

## Workshop on Research Methods in Linguistics: Corpus Data: TalkBank and CLAN

### 1. TalkBank

- TalkBank is a set of corpora established by Brian MacWhinney between 1999 and 2004.
- It contains databases from several subfields of linguistic research, including:
  - **First Language Acquisition** (+ Child-Directed Speech)
  - Child and Adult **Bilingualism**
  - **Second Language Acquisition**
  - Child and Adult **Clinical** Populations
- It currently includes corpora in **over 40 languages!**
- Some of these corpora include:

Corpus Name	Population type	Website
CHILDES	Children (and child-directed speech)	<a href="https://chilides.talkbank.org/">https://chilides.talkbank.org/</a>
CHILDES > Bilingual	Children acquiring two or more languages (and child-directed speech)	<a href="https://chilides.talkbank.org/access/Biling/">https://chilides.talkbank.org/access/Biling/</a>
BilingBank	Adult bilinguals	<a href="https://biling.talkbank.org/">https://biling.talkbank.org/</a>
SLABank	Children and adults acquiring a second language	<a href="https://slabank.talkbank.org/">https://slabank.talkbank.org/</a>
CHILDES > Clinical Corpora	Children with Specific Language Impairment or Speech Sound Disorders	<a href="https://chilides.talkbank.org/access/Clinical/">https://chilides.talkbank.org/access/Clinical/</a> <a href="https://chilides.talkbank.org/access/Clinical-MOR/">https://chilides.talkbank.org/access/Clinical-MOR/</a>
Aphasia Bank*	Adults with Aphasia	<a href="https://aphasia.talkbank.org/">https://aphasia.talkbank.org/</a>
ASD Bank	Children and adults with Autism Spectrum Disorder	<a href="https://asd.talkbank.org/">https://asd.talkbank.org/</a>
Dementia Bank*	Adults with Dementia	<a href="https://dementia.talkbank.org/">https://dementia.talkbank.org/</a>
Fluency Bank	Adults who stutter	<a href="https://fluency.talkbank.org/">https://fluency.talkbank.org/</a>
TBI Bank*	Adults with Traumatic Brain Injuries	<a href="https://tbi.talkbank.org/">https://tbi.talkbank.org/</a>

\*These corpora require a username and password. Brian ([macw@cmu.edu](mailto:macw@cmu.edu)) or myself can provide these upon request.

- You can find out **general information** about what each author transcribed (e.g., the age of the children, the context of the recordings...) by reading the corpus description on the author page.

- Some transcriptions also have **media files** (audio or video).
- The **Browsable Database** facility allows you to view and analyze transcripts as well. You may also playback transcripts with linked media directly from your browser.

childes.talkbank.org/browser/index.php?url=Spanish/Ornat/010900a.cha


Apps HUMNET ACCT Gmail Grammatical Theo... Anthem Insurance Pixton #Comics Course: Spanish &... Audio Visual Equi...

- 010900b.cha
- 010900c.cha
- 010900d.cha
- 010900e.cha
- 010900f.cha
- 010900g.cha
- 010900h.cha
- 010900i.cha
- 010900j.cha
- 010900k.cha
- 010900l.cha
- 011000a.cha

Command line: Spanish/Ornat/  
chains Run

Continuous playback: On: Off: ☐ ☒

Dependent tiers: %com: ☒ | %gra: ☐ | %mor: ☐ |  
Set options



```

0 @Loc: Spanish/Ornat/010900a.cha
1 @PID: 11312/c-00032726-1
2 @Begin
3 @Languages: spa
4 @Participants: CHI María Target_Child , MAD Mother , PAD Father
5 @ID: spa|Ornat|CHI|1:09.00|||Target_Child||
6 @ID: spa|Ornat|MAD|||Mother||
7 @ID: spa|Ornat|PAD|||Father||
8 @Media: 010900a, video
9 @Tape Location: Z-1;09-099
10 @Comment: DV Tape is maria03
11 @Situation: María en el baño.
12 @Types: long, toyplay, TD
13 *PAD: María se va a bañar .
14 *PAD: María , te vas a bañar y vas a hablar hoy , no ?
15 *CHI: sí .
16 *PAD: vale .
17 *MAD: suéltale un discurso a Papá por ser el día de los trabajadores ,
18 anda mi vida .
19 *CHI: sí ?
20 *MAD: sí , un discurso (.) xxx [% imita verbalizaciones de María] .
21 *MAD: todas esas cosas que dices tú .
22 *CHI: no oi(g)o [= oigo] .
23 *MAD: &=rie dile a Papá que te tengo dicho que "no sé que no sé cuantos"
24 .
25 *CHI: (m)(r)a acá [= caer/se cae] agua .
26 *CHI: 0 [% echa agua de un tapón a otro] (m)(r)a acá agua .
27 *MAD: qué ?
28 *CHI: nene acá no .
29 *CHI: mi(r)a agua no .
30 *MAD: no .
31 *CHI: e(l) nene acá [= caer/cae] no .
32 *MAD: a ti te gusta bañarte , María ?
33 *CHI: sí .

```

- However, in general, I recommend downloading the files over browsing online for a few reasons:
  - Long outputs are sometimes clipped in the browser version.
  - The site is sometimes down.
  - Your internet could go down!

## 1.1. Downloading TalkBank corpora

- Go to <http://talkbank.org/>
  - Select the corpora set you wish to look at.
  - Select **\*\*Index to Corpora\*\***
  - Select the collection/language you wish to look at.
  - Select a Corpus.
  - Click on 'Download transcripts'
  - Unzip the file.
  - Save it in any folder you want.
- Notice that all files have a .cha extension. They are written in CHAT format (more later).

## 1.2. TalkBank DataBase

- TalkBankDB lets you explore TalkBank's transcripts, specify data to be extracted, and pass these data on to other programs for further analysis: <https://talkbank.org/DB/>
- Step 1: Select transcripts by Collection, Language, Age, Gender...
- Step 2: Explore transcripts online or by downloading (click on "Save")
  - a) "Transcripts" tab: Shows a table with transcripts, media files available, languages spoken, date recorded, study design type...
  - b) "Participants" tab: Shows a table with speakers' IDs, role, age, gender, number of words and utterances produced by speaker per transcript...
  - c) "Token Types" tabs: Number of occurrences of all the words in the selected transcripts.
  - d) "Visualizations" tab: Creates a plot/table with specific word frequencies by age.

AutoSave

Home

Insert

Draw

Page Layout

Formulas

Data

Review

View

From HTML

From Text

New Database Query

Refresh All

Connections

Properties

Edit Links

Sort

Filter

Advanced

Clear

Reapply

Text to Columns

Flash Fill

Remove Duplicates

Data Validation

Consolidate

What-If Analysis

Group

Ungroup

Subtotal

Show De

Hide Det

A123

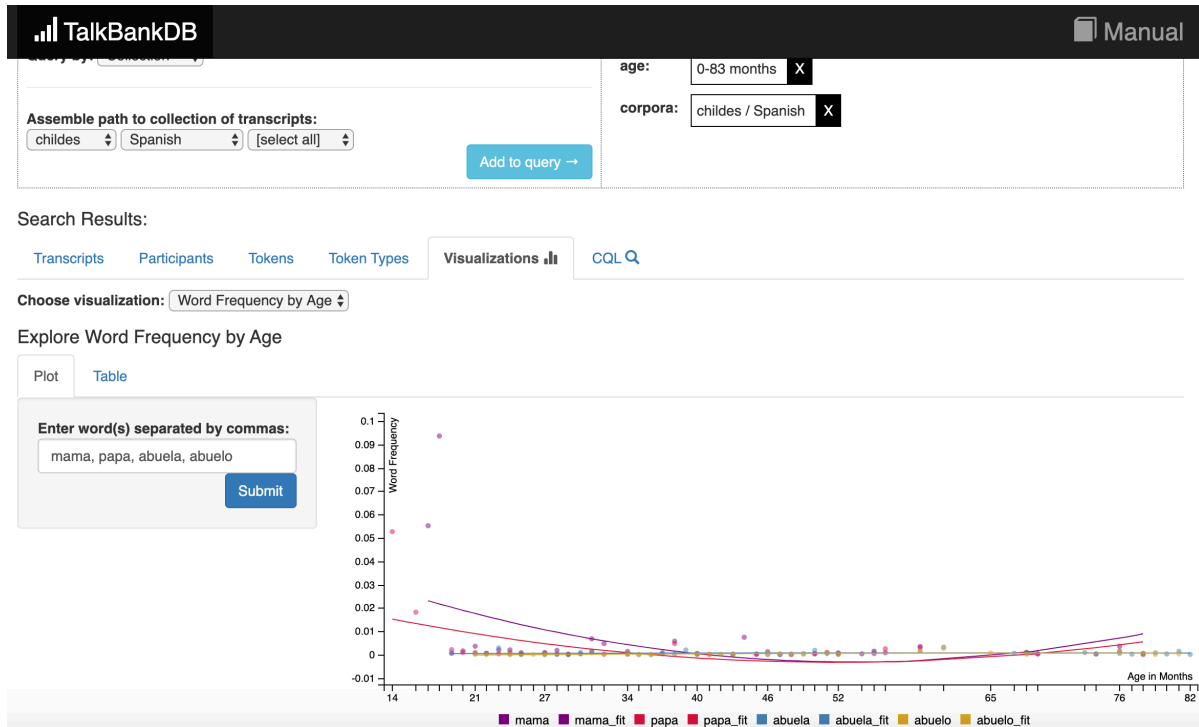
fx

03f07

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	File name	File Path	Participant ID	Participant Name	Role	Languages	Age in mos	Age in Y:M:D	Gender	Num words	Num utts	Avg wds/utt	Median wds/utt	
106	20726	chldes/Spanish/Aguirre/020726	FAM	null	Friend	spa	null	null	null	40	9	4.44444	4	
107	21024	chldes/Spanish/Aguirre/021024	CHI	Mag	Target_Child	spa	34	02;10;24	null	1498	573	2.61431	2	
108	21024	chldes/Spanish/Aguirre/021024	MOT	null	Mother	spa	null	null	female	2337	613	3.8124	3	
109	21024	chldes/Spanish/Aguirre/021024	FAT	null	Father	spa	null	null	male	31	6	5.16667	5	
110	21024	chldes/Spanish/Aguirre/021024	FAM	null	Friend	spa	null	null	null	262	47	5.57447	4	
111	02f10	chldes/Spanish/BecaCESNo/02f10	CHI	Helena	Target_Child	spa	34	02;10;23	null	315	78	4.03846	3	
112	02f10	chldes/Spanish/BecaCESNo/02f10	ARA	Aranzazu	Investigator	spa	null	null	null	379	101	3.75248	3	
113	03f04	chldes/Spanish/BecaCESNo/03f04	CHI	MariCarmen	Target_Child	spa,eng	40	03;04;01	null	2423	456	5.3136	4	
114	03f04	chldes/Spanish/BecaCESNo/03f04	FER	Fernando	Child	spa,eng	null	null	null	260	65	4	3	
115	03f04	chldes/Spanish/BecaCESNo/03f04	MAM	Madre	Girl	spa,eng	null	null	null	90	21	4.28571	4	
116	03f04	chldes/Spanish/BecaCESNo/03f04	PAT	Patricia	Investigator	spa,eng	null	null	null	2060	425	4.84706	4	
117	03f06	chldes/Spanish/BecaCESNo/03f06	CHI	Raquel	Target_Child	spa	42	03;06;	null	1184	188	6.29787	4	
118	03f06	chldes/Spanish/BecaCESNo/03f06	ANA	Ana	Investigator	spa	null	null	null	1568	229	6.84716	6	
119	03f06	chldes/Spanish/BecaCESNo/03f06	MAD	Madre	Mother	spa	null	null	null	70	13	5.38462	5	
120	03f07	chldes/Spanish/BecaCESNo/03f07	CHI	Clara	Target_Child	spa	43	03;07;19	null	1600	352	4.54545	3	
121	03f07	chldes/Spanish/BecaCESNo/03f07	MAR	Mari_Jose	Mother	spa	null	null	null	839	127	6.6063	5	
122	03f07	chldes/Spanish/BecaCESNo/03f07	JUA	Juan_Carlos	Father	spa	null	null	null	57	11	5.18182	5	
123	03f07	chldes/Spanish/BecaCESNo/03f07	CRI	Cristina	Investigator	spa	null	null	null	2151	436	4.93349	4	
124	03m04	chldes/Spanish/BecaCESNo/03m04	CHI	Sergio	Target_Child	spa	40	03;04;28	null	2534	536	4.72761	3	
125	03m04	chldes/Spanish/BecaCESNo/03m04	CAR	Carlos	Investigator	spa	null	null	null	1634	237	6.89451	6	
126	03m04	chldes/Spanish/BecaCESNo/03m04	TER	Teresa_Aunt	Relative	spa	null	null	null	1982	415	4.7759	4	
127	03m04	chldes/Spanish/BecaCESNo/03m04	PAD	Padre	Father	spa	null	null	null	8	5	1.6	1	
128	03m06	chldes/Spanish/BecaCESNo/03m06	CHI	David	Target_Child	spa	42	03;06;03	null	388	92	4.21739	3	
129	03m06	chldes/Spanish/BecaCESNo/03m06	EDU	Educadora	Investigator	spa	null	null	null	470	79	5.94937	5	
130	03m06	chldes/Spanish/BecaCESNo/03m06	PED	Pedro	Boy	spa	null	null	null	47	11	4.27273	2	
131	03m08	chldes/Spanish/BecaCESNo/03m08	CHI	Carlos	Target_Child	spa	44	03;08;06	null	403	70	5.75714	5	
132	03m08	chldes/Spanish/BecaCESNo/03m08	INT	null	Investigator	spa	null	null	null	39	6	6.5	5	
133	03m09	chldes/Spanish/BecaCESNo/03m09	CHI	Sergio	Target_Child	spa	45	03;09;10	null	1875	496	3.78024	3	
134	03m09	chldes/Spanish/BecaCESNo/03m09	MAI	Maitte	Investigator	spa	null	null	null	1857	425	4.36941	4	
135	03m09	chldes/Spanish/BecaCESNo/03m09	EVA	Eva	Mother	spa	null	null	null	56	14	4	4	
136	03m09	chldes/Spanish/BecaCESNo/03m09	JOS	Jose	Father	spa	null	null	null	9	3	3	2	
137	03m09	chldes/Spanish/BecaCESNo/03m09	ABU	Carmen	Grandmother	spa	null	null	null	115	25	4.6	4	
138	03m09	chldes/Spanish/BecaCESNo/03m09	OSC	Oscar	Child	spa	null	null	null	22	10	2.2	2	
139	04f01	chldes/Spanish/BecaCESNo/04f01	CHI	Marta	Target_Child	spa	48	04;00;11	null	2807	353	7.95184	6	
140	04f01	chldes/Spanish/BecaCESNo/04f01	MOT	Marta	Mother	spa	null	null	female	232	60	3.86667	4	

Corpus - Spanish Transcriptions

+



## 2. CLAN (Computerized Language Analysis)

- This is the program used to (i) **read, create** (§2.2) and (ii) **analyze transcriptions** (§2.3).
- It allows you to perform a large number of automatic analyses, such as frequency counts, word searches, morpheme searches, MLU (Mean Length of Utterance) calculations, morphosyntactic analyses, etc.

### 2.1. Installing CLAN

1. From the CHILDES website, click on 'CLAN', under 'Programs', or access this link:  
<https://dali.talkbank.org/clang/>
2. Click on Windows, Mac, or Unix
3. Save it where it suggests (i.e. in the C:\ directory in Windows, and the Applications folder in Mac)
4. Follow the directions from here (basically keep clicking "Continue")

### 2.2. Guidelines for transcription in CHAT

- Transcriptions on TalkBank are all in chat (.cha) format (Codes for the Human Analysis of Transcripts) they *should* follow certain formatting guidelines.
- Manual: <https://talkbank.org/manuals/CHAT.pdf>

#### 2.2.1. File tiers

- a) **Header lines**: Header lines give information about the participants and the setting. All headers begin with the "@" sign. This makes them invisible to CLAN for speech analysis. Headers that take an entry are followed by a colon, and a tab.
- b) **Speaker tiers**: Speaker tiers or main lines indicate what was actually said. All speaker tiers begin with an asterisk \*. Each main line should code *one and only one utterance*. When a speaker produces several utterances in a row, code each with a new main line.

After the asterisk on the main line comes a 3-letter code in upper case letters for the participant who was the speaker of the utterance being coded, child – CHI, patient – PAT, experimenter – EXP. After the three-letter code comes a colon and then a tab.

- c) **Dependent tiers**: Lines beginning with the % symbol can contain codes and commentary regarding what was said. They are called "dependent tiers". The % symbol is followed by a three-letter code in lowercase letters for the dependent tier type, e.g. phonology – pho, morphology – mor, comment – com, action – act; followed by a colon; and then a tab.

Example mini .cha transcription:

```
@Begin
@Languages: spa
@Participants: CHI Alfonso Target_Child, FAT Rafael Father
@ID: eng|macwhinney|CHI|2;10.10|||Target_Child||
@ID: eng|macwhinney|FAT||||Father||
*FAT: a_ver , enséñame tu dibujo .
*FAT: eso es plastilina , no ?
%act: señalando un trozo de plastilina pegado al dibujo .
*CHI: sí , esta es [/] esta es mamá .
@End
```

## 2.2.2. Structure

- a) **@Begin**: The first line in the file must be an @Begin header line.
- b) **@Languages**: The languages entered here use a three-letter ISO 639-3 code (e.g., English – eng; Spanish – spa; French – fra; German – deu; Mandarin Chinese – cmn; Korean – kor; for a complete list, visit: <http://www-01.sil.org/iso639-3/codes.asp>, check page 31 of the CHAT manual, or check the file *ISO-639.cut* inside *CLAN>lib>fixes* folder, e.g.,

```
@Languages: eng, spa
```

- c) **@Participants**: This line lists the three-letter codes for each participant, e.g., CHI for target child, MOT for mother, INV for investigator, PAT for patient, PAR for participant, etc., along with the name and their role, e.g.

```
@Participants: CHI Chip Target_Child, MOT Sharon Mother, FAT Doug
Father, INV Victoria Investigator
```

- d) **@ID**: Following these come a set of @ID headers, one for each participant, providing further details for each speaker. The form of the line is:

```
@ID:    language|corpus|code|age|sex|group|SES|role|education|custom|
```

- There must be one @ID field for each participant. Often you will not care to encode all of this information. In that case, you can leave some of these fields empty, e.g.  

```
@ID: spa|FernAguado|CHI|3;1.13|||Target_Child|Ainhua|
```
- Note age must be written in the format Y;M.D.

- For more control over creation and modification of these @ID headers, you can use the dialog system that comes up when you have an open CHAT file and select “Tiers” from the top Menu pulldown. Then click on “ID Headers”. Here is a sample version of this dialog box.

- e) Other **optional** lines that follow @ID include:

```
@Birth of #: dd-lll-yyyy      [NB: Month is first 3 letters, e.g., JUL]
@Birthplace of #: City, Country
@Location: City, Country [of recording]
@Date: dd-lll-yyyy [of recording]
@Time Duration: hh:mm:ss
@Transcriber: LastName, Name
@Situation: Details about situation during recording
@Comment: All-purpose comment line
```

- f) **@End**: The last line in the file must be an @End header line.

- In order to make sure that a file matches the requirements for correct analysis through CLAN, transcribers should run each file through the CHECK command: esc+l.

### 2.2.3. Speaker tiers: Words and Utterances

- a) **Utterance terminators:** Utterances must end with an utterance terminator. The basic utterance terminators are the period, the exclamation mark, and the question mark. These may be preceded by a space (but not necessarily), e.g.,

\*CHI:    leche .  
 \*CHI:    ves mami ?  
 \*CHI:    gatito malo !  
 \*MOT:    qué +!?

- b) **Intonation breaks:** Commas can be used as needed to mark intonation breaks, e.g.,

\*CHI:    mira un perro , mami .

- c) **Upper case use:** Use upper case letters only for proper nouns and the word “I” in English. Do not use uppercase for the first words of sentences.

- d) **Standardized spellings:** Marginal words should be spelled in standard ways. See *Oallwords* file in CLAN>lib>eng or *spa* folder (once you download the MOR dictionaries) for a complete list of standardized spellings (more later).

- no, nanay, mhm
- ee, eh, eh
- wow, guau
- oh, oo, oy, ui
- uay, guay
- okay, okey
- ssh, shh

- e) **Acronyms and Compounds:** Use underscore between letters, e.g. U\_S\_A, U\_C\_L\_A (exception: tv) or between words in compounds, e.g., arco\_iris (or + sign: video+juego) and fixed expressions, e.g., a\_ver, vaya\_por\_Dios, mecachis\_en\_la\_mar

- f) **Numbers:** They should be written in words, with no hyphens in between, e.g., 256 should be doscientos cincuenta y seis.

- g) **Omitted words:** They may be preceded by 0.

\*PAT:    voy 0a comprar .

- h) **Incomplete words:**

- When the complete word is known, it can be written with the omitted material in parentheses, as in (be)cause, tranqui(lo), or for non-standard shortenings, e.g. senta(da), ama(r)illo
- When the omitted segment is unknown or due to a disfluency/interruption, use &+ preceding the phonological fragment, e.g.  
 \*PAT: Es un &+ka gato .

i) **Unclear/Unintelligible words:**

- Make a best guess and follow with [?]
- Precede with an ampersand if they have a clear phonetic shape but no meaning, as in &guga.
- Use xxx if they have an unclear phonetic shape and you can't guess.

j) **Untranscribed material:** IFF you have a %pho tier, use yyy for non-standard pronunciations, otherwise, use IPA in speaker tier followed by @u (see below in 'special form markers') e.g.

\*CHI:      yyy queso .  
 %pho:      kiro keso

k) **Disfluencies:** Disfluencies or fillers can be preceded by & or &-, e.g., &ah, &eh, &uh, &uhm, &hml) **Unfilled pauses:** Pauses are coded with (.), (..), or (...) depending on length,

\*MOT:      No (...) creo que no .

m) **Pauses within words:** Indicate with circumflex ^, e.g., plá^tanosn) **Repetitions:** They are separated by [/]. Material being retraced is enclosed in angle brackets <>, unless there is a pause or filler, e.g.

\*PAT:      <yo quería> [/] yo quería invitar a Juan .  
 \*CHI:      es [/] (.) &ehm es difícil .

o) **Retractions:** When a speaker says something, stops, and reformulates the sentence. They are separated by [//], e.g.

\*CHI:      <una mesa> [//] mira , mami , una mesa .

p) **False starts:** When a speaker says something, stops, and starts with a complete separate idea.

\*PAT:      <te gusta> [/ -] quieres uno ?

q) **Replacements:** For replacements, use [: text] [\*]

\*PAT:      no cabo [: quepo ] [\*] .  
 \*PAT:      no kabo@u [: quepo ] [\*] .

r) **Lengthening within words:** Indicate with colon :, e.g. n:o, carr:eteras) **Trailing off:** Indicate with plus sign and three periods, +...

\*MOT:      huele bien +...

t) **Interruptions:** Indicate with plus sign, dash, and period, +/.

\*MOT:      qué has +/ .  
 \*CHI:      mami !  
 \*MOT:      +, dicho ?

- u) **Overlap**: Text enclosed in angle brackets is being said at the same time as the following speaker's bracketed speech. This code must be used in combination with the "overlap precedes" symbol, as in this example:

\*MOT: no (.) Sara (.) tienes que <parar , eh > [>] ?  
 \*SAR: <mhm , no me gusta> [<] !

- v) **Special form markers**:

- Babbling: @b e.g., bababa@b
- Echolalia, repetition: @e e.g., quieres@e más@e
- Singing: @s e.g., lalala@s
- Onomatopoeia: @o e.g., guau+guau@o
- IPA string: @u e.g., tʃatʃitʃon@u [: salchichón ] [\*]
- Use of @u in speaker tier should be limited. If many words are mispronounced, consider adding a %pho tier for the child.
- NB: Not all transcribers use @u when using IPA.

- w) **Special events**: They can be preceded by an ampersand followed by an equal sign, &=event or in square brackets followed by an equal sign and an exclamation point, [=! event] .

- &=coughs
- [=! llorando]
- &=laughs
- &=mumbles
- [=! points]
- &=canta
- &=sneezes
- &=bosteza
- [=! yells]

- x) **Giving details on special events**:

- &=head:yes
- &=hands:hello
- &=ges:come
- &=shows:picture
- &=moves:doll
- &=eats OR &=eats:cookies
- &=drinks:milk
- &=points:dog
- &=turns:page
- &=hits:table
- &=imit:lion

## Summary

Error/Disfluency	Code	Example
Omitted words	0	voy 0a salir
Word fragment	()	cole(gio)
Phonological fragment	&+	&+tri tigre
Non-word strings	&	&gaga
Unclear words	[?]	quiero una rana [?]
Unintelligible words	xxx	quiero xxx .
Untranscribed (+ %pho)	yyy	*CHI: una yyy ! %pho: una galatixa
Disfluencies	& or &-	&uhm
Pauses	(.) or (...) or (...)	pues (.) no sé
Pauses within words	^	ele^fante
Repetitions	<> [/]	<me> [/] me parece que no
Retractions	<> [//]	<creo> [//] creemos que es lo mejor
False starts	<> [/ -]	<me gustaría> [/ -] quieres venir ?
Replacements	[ : ] [ * ]	abaso [ : abrazo ] [ * ]
Lengthening within words	:	s:erpiente
Trailing off	+...	me gustan los perros +...
Interruptions	+/	me gusta dibujar +/
Overlap	<> [>][<]	*CHI: <mami> [>] . *MOT: <ssh> [<] !
Special form markers	@b or @e or @o or @u or @s	guau@o rana@u [ : rana ] [ * ]
Special events	&= or [=!]	&=tose [=! grita]

### 2.2.4. Dependent Tiers

- a) **Action tier**: This tier describes the actions of the speaker or the listener. Here is an example of text accompanied by the speaker's actions:

\*CHI: yo quiero !  
%act: runs to toy box

- b) **Comment tier**: This is a general-purpose tier:

\*CHI: qué asco: !  
%com: child is talking about baby sister's diaper

- c) **Phonological tier:** You must use Unicode characters. To facilitate this, you can use one of these websites: <http://ipa.typeit.org/> or <https://westonruter.github.io/ipa-chart/keyboard/> . If you have Mac, you may install the IPA Palette: <https://www.blugs.com/IPA/>

\*CHI: me he hecho pupa .

%pho: me e 'etʃo 'pupa

- d) **Morphological tier**: This tier codes morphemic segments by type and part of speech. This tier may be created automatically(!) through the command MOR. Here is an example of the %mor tier, e.g.

\*MOT: leemos el cuento (.) quieres ?

```
%mor:  v|lee-1P&PRES=read det:art|el&m&SG=the n|cuento&m=story
       v|quiere-2S&PRES=want ?
```

- For a full description of the morphological analysis:  
<https://talkbank.org/manuals/MOR.pdf>
- **Words:** Beneath the level of the word group is the level of the word. The structure of each individual word is (parenthesis indicates it's optional):

```
prefix#part-of-speech|stem{&fusionalsuffix  
-suffix  
~clitic}=English_translation
```

E.g., deshacer = des#v|hac-INF=undo  
ardillita = n|ardilla&f-DIM=squirrel  
niños = n|niñ-m-PL=child OR n|niño-m-PL=child  
fui = v|i&1S&PRET=go OR v|i-1S&PST=go  
leerlos = v|lee-INF~pro:clit|OBJ&m&PL=read

- **Part of speech codes:** The basic scheme is 'category:subcategory' (capital letters or not, it doesn't matter). The most commonly used labels are included in the table below, but you can always add more information, e.g., `V: INTRANS` for intransitive verbs.

Category	Code	Example
Adjective	ADJ	ADJ   big
Adverb	ADV	ADV   well
Communicator	CO	CO   aw
Conjunction	CONJ	CONJ:COO   and CONJ:SUB   if
Determiner	DET	DET:ART   a
Filler	FIL	FIL   uh
Interjections	INT	INT   hi
Noun	N	N   cat

Proper Noun	N:PROP	N:PROP   Mary
Number	DET:NUM	DET:NUM   three
Onomatopoeia	O	O   woof
Preposition	PREP	PREP   in
Pronoun	PRO	PRO   I
Quantifier	QN	QN   all
Verb	V	V   walk
Auxiliaries AND modals	V:AUX	V:AUX   can
Wh-words	WH	WH   who

- **Stem:** This is the basic form of the word. For nouns, this is the singular. For verbs, it is the bare form in English (e.g., *play*), the infinitival form minus the final -r in Spanish (e.g., *se* for *ser*).
- **Affixes and clitics:** are coded in the position in which they occur with relation to the stem. The morphological status of the affix should be identified by the following markers or delimiters: - for a suffix, # for a prefix, & for fusional or infixed morphology, ~ for clitics.

Inflectional affixes in English (cells in grey = only necessary when verbs are overtly marked for that morpheme)

Function	Code	Example
plural	PL	N   flor-PL, N   man&PL
masculine	M	N   gato-M
feminine	F	N   mesa&F
person and number	1S... 3P	V   walk-3S, V   be&1S, V   salta-3P
present	PRES	V   pone-3P&PRES
past	PAST or PRET	V   walk-PAST, V   despertá-3S&PRET
future	FUT	V   ve-2S&FUT
present participle	PRESP	V   anda-PRESP
past participle	PASTP	V   encontra-PASTP, V   rompe&PASTP
diminutive	-DIM	N   casa&F-DIM

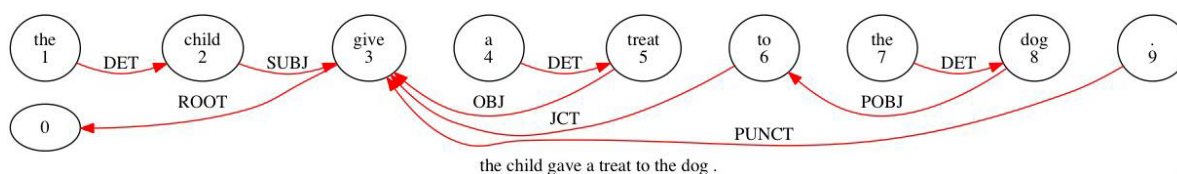
- **Compounds:** They're indicated with '+'. First you indicate what the whole word is in terms of part of speech, then you indicate each of the two parts preceded by '+'. E.g. playground  
n | +v | play+n | ground

e) **Grammatical tier:** This tier codes grammatical relations. This tier may be created through the command MOR. Here is an example of the %gra tier, e.g.

\*MAR: The child gave a treat to the dog .

%mor: det | the n | child v | give&PAST det | a n | treat prep | to det | the n | dog .

%gra: 1 | 2 | DET 2 | 3 | SUBJ 3 | 0 | ROOT 4 | 5 | DET 5 | 3 | OBJ 6 | 3 | JCT 7 | 8 | DET 8 | 6 | POBJ  
9 | 3 | PUNCT



- 1) The first part is a number that indicates word order within the utterance.
- 2) The second part is a number that indicates the dependent relationship between the word and the word it modifies (adjuncts to heads, subjects and objects to verbs; verbs are the root and hold a 0)
- 3) The third part is the grammatical category, e.g.,
  - ROOT = ROOT is the “head” of the clause, typically the main verb.
  - SUBJect = SUBJ is the subject of the clause.
  - OBJect = OBJ is the first complement (direct object) of the verb.
  - OBJect2 = OBJ2 is the second complement (indirect object) of the verb.
  - PrepositionalOBJect = POBJ is the complement or adjunct noun that follows a P.
  - adJunct = JCT is an adjunct that modifies the verb (including Ps in complement PPs).
  - MODifier = MOD is a nominal modifier.
  - DETerminer = DET is an article, possessive pronoun, or demonstrative.
  - QUANTifier = QUANT is a number or quantifier determiner.
  - NEGation = NEG is a verbal negative particle.
  - ...

➤ For a full description of the syntactic dependency analysis, see:

<https://talkbank.org/manuals/MOR.pdf> (Chapter 10)

NB: This function is still *very much* in development.

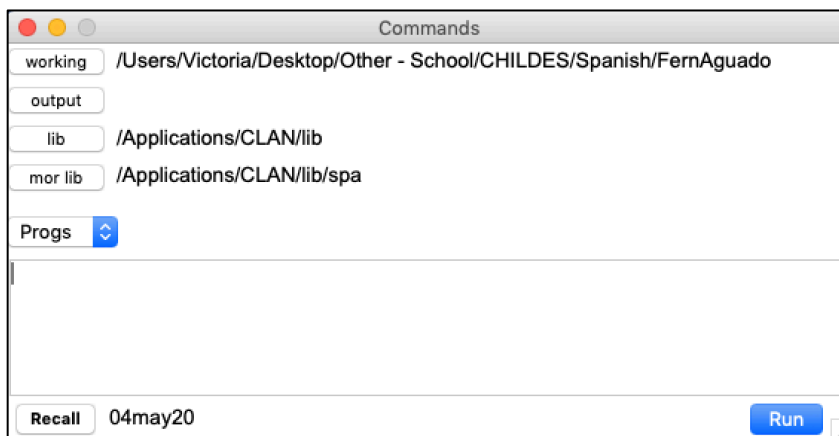
## 2.3. Guidelines for analysis with CLAN

- CLAN also allows users to conduct the basic operations of corpus analysis on TalkBank, such as frequency profiling, co-occurrence computation, utterance length calculation, automatic tagging for part of speech and grammatical dependency structures, among other functions.
- Manual: <https://talkbank.org/manuals/CLAN.pdf>

### 2.3.1. Setting up CLAN for analysis

1. Select “Windows” from the upper Menu. Then click on “Commands”, and you should get the command box.
2. Click on “working” and select the *folder* where the .cha files you want are.
3. You do not need to worry about setting the “output” directory. By default, it is the same as the working directory.

4. Click “lib” and select the CLAN library within the CLAN folder.
5. Click “mor lib” and select the morphological dictionary of the language of your transcriptions.  
To obtain those dictionaries:
  - i) Download the MOR grammar in one of the available languages here: (<http://talkbank.org/morgrams/>)
  - ii) Save it in the CLAN “lib” folder, inside the CLAN application folder (in “~Applications/CLAN” in Mac, or inside C:/TalkBank/CLAN in Windows).
  - iii) Go to the CLAN command window, click on “more lib” and select the language folder where the MOR grammar is (e.g. if you downloaded it for English, select the CLAN>lib>eng folder).
6. In the command box itself, you can type commands.



### 2.3.2. Commands

- Commands have several different build-up blocks, but before we get to them you should keep in mind the following points:
  - **IMPORTANT NOTE:** Although there have been considerable efforts to make all of TalkBank as homogeneous as possible, each corpus has been transcribed and perhaps coded by a different person and they may use different annotation standards (e.g., some may use “PRET” for preterite/past, while others may use “PAST”, some may code the target child as CHI others as ADA [for Adam]).  
So, in order to make productive searches, you have to make sure that what you’re looking for actually exists in the files!  
*The first step before running a CLAN analysis will **ALWAYS** be to open the file(s).*

▪ **Before getting started:**

- **Spaces:** Leave a space between each command part (e.g., between the type of analysis, and the speaker where you are restricting the analysis to).

```
mlu +t*CHI 020600.cha
```

- **Order** of components within a command is **irrelevant**.

- **Wildcards:** The asterisk symbol (\*) can be used in place of something else. For example, if you want to use your command across a group of 10 files all ending in .cha, you may type \*.cha

```
mlu +t*CHI *.cha
```

- **Recalling commands:** Click on “Recall” or press the ↑ button on your keyboard while you’re on the CLAN command window. It will show you the last command(s) you ran.

▪ **Basic parts of a command:**

1. **Specify the type of analysis** you want to do:

- **freq** Gives you the of times the string you specified was said by a speaker
- **kwal** Gives you the utterances containing the string you specified said by a speaker
- **combo** Gives you the of times the string you specified was said by a speaker and also lists the actual utterances
- **mor** Automatically creates a morphological tier (%mor) (and a grammatical tier (%gra) for some languages).
- **mlu** Gives you the Mean Length Utterance of a speaker. It requires a %mor tier.
- **kideval** Also eval. Creates a spreadsheet with a series of language development markers, e.g., number of utts, MLU, MLUw, number of different words, number of clauses per utterance... (More later). Also requires a %mor tier.

2. **Specify the tier(s)** you want to work with (make sure they exist in the specified file!). You may include one or more speakers with a + sign or exclude them with a – sign.

- **+t\*CHI** Searches the \*CHI utterances (IFF there is a \*CHI tier!)
- **-t\*CHI** Searches all utterances produced by all speakers except for the target child (make sure the target child is identified as \*CHI!)
- **+t%mor** Searches the %mor tier (IFF there is a %mor tier!)

3. **Specify the search** (if applicable to the type of analysis):

- **+s“yo”** Searches for the string “yo” (surrounded by spaces)
- **+s“gust\*”** Searches for strings starting with “gust-” (followed by space or any number of additional characters)
- **+s“n|\*”** Searches for the character string “n|” followed by anything (i.e. all nouns)
- **+s“\*PRET\*”** Searches for the string “PRET” preceded and followed by anything (i.e. all verbs in the past)

Some useful symbols for searches:

- ^ Immediately followed by
- ^\*^ Eventually followed by
- + Inclusive 'or' (NB: + does *not* work along with the command "freq". Use "combo" every time instead, or use "freq" and have two separate searches in the command line, e.g. `freq +t*CHI +s"mama" +s"papa" *.cha`)

#### 4. Specify the file(s):

- `sample.cha` Your search will be conducted on this specific file
- `*.cha` Your search will be conducted on all the files that end in `.cha` in the folder specified in "working"
- `02*.cha` Your search will be conducted on all the files that start with 02 (age 2;X) and end in `.cha` in the folder specified in "working"

#### ▪ Optional parts of a command:

1. Specify the window surrounding the data (do you want to see the line before the target data, and the line after? Then use `+w1` for the *speaker* line after and `-w1` for the *speaker* line before)
  - `-w3` Your search will include the target utterance as well as the 3 immediately preceding utterances, i.e., speaker tiers.
  - `-w1 +w1` Your search will include the target utterance as well as the immediately preceding utterance and the immediately following one.
2. Run the command recursively on all subdirectories of a folder, e.g., if the corpus FernAguado has 47 subfolders and you want to analyze the CHAT files within them, you may select "FernAguado" in "working" and the `+re` function will allow CLAN to "see" inside the subfolders as well.
  - `+re`
3. Merge results together in the output (useful for `freq` analyses).
  - `+u` Merges file results
  - `+o3` Merges speaker results
4. Sort `freq` output by descending frequency as opposed to alphabetically.
  - `+o`
5. Save the data to a document
  - `>sample.txt`

NB: The file will be saved in the folder you are working with unless you specify a different output location.

*Tip:* For frequency counts, it may be helpful to import the `.txt` file into excel and select the delimited data to be separated by spaces/tabs or through a fixed width.

### 2.3.3. Running commands: Examples

To run these examples, download the BecaCESNo and FernAguado corpora, unzip them, and save them in your computer. Then select the folder you want as your working directory.

<https://chilides.talkbank.org/access/Spanish/BecaCESNo.html>

<https://chilides.talkbank.org/access/Spanish/FernAguado.html>

➤ Let's start with BecaCESNo.

**a. `mor +t* 03f04.cha`**

This creates morphological %mor and grammatical %gra tiers for all speakers in this file.

NB: The file will be overwritten. If you want to preserve the original file, make a copy first.

**b. `mlu +t*CHI 03f04.cha`**

This gives you the number of utterances, morphemes, and the Mean Length of Utterance (MLU) of the tier corresponding to \*CHI in each of the .cha files that CLAN found in the selected folder.

- Mini exercise: What's the MLU in morphemes of this child?

**c. `freq +t* 03f04.cha +o >freq.txt`**

This gives you a list of all the words all speakers produced in the file 03f04.cha, along with the word frequencies divided by speaker, and in descending frequency order.

- Mini exercise: What's the most frequent word the investigator (\*PAT) produces?

*Tip*: you can also open the file on excel and sort by other categories.

**d. `freq +t*CHI +s"pequeñ*" *.cha`**

This asks how many words that begin with the string "pequeñ-" (i.e., *pequeño*, *pequeña*, *pequeños*, *pequeñas*...) are produced by the target child in all the .cha files contained in the specified folder.

**e. `combo +t*CHI +s"mama+mamá+papa+papá" 03f04.cha`**

This asks how many times the words "mama", "mamá", "papa" or "papá" are produced by the CHI in the file 03f04.cha and give you a list of the utterances themselves.

- Mini exercise: Does the child produce 'mama' or 'papa' more often? In what context does he use 'papa'?

**f. `freq -t*CHI +s"muñeca" +s"muñeco" *.cha +re +u +o3`**

(Set the working directory to FernAguado, so you can search within all its folders.) This asks how many times the words "muñeco" or "muñeca" are produced by all the speakers *except* the children identified as CHI in all the .cha files within the folders inside FernAguado. The results will merge all the files (+u) and all the speakers (+o3) together.

- Mini exercise: What do children hear more often, 'muñeca' or 'muñeco'?

**g. freq +t\*CHI +t%mor +s"det:art|\*" 02\*.cha +re**

This asks how many times the child produces articles (coded as `det:art|e1&m&SG=the` or `det:art|un-F&SG=one...`) by asking how often the string "`det:art|`" appears in the %mor tier associated with the \*CHI speaker tier in all the files that start with 02- (i.e. where target child is 2yo) within the folders inside FernAguado.

NB: This one will take a second! Be patient.

- Mini exercise: Skim over the results. What do 2-yos use more often, definite or indefinite articles?

**h. combo +t\*CHI +s"pr\*+tr\*+cr\*+br\*+dr\*+gr\*+fr\*" \*.cha**

(Set the working directory to Ainhua.) This searches for all the \*CHI tiers where there are words that start with any of the onset consonant clusters available in Spanish: [pr-, tr-, kr-, br-, dr-, gr-, fr-].

NB: This search will also include cases in which the child dropped a consonant, i.e., instances transcribed by using parentheses, e.g., t(r)es.

- Mini exercise: Skim over the results. In cases of consonant cluster reduction, which consonant is omitted more often?

**i. combo +t\*CHI +t%mor +s"n|^adj\*" \*.cha**

**combo +t\*CHI +t%mor +s"adj|^n|^\*" \*.cha**

These search all the cases in which a noun is followed by an adjective or preceded by an adjective, i.e., it looks inside %mor tiers for the string "n|" followed by whatever (\*) and immediately followed/preceded by the string "adj|" followed by whatever (\*).

- Mini exercise: Which order is more common?

**j. combo +t\*CHI +t%mor +s"cop|^adj\*" \*.cha**

This searches all the times the child produces a copula immediately followed by an adjective, i.e., all %mor tiers associated with the child containing the string "cop|" followed by whatever (\*) eventually followed by "adj" followed by whatever (\*).

- Mini exercise: Does the child use 'ser' and 'estar' appropriately?

**k. kideval +t\*CHI \*.cha +re**

(Set the working directory to FernAguado again.) This creates an excel spreadsheet with the following information (among others):

- File: name of the file
- Age: age of the speaker in months
- Sex: sex of the speaker
- Group: speaker group he/she belongs to, e.g., SLI
- Total Utts: total utterances (excludes non-verbal utterances)
- MLU Utts: number of utterances, as used for computing MLU (excludes disfluencies, xxx...)
- MLU Words: MLU in words
- MLU Morphemes: MLU in morphemes
- FREQ types: total word types, as used for computing FREQ

- **FREQ tokens:** total words (tokens)
- **FREQ TTR:** type/token ratio
- **NDW 100:** number of different words in the first 100 words in the sample
- **Verbs/Utt:** verbs per utterance. This can be less than 1.0 for young children
- **TD Utts:** total number of utterances for each speaker (no exclusionary criteria)
- **Word Errors:** number of words involved in errors (marked with [\*])
- **Utt Errors:** number of utterances involved in errors
- **Retracing [/]:** number of retracings
- **Repetition [/]:** number of repetitions
- **The frequencies of each of Brown's (1973) 14 grammatical morphemes:** progressive -ing, plural -s, 3sg -s, 3sg irregular, possessive -s, past -ed, past irregular, articles, P in, P on, copula (contracted and not), auxiliaries (contracted and not).
- **Mini exercise:** What's the MLU in morphemes of 2;10-3yos (34-36mos) vs. 3;10-4yos (46-48mos)?  
*Tip:* Use the Excel to freeze panes, add rows, and calculate the average with the formula =AVERAGE(cells)
- **Extra exercise:** Download the corpora FerFuLice under the Bilingual corpora in CHILDES. Add a %mor tier to the files in the Spanish folder and then run kideval. Do you observe a significant difference between monolingual and bilingual children with respect to their MLU in those same age ranges? What about their vocabulary size (NDW)?