
TalkBank

Brian MacWhinney
CMU - Psychology

TalkBank

	CHILDES	TalkBank	Aphasia	PhonBank	HomeBank
Years	33 years	15	8	6	1
Words (mil)	59	47	1.5	0.7	2.0
Media	2.8 TB	1.1 TB	.4 TB	.6 TB	2TB
Languages	34	18	6	13	3
Publications	7500	220	140	76	8
Active users	2500	634	520	154	42
Web Hits	4,577,61	1,329,92	434,170	92,110	278,347

Focus Areas

Children	CHILDES	PhonBank	Narrative	Bilingual
Clinical	Aphasia	TBI	Dementia	Fluency
Adult	CABank	Tutoring	Medical	ClassBank
Multilingualism	BilingBank	SLABank	Online Tutors	CapVid

TalkBank Principles

- ❖ Standard format — CHAT, variable levels of detail
- ❖ Transcripts linked to media
- ❖ Multilingual (43 languages), multiscrypt
- ❖ Open and free access
- ❖ Analytic programs — CLAN, tutorials, MOR grammars
- ❖ TalkBank is a CLARIN-B Center, Core Trust Seal
- ❖ Metadata: OLAC, CMDI, VLO
- ❖ Interoperable with other resources: R, Elan, Praat, SALT

A Tour of the Websites

- ❖ All reachable from <https://talkbank.org>
- ❖ Also <https://talkbank.org/screenscasts>

4 Major Methods

1. Corpus Analysis - CLAN, R, ShinyServer, etc
2. Profiling - EVAL, KIDEVAL, FluCalc
3. Microanalysis - CA, gesture
4. Web-based Tutors, Experiments - eCALL

Method #1: Corpus Analysis

- ❖ **FREQ** - Frequency analysis
 - ❖ wild cards
 - ❖ word files (morality words, LIWC, medical)
- ❖ **KWAL** - Key word and line
 - ❖ matches highlighted
- ❖ **COMBO** - Regular expression matching
- ❖ Hits can be triple-clicked to go back to transcript and play

MOR, POST, GRASP

- ❖ 41 languages, but only 11 have MOR/POST
- ❖ Cantonese, Danish, Dutch, English, French, Italian, Hebrew, Japanese, German, Mandarin, Spanish
- ❖ GRASP for English, German, Hebrew, Spanish, Mandarin

MOR

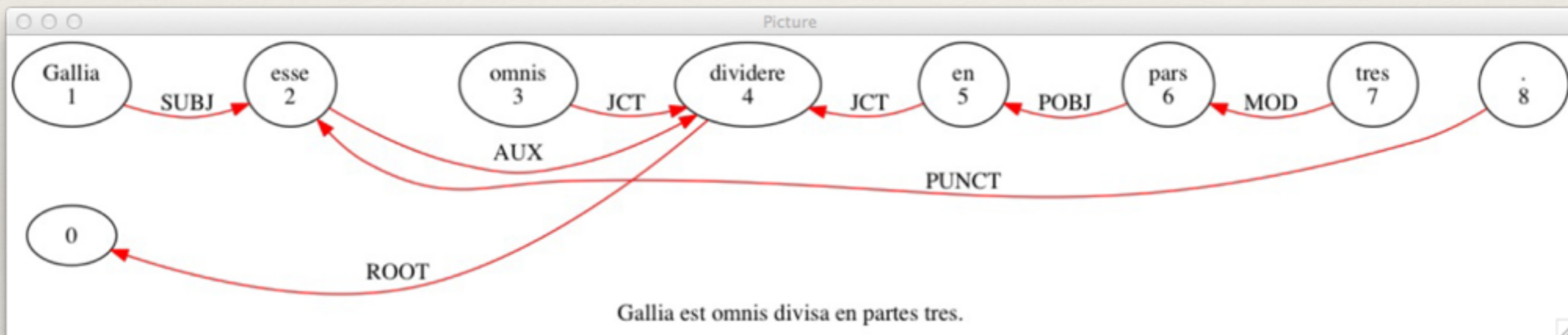
- ❖ More declarative than FST
- ❖ Part-of-speech tuned to spoken language
- ❖ Easy to use once there is a grammar
- ❖ Hard to build the grammar (A-rules, C-rules)
- ❖ 98% accuracy for English
- ❖ POSTMORTEM rules (as for German declension)

Bilingual MOR

- ❖ *CHL: +" [- spa] <yo no la> [/] yo no la desmentí porque. [+ break]
- ❖ *CHL: what's my word against hers &ladadada .
- ❖ *CHL: +" [- spa] todos estamos con un calor and@s working@s .
- ❖ All words are tagged implicitly; can be made explicit.
- ❖ Coding system makes code-switching junctures evident.
- ❖ Run English MOR, excluding [- spa], then Spanish MOR including [- spa]

Dependency Graphs

Web service runs by triple-clicking on %gra line



Using TalkBank data

- ❖ Standard statistical tests in Excel and R
- ❖ R routines - LuCiD Shiny server, childesr, rbrul
- ❖ Collostructional analysis
- ❖ CHILDES corpora inside SketchEngine

Method #2: Profiling

- ❖ EVAL and KIDEVAL
- ❖ Depends on MOR and GRASP
- ❖ Crucial for Clinicians

EVAL

MLU, TTR

Verbs / Utt

% errors

% N, V, Aux, Adv, Conj,

Pro

% PAST, PASTP, PL

Retracing, repetition

Select eval options

PLEASE SELECT AT LEAST ONE SPEAKER
Speaker: *PAR *INV *CLI

Database types:

Anomic Broca TransSensory
 Global Wernicke TransMotor
 Control Conduction NotAphasicByWAB

Age range: Male only Female only

Gem choices:

Speech Cinderella Important_Event
 Cat Umbrella Stroke
 Flood Sandwich Window

Sample Output

Comparing adler01a to 91 Broca PWA on all parts of protocol

The screenshot shows an Excel spreadsheet titled 'ACWT01a.eval'. The spreadsheet compares a subject (adler01a) to a database of 91 Broca PWA. The columns represent various linguistic metrics, and the rows represent the subject's performance and the database statistics.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	
1	File/DB	Language	Corpus	Code	Age	Sex	Group	Total Utts	MLU Utts	MLU Words	MLU Morph	FREQ types	FREQ tokens	FREQ TTR	Words/Min	Verbs/Utt	% Word Err	Utt Errors	density	% Nouns	% Plurals	% V
2	ACWT01a.ch	eng	ACWT	PAR	69;11.	female	Broca	67	57	3.702	4.105	100	256	0.391	38.886	0.358	11.719	34	0.359	23.438	2.344	
3								-0.748	-0.865	0.586	0.568	-0.055	-0.391	0.597	0.024	-0.134	0.228	-0.502	0.065	0.518	-0.041	
4																						
5	Mean Database							112.747	109.22	2.967	3.301	102.912	368.066	0.327	38.34	0.399	9.255	49.934	0.352	18.415	2.499	
6	Min Database							17	17	1.05	1.091	6	26	0.113	5.697	0.009	0.536	1	0.047	2.903	0.781	
7	Max Database							306	298	6.842	7.825	240	1617	0.692	131.834	1.456	72.566	158	0.53	50	26.923	
8	SD Database							61.18	60.403	1.255	1.416	52.846	286.605	0.107	22.749	0.307	10.822	31.761	0.109	9.693	3.796	
9	+/- SD * = 1 SD, ** = 2 SD																					
10	Database keywords: Broca																					
11	# files in database: 91																					

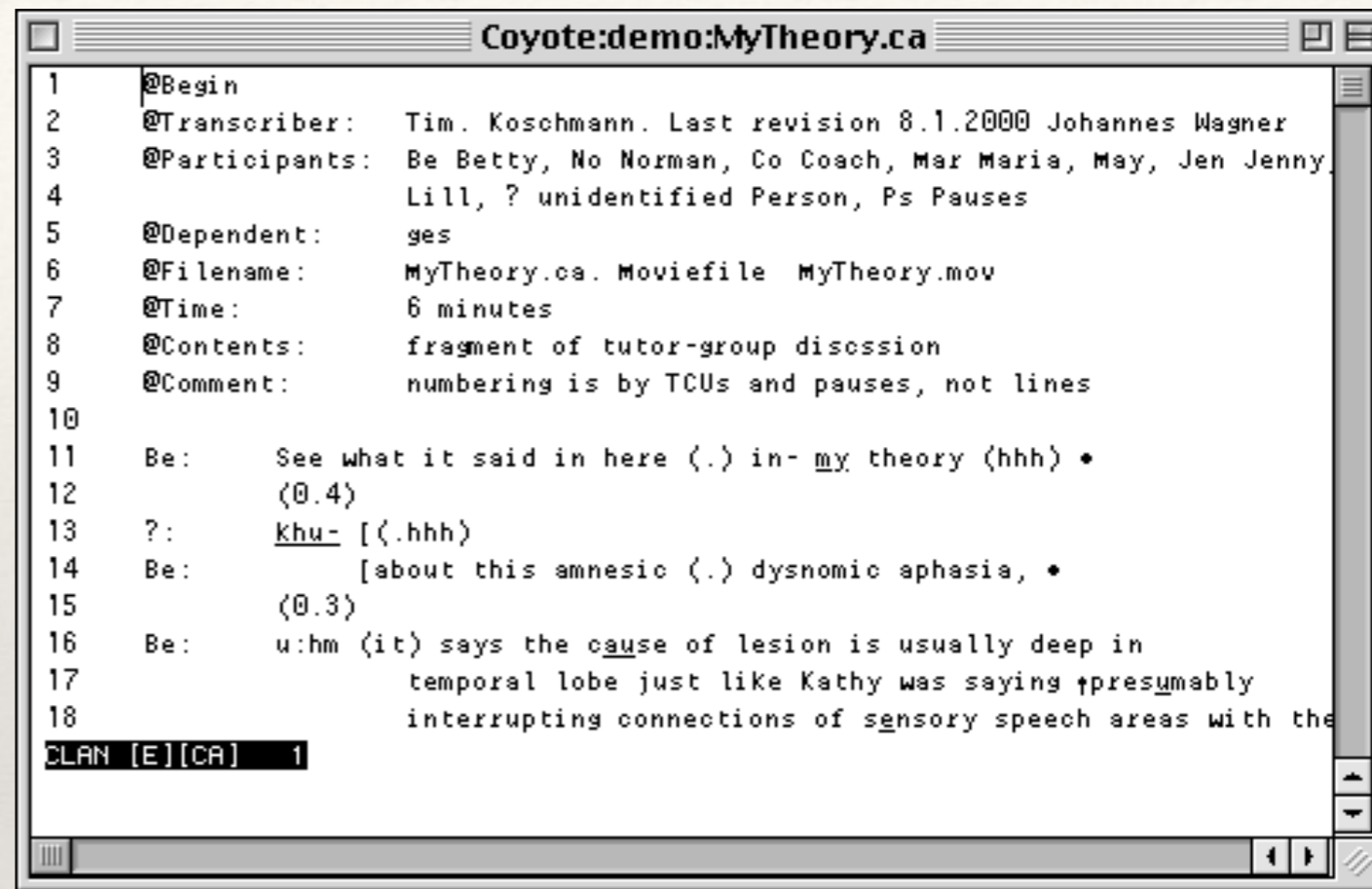
Error Analysis

- ❖ [*p] phonological p:w, p:n, p:m
- ❖ [*s] semantic s:r, s:ur, s:uk, s:per
- ❖ [*n] neologism n:k, n:uk, n:k:s, n:uk:s
- ❖ [*d] dysfluency
- ❖ [*m] morphology m:a:0es etc.
- ❖ [*f] formal lexical
- ❖ [+gram] [+jar] [+es] [+per] [+cir]

Method #3: Microanalysis

- ❖ Process frames show their effects in specific moments in time and space — on video.
- ❖ Consolidation is revealed across times.
- ❖ Microanalysis (CA) looks for practices, devices to see how they are conditioned

A sample moment: Transcript linked to video



You flip up that little temporal lobe

- ❖ Goal Stack: Med School, PBL, differential diagnosis, amnesic dysnomic aphasic, anterior cerebral circulation
- ❖ Where is the hippocampus?
- ❖ Lot more medial ,Pointing to diagram
- ❖ Finding right section
- ❖ Dealing with interaction
- ❖ Linking to CMaps

CA Coding

Special Characters	
↑	shift to high pitch; F1 up-arrow
↓	shift to low pitch; F1 down-arrow
↗	rising to high; F1 1
↖	rising to mid; F1 2
→	level; F1 3
↘	falling to mid; F1 4
↙	falling to low; F1 5
∞	unmarked ending; F1 6
≈	≈continuation; F1 +
.	inhalation; F1 .
≈	latching≈; F1 =
≡	≡uptake; F1 u
⌈	top begin overlap; F1 [
⌋	top end overlap; F1]
⌌	bottom begin overlap; F1 {
⌍	bottom end overlap; F1 }
Δ	ΔfasterΔ; F1 right-arrow
∇	∇slower∇; F1 left-arrow
*	*creaky*; F1 *
?	?unsure?; F1 /
°	°softer°; F1 0
⊙	⊙louder⊙; F1)
=	=low pitch=; F1 d
≡	≡high pitch≡; F1 h
☺	☺smile voice☺; F1 l
☼	☼breathy voice☼ marker; F1 b
	whisper ; F1 w
ÿ	ÿyawnÿ; F1 y
‡	‡singing‡; F1 s
§	§precise§; F1 p
~	~constriction~; F1 n
○	○pitch reset; F1 r
ℋ	ℋlaugh in a word; F1 c
„	„Tag or sentence final particle; F2 t
‡	‡Vocative or summons; F2 v

CHAT2ELAN

The screenshot displays the Elan software interface for a video file named 'mytheory.eaf'. The interface includes a menu bar (File, Edit, Annotation, Tier, Type, Search, View, Options, Window, Help), a video player window showing a classroom scene, and a control panel with sliders for Volume (0-100) and Rate (0-200). Below the video player is a playback control bar with various navigation icons and checkboxes for Selection Mode and Loop Mode. The main area is a transcription timeline with a time axis from 00:00:41.000 to 00:00:49.000. The timeline shows several tiers of annotations:

- *BET**: A pink bar spanning from 00:00:41.000 to 00:00:49.000.
- *UNK**: A light blue bar spanning from 00:00:41.000 to 00:00:42.000.
- *NOR**: A light green bar spanning from 00:00:41.000 to 00:00:42.000.
- *COA**: A light blue bar with the text 'You can you can point to it on' starting at 00:00:43.000.
- %gpx@NOR**: A light green bar spanning from 00:00:41.000 to 00:00:42.000.
- *MAR**: A light green bar with the text 'if you lift up +/.' starting at 00:00:41.000, 'that little temporal lobe' at 00:00:42.000, 'insid' at 00:00:43.000, '#0 Middle top?' at 00:00:44.000, and '0.' at 00:00:49.000.
- %gpx@MAR**: A light green bar with the text 'brings R hand in' starting at 00:00:41.000, 'lifts R hand above head' at 00:00:42.000, and 'Maria poin' at 00:00:49.000.
- %gpx@COA**: A light blue bar with the text 'Points with R hand from seat t' starting at 00:00:43.000.

CHAT2PRAAT - sociophonetics

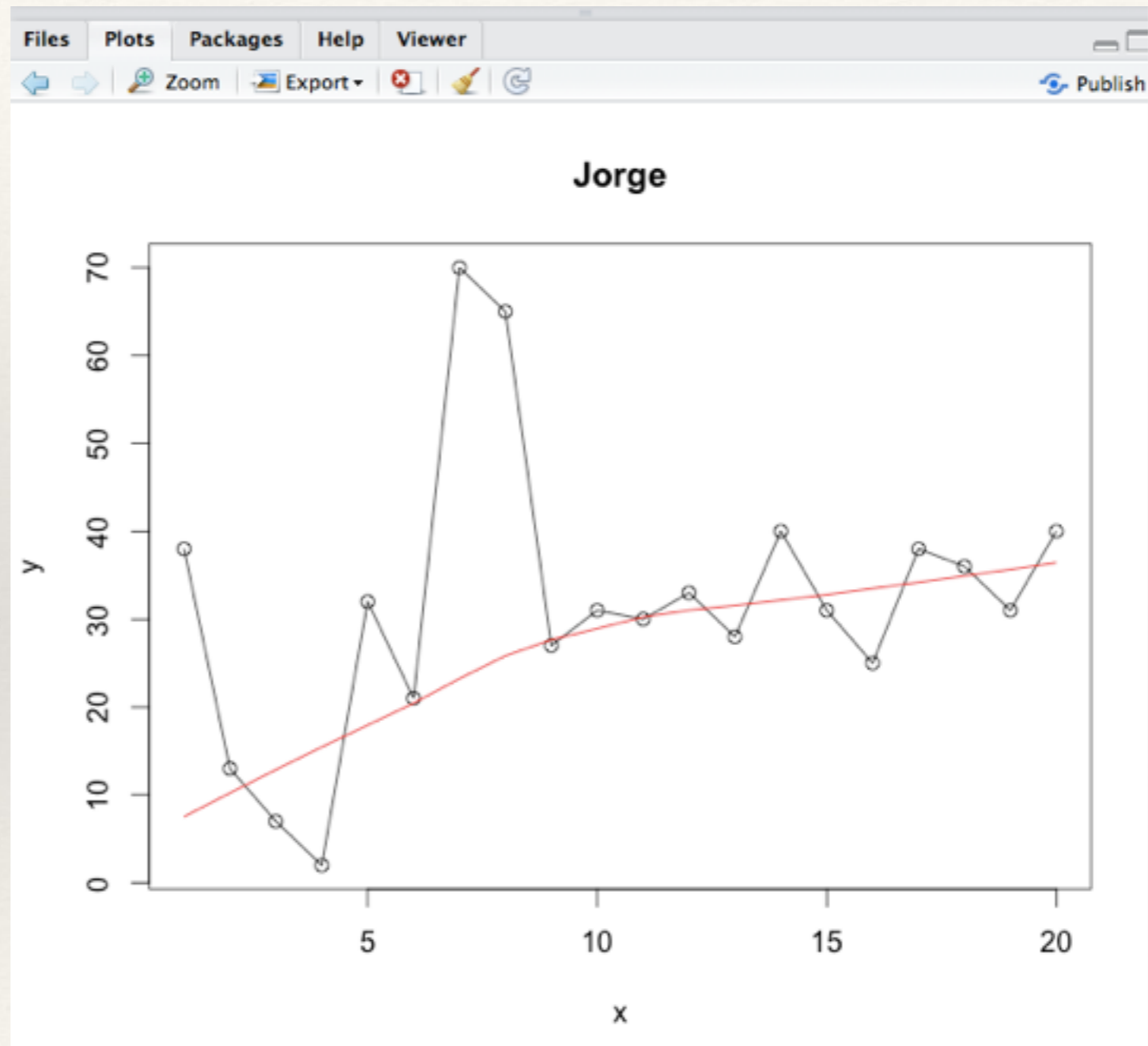
- ❖ Highlight utterance bullet
- ❖ Send to sound analyzer
- ❖ Extracts audio from video
- ❖ In Praat, draw a picture

The screenshot displays the Praat software interface. On the left, a list of applications is visible, including Maps.app, Messages.app, Microsoft Excel.app, Microsoft OneNote.app, Microsoft Outlook.app, Microsoft PowerPoint.app, and Microsoft Word.app. Below this, the Praat Objects window shows a list of objects, with '2. Sound ,äält_s_right_,Üdown_there,Ü' selected. The main window displays a waveform and a spectrogram. The waveform shows amplitude over time, with a vertical red dashed line indicating a specific time point. The spectrogram shows frequency (0 Hz to 5000 Hz) over time, with a blue line indicating a specific frequency component. The visible part of the audio is 0.678000 seconds, and the total duration is 0.678000 seconds. The transcript at the bottom shows the following text:

```
*BET: I don(t) do we have a picture up there → •
*BET: on the → •
*NOR: (It's right ↓down there ↓) •
%gpx: pointing with hand toward atlas from seat.
*NOR: it's the bottom of this thing → •
```

Time Series and R

Alberto and
Jorge — I no go.



Method #4: Web-based Tutors

1. E-CALL Tutors
2. Learning from Corpora
3. Language Learning in the Wild

PinyinTutor

<http://sla.talkbank.org/pinyin/#>

Features:

- Immediate Corrective Feedback
- Initial-final-tone separation
- Target-Your attempt comparison
- Recycling/Scheduling
- Linkage to textbook / or not
- Instructor report
- Data logged to server etc.
- Pages with rules of Pinyin
- Playable sound chart
- Multiple speakers

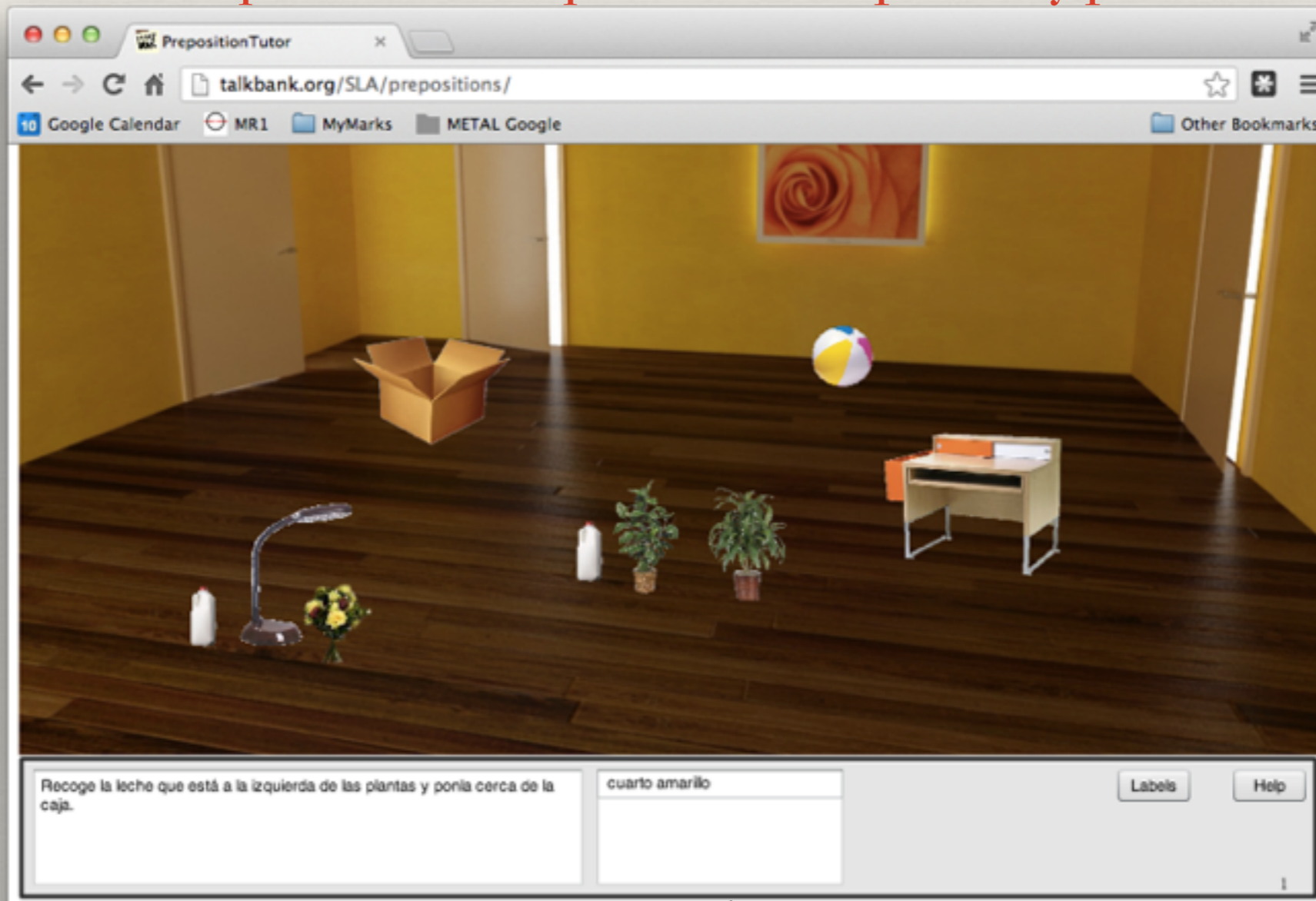


Virtual Reality Tutor

Spanish Prepositions and Relative Clause Processing

Take the milk to the left of the plants and put it next to the box

Recoge la leche que está a la izquierda de las plantas y ponla cerca de la caja.



Online Individual Difference Measures

The screenshot shows a web browser window with the URL `talkbank.org/SLA/tasks/test/`. The page features a navigation bar with links for 'Contribute', 'Calendar', 'CHILDES', 'Blackboard', and 'Kindle'. Below the navigation bar is a 'TalkBank' header with a search bar for 'Email' and 'Password', and a 'Sign in' button. The main content area is titled 'Cognitive Test Library' and includes a description: 'A set of free cognitive tests in eight languages. These tests are designed to work on all modern browsers and platforms including cellphones and iPads.' A blue 'Learn more »' button is positioned below the description. The page is divided into three columns, each featuring a test title, a brief description, a language dropdown menu (set to 'English'), and a green 'Launch Demo »' button. The tests listed are 'Digit Span', 'Number-Letter', and 'Flanker'.

TalkBank

Cognitive Test Library

A set of free cognitive tests in eight languages. These tests are designed to work on all modern browsers and platforms including cellphones and iPads.

[Learn more »](#)

Digit Span

In this test, participants hear a sequence of numbers. After the sequence finishes, the task is to type the numbers in their original order. For example, if participants hear "5, 2", their task is type 52 (without spaces).

English

Number-Letter

In this test, participants hear a mixed sequence of numbers and letters. After the sequence finishes, the task is to first type the numbers in numeric order, then the letters in alphabetic order. For example, if participants hear "7-c-3-a", their task is to type 37ac (without spaces).

English

Flanker

In this test, participants are asked to respond quickly and accurately to a series of images. Based on the image presented, participants are to respond/not respond to the direction of a red arrow.

English

Aligned Parallel Corpora

<http://sla.talkbank.org/latin/originalAlignmentDemo.html>

Sentence: 1 ↕

verb, 3rd person, plural, present, indicative mood, active voice

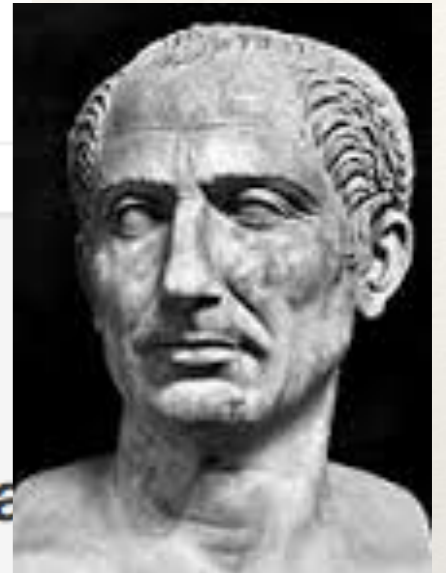
Gallia est omnis divisa in partes tres, quarum unam incolunt Belgae, a
ipsorum lingua Celtae, nostra Galli appellantur.

Word-by-word English translation (missing translations in parens):

Gaul (est) all divided into parts three, (quarum) one inhabit the Belgae
who own language called Celts, our Gauls third.

Free English translation:

All gaul is divided into three parts, one of which the belgae inhabit, the
language are called celts, in our gauls, the third.




Captioned Video

Hummus | DOVE Caption Browser

talkbank.org/DOVE/?file=Yinzers/hummus/

DOVE/ YINZERS/ HUMMUS



00:20 -02:03

« Last :: Pause Next » | Loop current utterance

EXPERIMENTAL: German : Translate

ENGLISH -

It's mine and Gina's 'round the world' night.

German and English through Wikipedia

Artikel Tages



Die **Wasseramseln** (*Cinclus*) bilden mit fünf Arten **die** einzige Gattung **der** Familie Cinclidae. Sie sind **der** Ordnung **der** Sperlingsvögel (Passeriformes) und **der** Unterordnung **der** Singvögel (Passeri) zugeordnet.

rundlich wirkenden,

finken- bis starengroßen Vögel kommen in Europa und Asien sowie in Nord-, Mittel- und Südamerika vor.

Eurasische Wasseramsel oder kurz Wasseramsel (*Cinclus cinclus*) brütet auch im Nordwesten Afrikas. Alle Arten leben entlang von schnellfließenden, sauerstoffreichen Gewässern, wo sie sich meist von Wasserinsekten und anderen aquatisch lebenden Wirbellosen ernähren, zum Teil tauchend und schwimmend erbeutet werden.

From today's featured article



The **Dorset Ooser** is **a** wooden head that featured in **the** nineteenth-century folk culture of **Melbury Osmond**, **a** village in southwestern **English** county of **Dorset**. **The**

head was hollow, thus perhaps serving as **a** mask, and included humanoid face with horns, beard, and hinged jaw. Although sometimes used to scare people during practical jokes, its main recorded purpose was as part of local variant of custom known as "**rough music**", in which it was used to humiliate those who were deemed to have behaved in immoral manner. It was first

Interesting Issues

- A Federation linked to CLARIN?
- Uniform Format, Data Types
- Open Access, IRB
- Metadata - Access
- Federated Content Search
- Sustainability

Federations

- CLARIN as an example
- Core Trust Seal
- Can the Americas support one?
- Suggestion: Recruit Linguistic departments
- Joining CLARIN vs. parallel to CLARIN
- What would be the benefits?
- Who could support this?

Uniform Format

- User only has to learn one set of programs
- Analyses can use multiple corpora, various ages, language backgrounds, etc.
- Clear definition of codes, verbal features, categories
- What data types are we covering?

Open Access, Data-Sharing, IRB

- TalkBank data are fully open and shareable
- Why aren't we all sharing data?
- Why haven't we adopted (at least) interoperable formats?
- Perhaps there is a credit assignment problem
- But we have: web pages, DOI, coauthorship, shared grants, citations in articles etc.

Metadata, Access

- My sense is that the emphasis on metadata has occluded the need to standardize formats
- Metadata is certainly important
- But what good is metadata if you can't actually access the materials?

Federated Content Search

- This seems to be a CLARIN idea — it does make sense
- But is it just access through metadata or can you really search and analyze multiple corpora across multiple sites
- Can we standardize search engines: CQL, ANNIS, MTAS, SearchEngine

Sustainability

- Federation, standardization, and the Cloud can help
- We seem to have lots of contacts with enterprise. How can this be leveraged?
- Motivated groups and projects are crucial
- Generational transition