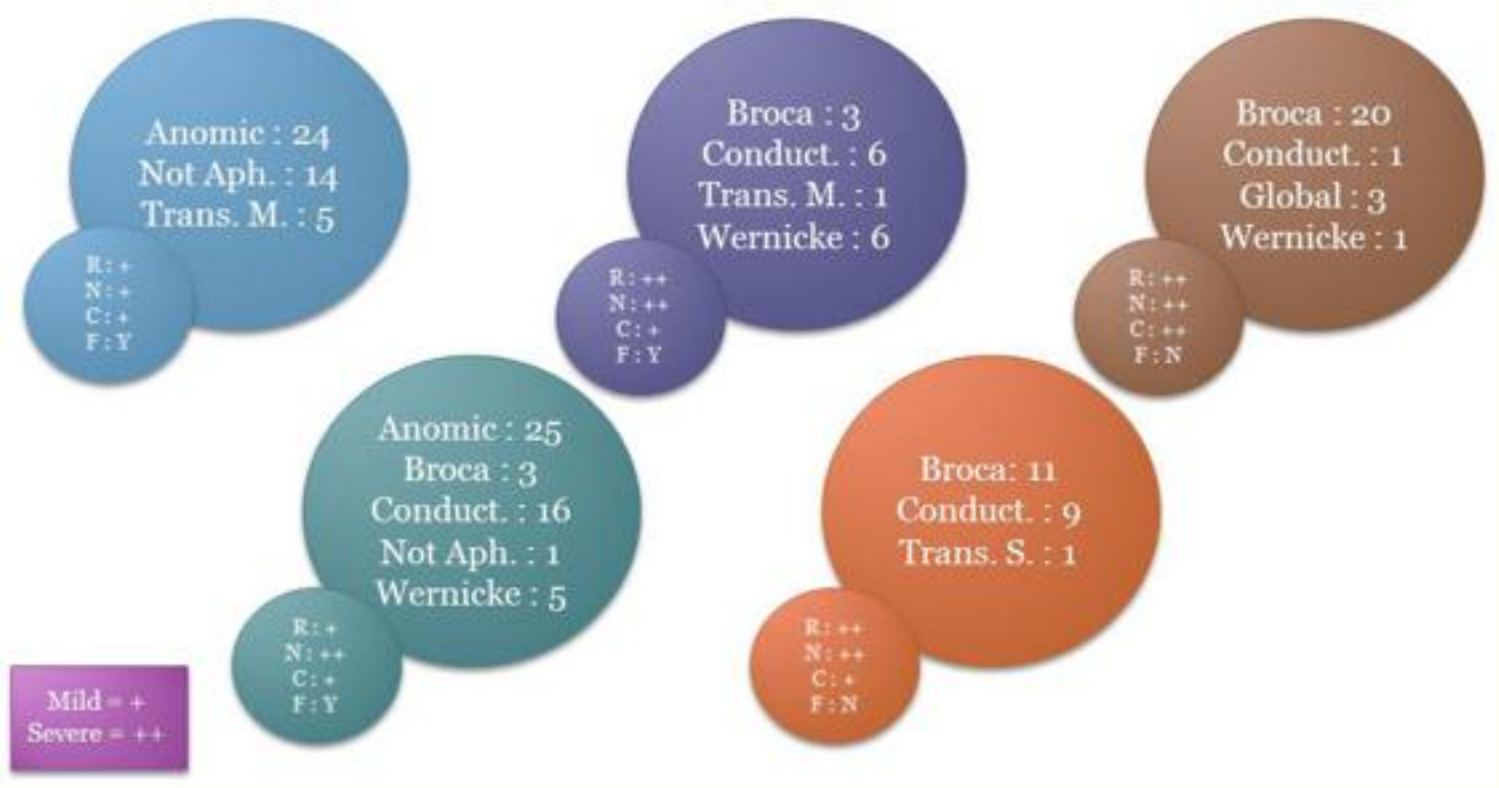
- ELAN
- Transcriber
- ANVIL
- CoNNL
- Praat
- Phon
- OpenSHAPA
- CHAT XML and RoundTrips

2. The Speechome Challenge

- AphasiaBank
- DementiaBank
- FluencyBank -- stuttering, SLA, etc.
- ACCBank - device models

- Standardized Protocol, Format
- Automatic analysis: errors, morphosyntax
- Measures from tests and corpora
- Clustering patient types
- Evaluating treatment outcomes

New Typology for Aphasia



Anomic : 24
Not Aph. : 14
Trans. M. : 5

R : +
N : +
C : +
F : Y

Broca : 3
Conduct. : 6
Trans. M. : 1
Wernicke : 6

R : ++
N : ++
C : +
F : Y

Broca : 20
Conduct. : 1
Global : 3
Wernicke : 1

R : ++
N : ++
C : ++
F : N

Anomic : 25
Broca : 3
Conduct. : 16
Not Aph. : 1
Wernicke : 5

R : +
N : ++
C : +
F : Y

Broca: 11
Conduct. : 9
Trans. S. : 1

R : ++
N : ++
C : +
F : N

Mild = +
Severe = ++

3. The Metaphor Challenge

- Automatic detection and interpretation of metaphors
- Conceptual Metaphor Theory
- English, Spanish, Farsi, Russian
- Relies on: WN, FN, PropBank, corpora, parsers, taggers, concept repositories, theory, experiment
- Everything requires harmonization

The Resources Reality

- Complete solutions (FreeLing, CLARIN) still have rough edges
- Harmonization is rapidly progressing, but mostly for lexical resources
- Corpus availability is good, but genre sampling is poor and incomplete
- The situation for parsing is unclear
- Automatic interpretation is *terra incognita*

4. The SLA Challenge

- Learner corpora: ICAME, ICE, TalkBank
- DuoLingo
  - Luis von Ahn (Captcha)
  - Freely available -- 300,000 users
  - “Crowdsourcing” of Spanish, English, German, French translations
  - Learner reading, translation, and dictation
- The iPad Language Partner
- Funding is a big issue

Where now?

- Challenges will drive innovation
- It would be great if funding could match the challenges
- LDC, TalkBank, CLARIN, ELRA and others will move to adapt to the new infrastructure