Transcribing

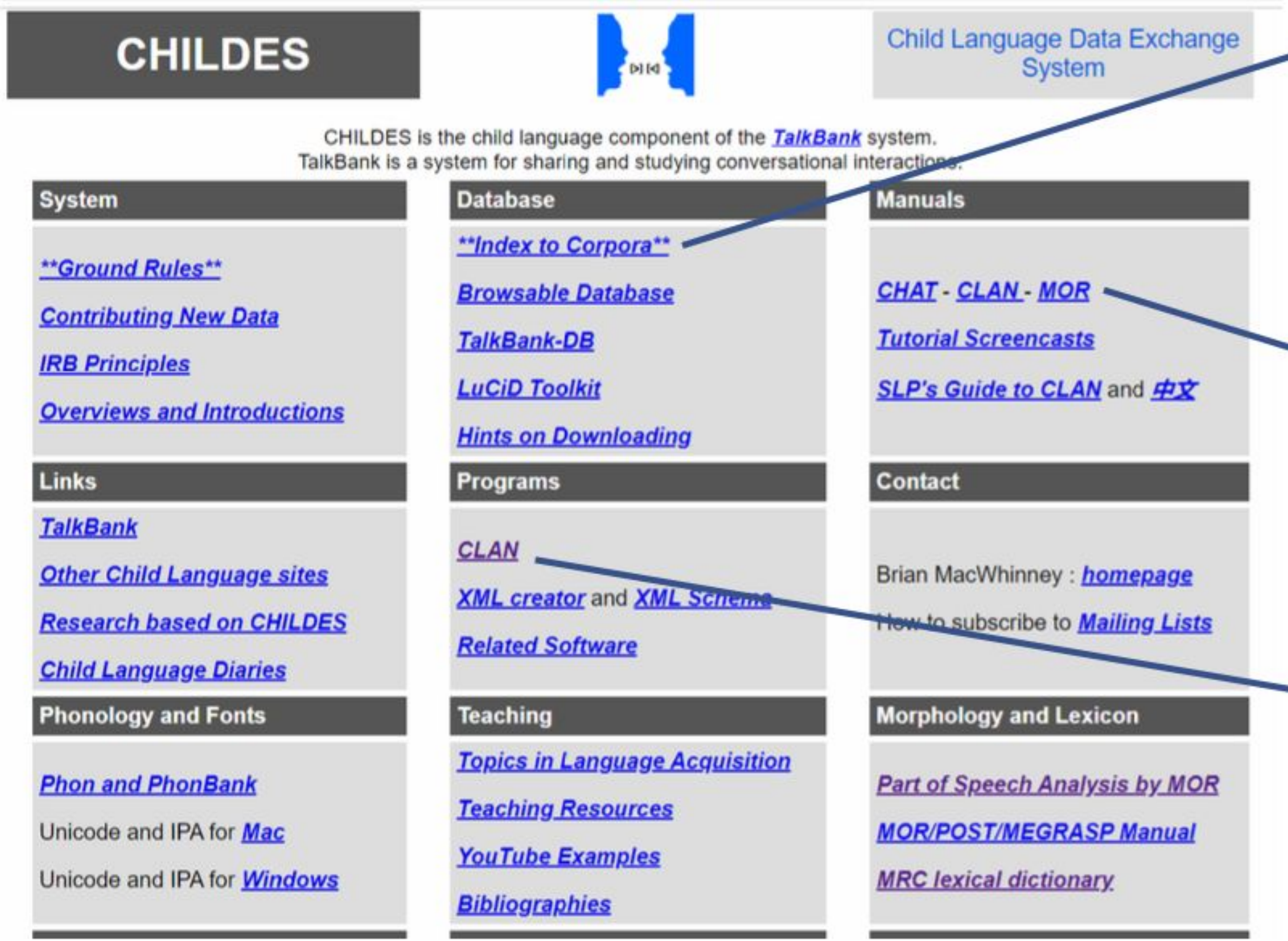
1. Introduction to CHILDES and CLAN

Child Language Data Exchange System (CHILDES)

- The Child Language Data Exchange System (CHILDES, pronounced as a single syllable) was established by Brian MacWhinney and Catherine Snow in 1984 to facilitate exchange of child language data
- Beginning in 2001, the CHILDES database concept was extended to a series of additional fields such as the studies of aphasia, dementia and second language acquisition. The idea of TalkBank was introduced (https://talkbank.org).

CHAT & CLAN

- TalkBank is the largest open repository of data on spoken language. All of the data in TalkBank are transcribed in the CHAT (Codes for the Human Analysis of Transcripts) format which is compatible with the CLAN programs.
- CHILDES, which is the oldest and most widely recognized in TalkBank, has been used in over 7000 published articles.
- For 10 of the languages in the database, automatic morphosyntactic analysis using a series of programs built into CLAN. Cantonese, Mandarin and English are three of them.
- The CHAT system is comprehensive, but it is not ideal for all purposes. The programs are powerful, but they cannot solve all analytic problems.



CHILDES

Child Language Data Exchange System

CHILDES is the child language component of the *TalkBank* system.
TalkBank is a system for sharing and studying conversational interactions.

System

- ***Ground Rules***
- *Contributing New Data*
- *IRB Principles*
- *Overviews and Introductions*

Database

- ***Index to Corpora***
- *Browsable Database*
- *TalkBank-DB*
- *LuCiD Toolkit*
- *Hints on Downloading*

Manuals

- *CHAT* - *CLAN* - *MOR*
- *Tutorial Screencasts*
- *SLP's Guide to CLAN* and *中文*

Links

- *TalkBank*
- *Other Child Language sites*
- *Research based on CHILDES*
- *Child Language Diaries*

Programs

- *CLAN*
- *XML creator* and *XML Schema*
- *Related Software*

Contact

Brian MacWhinney : *homepage*

How to subscribe to *Mailing Lists*

Phonology and Fonts

- *Phon and PhonBank*
- Unicode and IPA for *Mac*
- Unicode and IPA for *Windows*

Teaching

- *Topics in Language Acquisition*
- *Teaching Resources*
- *YouTube Examples*
- *Bibliographies*

Morphology and Lexicon

- *Part of Speech Analysis by MOR*
- *MOR/POST/MEGRASP Manual*
- *MRC lexical dictionary*

- Find and download existing corpora here

- View and download updated manuals here

- Download CLAN here

https://childes.talkbank.org/

2. General principles and procedures

- Header lines
- Main body
- Checking syntax and lexicon

(1). Header lines

A CLAN transcript

- **File headers**

lines of text about the participants and the setting. All headers begin with the “@” sign.

关于录像、参与者以及转写文档的背景信息

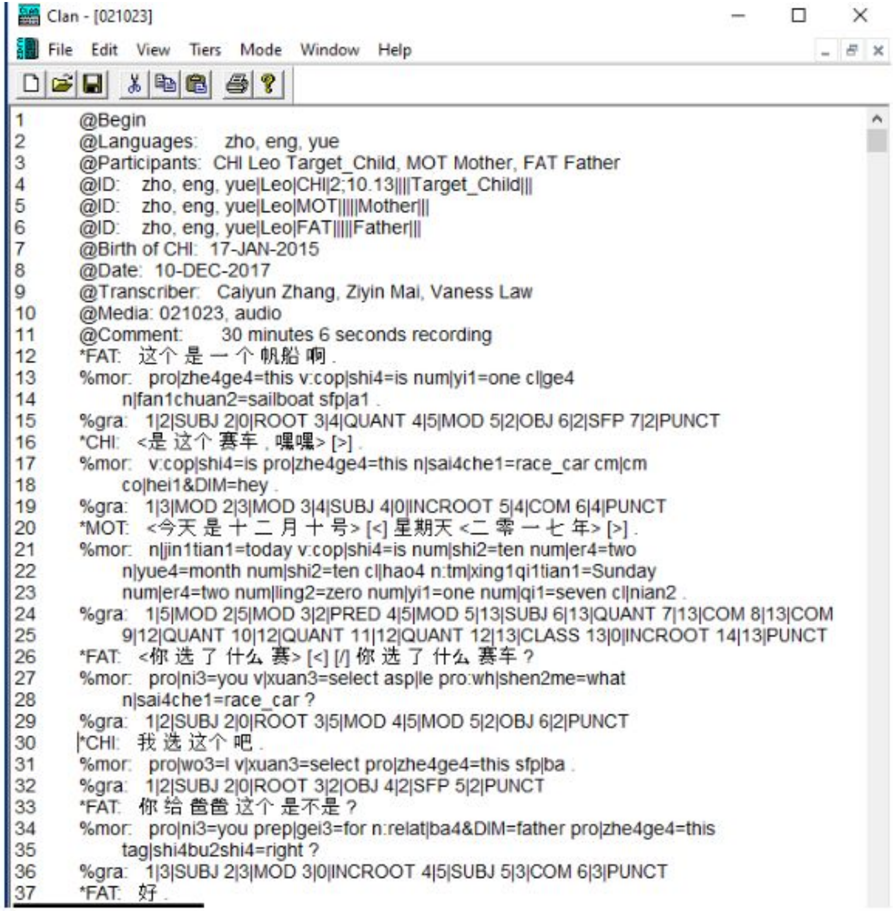
- **Body**

main tiers (lines beginning with *)

dependent tiers (lines beginning with %)

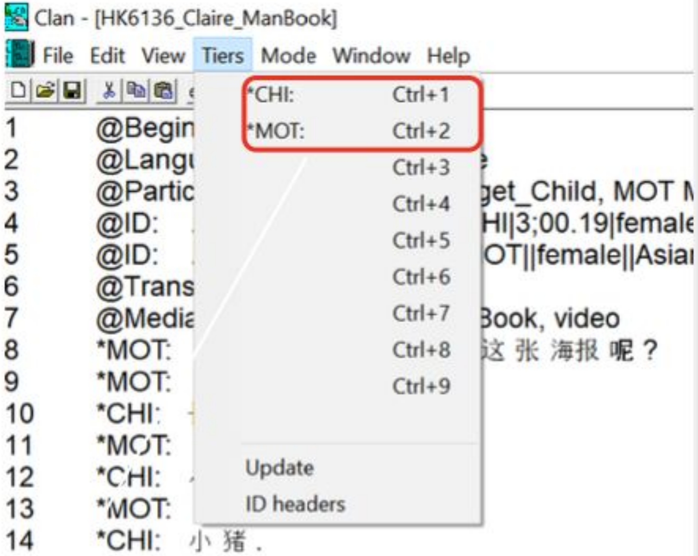
ends with @End.

语言数据



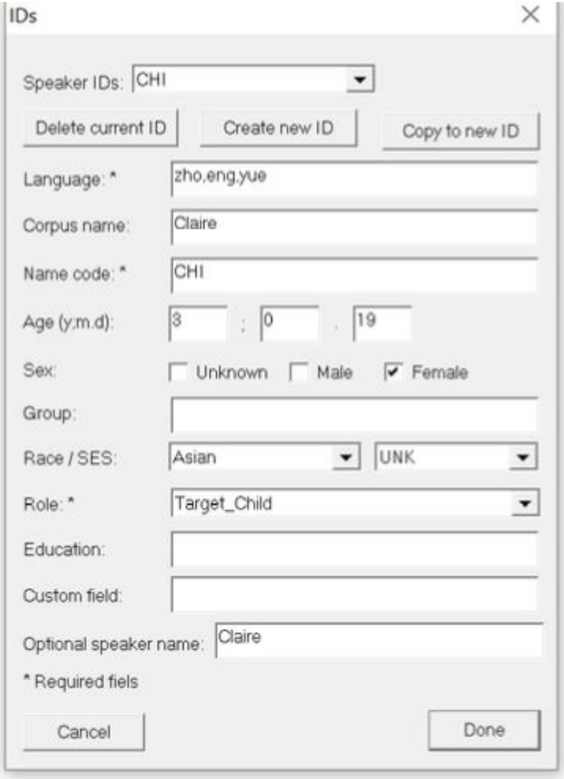
```
@Begin
@Languages:    zho, eng, yue
@Participants:  CHI Leo Target_Child, MOT Mother, FAT Father
@ID:   zho, eng, yue|Leo|CHI|2;10.13|||Target_Child|||
@ID:   zho, eng, yue|Leo|MOT|||||Mother|||
@ID:   zho, eng, yue|Leo|FAT|||||Father|||
@Birth of CHI:  17-JAN-2015
@Date:  10-DEC-2017
@Transcriber:   Caiyun Zhang, Ziyin Mai, Vaness Law
@Media: 021023, audio
@Comment:      30 minutes 6 seconds recording
*FAT:   这个 是 一 个 帆船 啊 .
%mor:   pro|zhe4ge4=this v:cop|shi4=is num|yi1=one cl|ge4
        n|fan1chuan2=sailboat sfp|a1 .
%gra:   1|2|SUBJ 2|0|ROOT 3|4|QUANT 4|5|MOD 5|2|OBJ 6|2|SFP 7|2|PUNCT
*CHI:   <是 这个 赛车 , 嘿嘿> [>] .
%mor:   v:cop|shi4=is pro|zhe4ge4=this n|sai4che1=race_car cm|cm
        co|hei1&DIM=hey .
%gra:   1|3|MOD 2|3|MOD 3|4|SUBJ 4|0|INCROOT 5|4|COM 6|4|PUNCT
*MOT:   <今天 是 十 二 月 十 号> [<] 星期天 <二 零 一 七 年> [>] .
%mor:   n|jin1tian1=today v:cop|shi4=is num|shi2=ten num|er4=two
        n|yue4=month num|shi2=ten cl|hao4 n:tm|xing1qi1tian1=Sunday
        num|er4=two num|ling2=zero num|yi1=one num|qi1=seven cl|nian2 .
%gra:   1|5|MOD 2|5|MOD 3|2|PRED 4|5|MOD 5|13|SUBJ 6|13|QUANT 7|13|COM 8|13|COM
        9|12|QUANT 10|12|QUANT 11|12|QUANT 12|13|CLASS 13|0|INCROOT 14|13|PUNCT
*FAT:   <你 选 了 什么 赛> [<] [/] 你 选 了 什么 赛车 ?
%mor:   pro|ni3=you v|xuan3=select asp|le pro:wh|shen2me=what
        n|sai4che1=race_car ?
%gra:   1|2|SUBJ 2|0|ROOT 3|5|MOD 4|5|MOD 5|2|OBJ 6|2|PUNCT
*CHI:   我 选 这个 吧 .
%mor:   pro|wo3=I v|xuan3=select pro|zhe4ge4=this sfp|ba .
%gra:   1|2|SUBJ 2|0|ROOT 3|2|OBJ 4|2|SFP 5|2|PUNCT
*FAT:   你 给 爸爸 这个 是不是 ?
%mor:   pro|ni3=you prep|gei3=for n:relat|ba4&DIM=father pro|zhe4ge4=this
        tag|shi4bu2shi4=right ?
%gra:   1|3|SUBJ 2|3|MOD 3|0|INCROOT 4|5|SUBJ 5|3|COM 6|3|PUNCT
*FAT:   好 .
```

Creating ID headers



```
1   @Begin
2   @Langu                        e
3   @Partic                  get_Child, MOT M
4   @ID:                     HI|3;00.19|female
5   @ID:                     OT||female||Asian
6   @Trans
7   @Media                   Book, video
8   *MOT:                    这 张 海报 呢 ?
9   *MOT:
10  *CHI:
11  *MOT:
12  *CHI:
13  *MOT:
14  *CHI:   小 猪 .
```

These become shortcuts for later data entry



Speaker IDs: CHI

Language: * zho,eng,yue

Corpus name: Claire

Name code: * CHI

Age (y;m.d): 3 ; 0 . 19

Sex: □ Unknown □ Male ☑ Female

Group:

Race / SES: Asian UNK

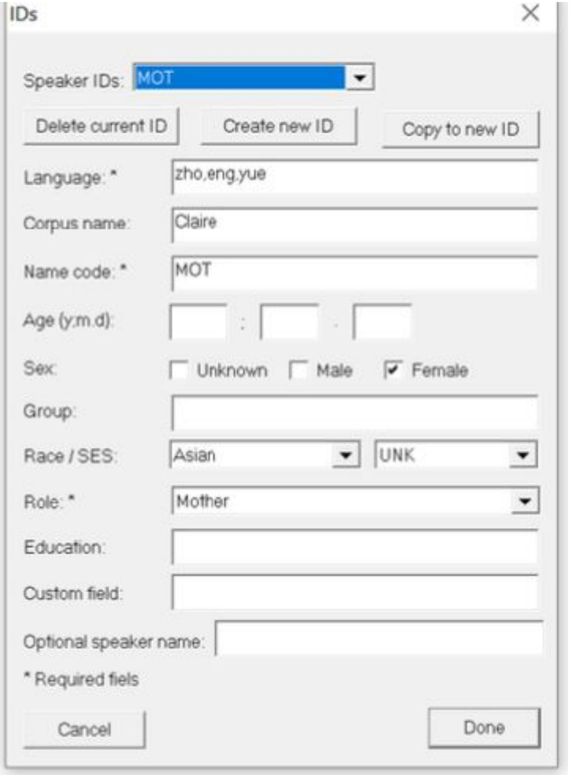
Role: * Target_Child

Education:

Custom field:

Optional speaker name: Claire

* Required fiels



Speaker IDs: MOT

Language: * zho,eng,yue

Corpus name: Claire

Name code: * MOT

Age (y;m.d): ; .

Sex: □ Unknown □ Male ☑ Female

Group:

Race / SES: Asian UNK

Role: * Mother

Education:

Custom field:

Optional speaker name:

* Required fiels

(2). Main body

Words, utterances and non-verbal communications between the target participants

Segmenting words based on MOR grammar

Some examples of “arbitrary” segmentation in Mandarin:

- Single words:

打字 看到 好朋友 小张阿姨 好啦

- Separate words:

写 字 见 到 我 们 好 呀

- N.B. Some lexical strings are treated differently in different context:

是不是/好不好/行不行

*MOT: 一会儿 再 看 , 好不好(okay)？

*MOT: 刘老师 好(good) 不 好？

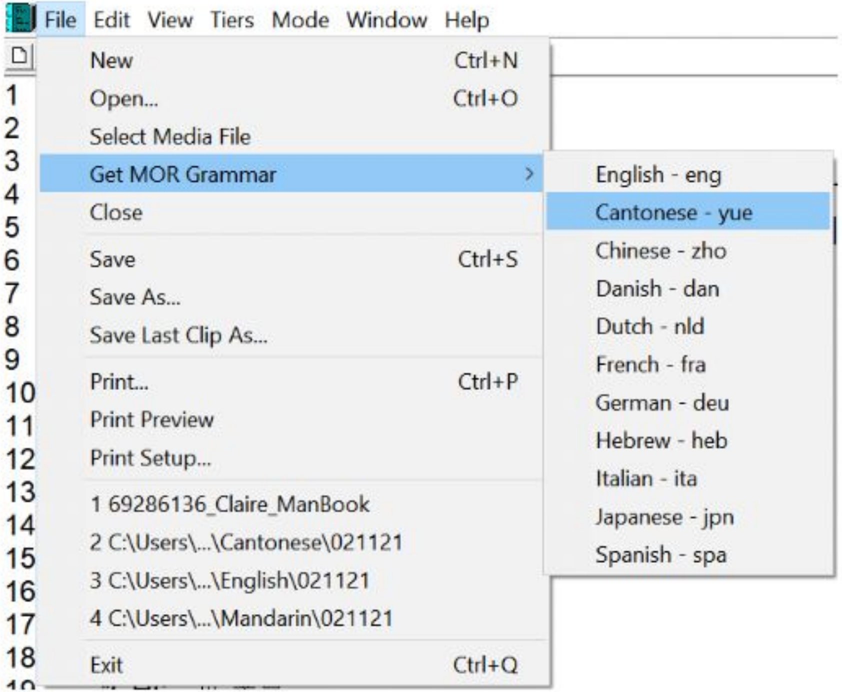
Downloading MOR grammar

CLAN > File > Get MOR Grammar

Select English, Cantonese and Chinese

Files will be downloaded and unzipped automatically (usually to a folder called MOR on the **desktop** in your computer).

You may want to cut the MOR folder and paste it to the default TalkBank folder in your computer (e.g., This PC > windows (C:) > TalkBank > CLAN > Lib > MOR).



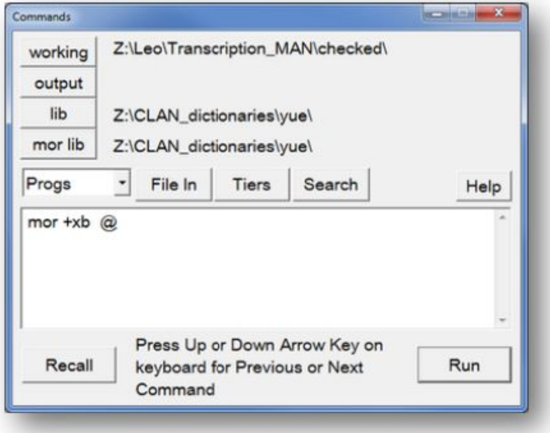
Run “mor +xb @” to check lexicon

WINDOW → COMMANDS

- Choose the yue dictionary to check Cantonese words
- Type the command **“mor +xb”**
- Click **FILE IN** to choose the file you want to check
- A new clan file will be generated to show a list of undefined words
- Correct your transcript based on the feedback (typos and mis-segmented words)

Repeat the same using **zho** and **eng** dictionaries to check Mandarin and English lexicons

If you want to go ahead and tag the transcript. Feel free to run **“mor *.cha”** in commands. You will find a new transcript with %mor and %gra in it. Refer to MOR manual for details.



working Z:\Leo\Transcription_MAN\checked\
output
lib Z:\CLAN_dictionaries\yue\
mor lib Z:\CLAN_dictionaries\yue\

```
mor +xb @
```

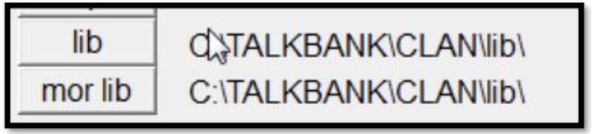
Press Up or Down Arrow Key on keyboard for Previous or Next Command

MOR commends

Run “CHECK” to check syntax

- **MODE → Check Opened file** or: ESC + L
- Run CHECK from the top of the transcript to the end
- Correct everything.

- If you see ‘can’t open file ‘ISO-639’/’depfile’, locate the library directory in the command windows to the correct location:



lib	C:\TALKBANK\CLAN\lib\
mor lib	C:\TALKBANK\CLAN\lib\

- Continue the process until you see "Success! No Errors Found“

Delimitating utterances

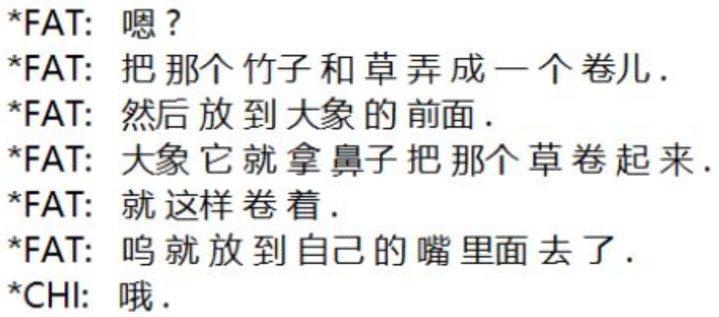
Typical utterance delimitators:

1. Silence or pause longer than 2 seconds (except for lexical retrieval)
2. Utterance-ending prosody (rising or falling)
3. Grammatically complete sentences

WHICHEVER IS SHORTER!

Example: *take out the plate / then put the egg on it / then cook the egg (3 utterances)*
拿出盘子 / 把鸡蛋放里边 / 然后就开始煮 (3个语句)

*CHI: I do it every +…
*CHI: not right .
*CHI: you should put yellow here .



*FAT: 嗯 ?
*FAT: 把 那个 竹子 和 草 弄 成 一 个 卷儿 .
*FAT: 然后 放 到 大象 的 前面 .
*FAT: 大象 它 就 拿 鼻子 把 那个 草 卷 起 来 .
*FAT: 就 这样 卷 着 .
*FAT: 呜 就 放 到 自己 的 嘴 里面 去 了 .
*CHI: 哦 .

Formatting utterances

- Only one utterance on each main line
- End the utterance with one of the three “terminators”
  - Period .
  - Question mark ?
  - Exclamation mark !

Non-verbal communications

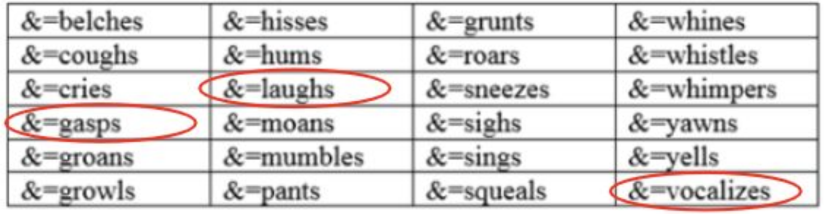
1. Paralinguistic materials

   *MOT: [=! laughs] = happens with in an utterance

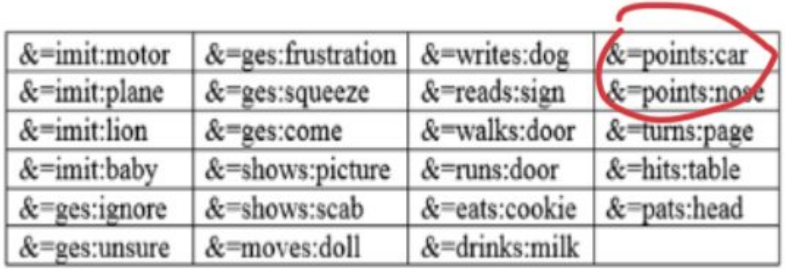
   *MOT: &=laughs . = happens between utterances

Non-verbal communications

N. B. Only one single lexical string is allowed after "&=“. Here are some examples.



&=belches	&=hisses	&=grunts	&=whines
&=coughs	&=hums	&=roars	&=whistles
&=cries	&=laughs	&=sneezes	&=whimpers
&=gasps	&=moans	&=sighs	&=yawns
&=groans	&=mumbles	&=sings	&=yells
&=growls	&=pants	&=squeals	&=vocalizes



&=imit:motor	&=ges:frustration	&=writes:dog	&=points:car
&=imit:plane	&=ges:squeeze	&=reads:sign	&=points:nose
&=imit:lion	&=ges:come	&=walks:door	&=turns:page
&=imit:baby	&=shows:picture	&=runs:door	&=hits:table
&=ges:ignore	&=shows:scab	&=eats:cookie	&=pats:head
&=ges:unsure	&=moves:doll	&=drinks:milk	

&=nods.

&=head:yes	&=hands:no	&=mouth:open
&=head:no	&=hands:hello	&=mouth:close
&=head:shake		

- &=imit: the noise source being imitated vocally
- &=ges: meaning of the gestures being used
- &=walks:door, &=runs:door: the direction or goal of the walking or running
- &=slurps and &=eats: the sounds of slurping or eating

Non-verbal communications

2. Explanation tiers (%exp) 另行解释

*FAT: 这边 再 亲 一 个 .

%exp: the child kissed the other side of his father’s face .

- Whenever possible, try to use “&=text” or “[=! text] as a substitute for writing longer comments using dependent tiers.

*FAT: 这边 再 亲 一 个 .

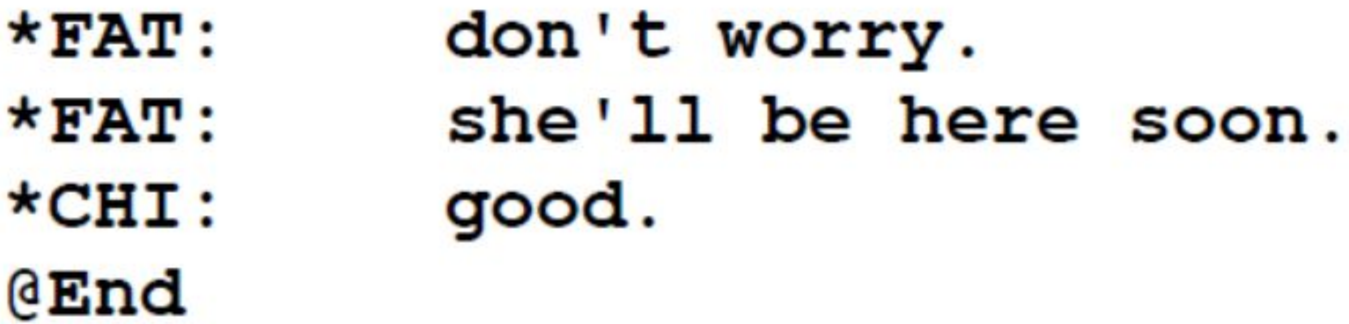
*CHI: &=kisses:face .

%exp: the child kissed the other side of his father’s face .

Ending the transcript

The last line in the file must be an @End header line.

转写结束，另起一行用@End标注。这一行必须是全文最后一行。



```
*FAT:	don't worry.
*FAT:	she'll be here soon.
*CHI:	good.
@End
```

Symbols and rules

How do you handle capitalization in English?

Do not capitalize the first letter of the first word in a line.

Capitalize **proper nouns** & the **pronoun “I”** only.

Except for the first letter of proper nouns and the first-person pronoun "I", the first letter of each line does not need to be capitalized.

Common names are not capitalized. For example, t_v, v_c_r

Example:

*CHI: you and I love Barney , Pooh_Pooh and Buzz_Lightyear .

How are proper nouns segmented?

In English: Connect the words with underscore: _ 英文中，以短横线连接专有名词:

In Chinese: No spaces between characters

中文里，专有名词(如人名、人物称呼、角色名、书名等)各个字符之间不 空格

*CHI: <爸爸, 我 要 唱> [<] 摇一摇 这 个 歌.

*CHI: 经典 童话 丑小鸭.

*CHI: 陈老师 教 我 们 的.

*CHI: 还 有 太奶奶.

What if you're pronouncing single letters or spelling out words?

@l or spelling sequence: @k

*CHI: a@l.

*CHI: m@l a@l 马.

*CHI: apple@k.

How should numbers and titles be handled?

No Arabic numerals in transcripts.

Transcribe numbers as it is pronounced. For example, “two hundred fifty six” for 256; “twenty nine point five percent” for 29.5%. Do not include hyphens or underscores.

中文数字使用汉字表示:

二零二一年五月十八日

二零二一年五月十八號

How to transcribe reduplicated words?

Transcribe the words as they are

Examples:

*MOT: 你 睇睇 先啦.

*CHI: no@s:eng, 我 要 車車.

*MOT: 哩樣嘢 黃黃 哋色嘅?

What if the speaker omits certain syllables?

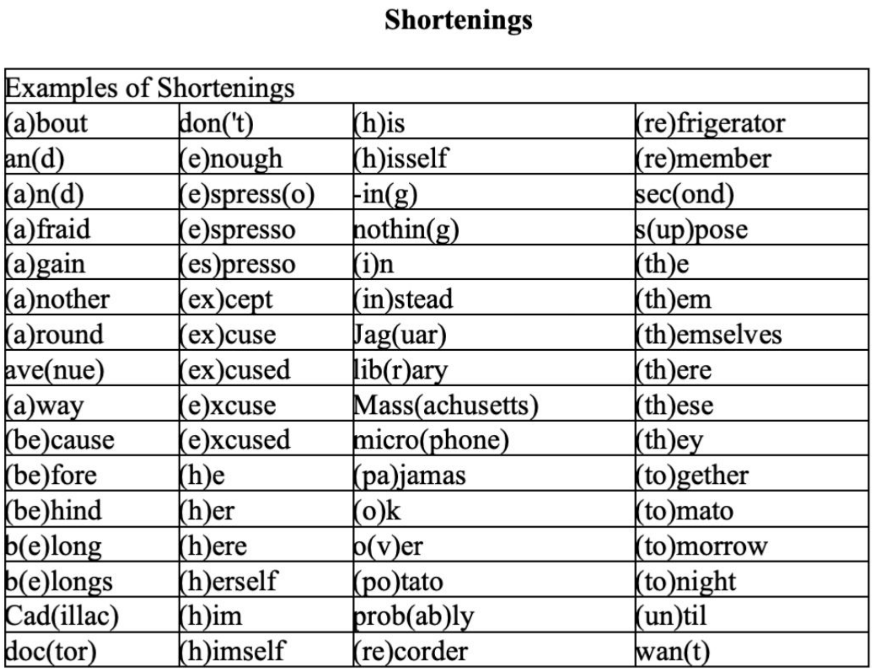
(xxx)

Examples:

*CHI: [- eng] (o)kay.

*CHI: 跟(住) +...

*CHI: 东(西) [/] 东西.



Shortenings

Examples of Shortenings			
(a)bout	don('t)	(h)is	(re)frigerator
an(d)	(e)nough	(h)isself	(re)member
(a)n(d)	(e)spress(o)	-in(g)	sec(ond)
(a)fraid	(e)spresso	nothin(g)	s(up)pose
(a)gain	(es)presso	(i)n	(th)e
(a)nother	(ex)cept	(in)stead	(th)em
(a)round	(ex)cuse	Jag(uar)	(th)emselves
ave(nue)	(ex)cused	lib(r)ary	(th)ere
(a)way	(e)xcuse	Mass(achusetts)	(th)ese
(be)cause	(e)xcused	micro(phone)	(th)ey
(be)fore	(h)e	(pa)jamas	(to)gether
(be)hind	(h)er	(o)k	(to)mato
b(e)long	(h)ere	o(v)er	(to)morrow
b(e)longs	(h)erself	(po)tato	(to)night
Cad(illac)	(h)im	prob(ab)ly	(un)til
doc(tor)	(h)imself	(re)corder	wan(t)

How should singing or making meaningless sounds be marked?

1. Phonological fragments
   - &+
2. Filler
   - &-
     - Eg. *CHI: &- 嗯 .
3. Nonwords
   - &~

If I really can't hear it, what should I do?

Use “xxx” if you have absolutely no idea what was said.

If you can't hear or guess what the speaker is saying at all, you can use "xxx".

Regardless of how long the unheard content is, use "xxx" to indicate it.

What should you do when you're unsure if what you've heard is accurate?

- When you are unsure about part of the transcribed content, you need to mark the guess with the symbol "[?]".

How to use punctuation marks?

Punctuation marks: “,” “.” “?” “!”

- All of the symbols used in CLAN must be English (rather than Chinese) punctuation marks

All punctuation marks are in English.

- “,” commas should only be used within an utterance “,”Commas can only be used in the middle of a sentence, not at the end.
- “.” “?” “!” punctuation marks like the period, the question mark, and the exclamation mark should only appear at the end of an utterance.
- “.” “?” “!” Periods, question marks, and exclamation marks are only used at the end of a sentence.

What if two people are saying the same thing?

For declarative sentence:

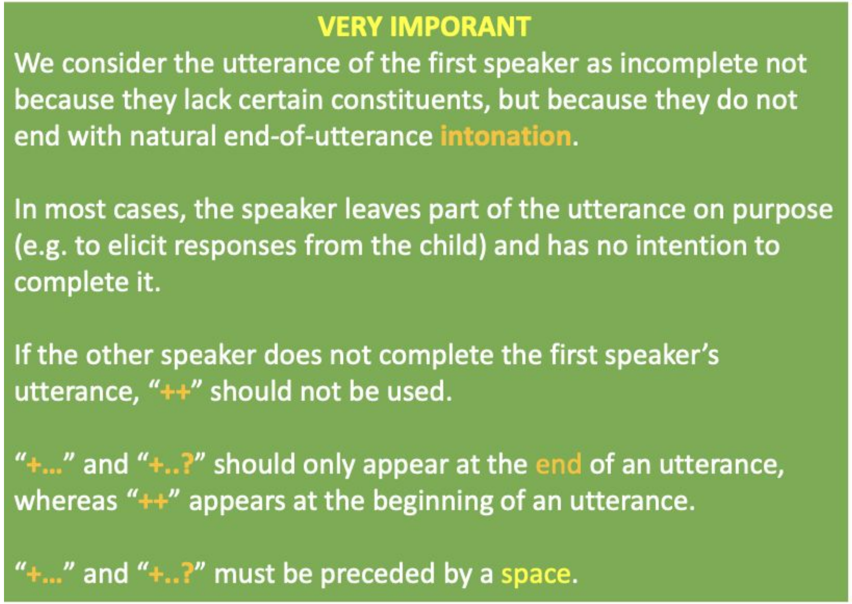
*MOT: 你看, 小 狗狗 [: 狗] 正在 追 +...

*CHI: ++ 小 兔子 .

For questions:

*CHI: <係咪 你> [/] 係咪 你 +..? *

MOT: ++ 食 苹果 ？



VERY IMPORANT

We consider the utterance of the first speaker as incomplete not because they lack certain constituents, but because they do not end with natural end-of-utterance intonation.

In most cases, the speaker leaves part of the utterance on purpose (e.g. to elicit responses from the child) and has no intention to complete it.

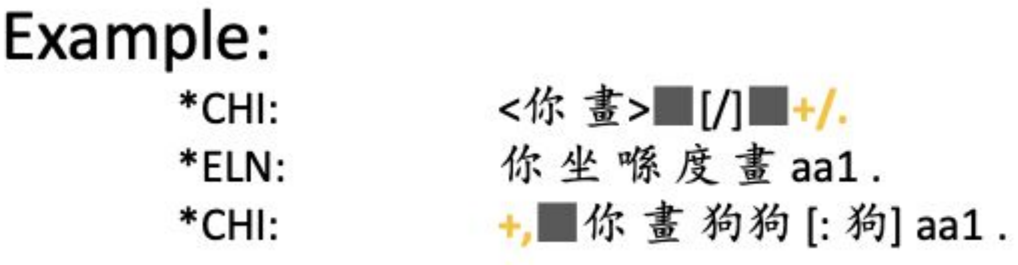
If the other speaker does not complete the first speaker's utterance, "++" should not be used.

"+..." and "+..?" should only appear at the end of an utterance, whereas "++" appears at the beginning of an utterance.

"+..." and "+..?" must be preceded by a space.

What should you do if you are interrupted by another speaker?

- Use “+/.” to end the interrupted utterance, and “+,” to begin the utterance that completes the interrupted utterance by the SAME speaker, if any. If the utterance being interrupted has the shape of a question, use “+/?” .
- If the interrupted speaker drops the topic and start another utterance, do not mark it with “+,”.



Example:

```
*CHI:        <你 畫>■[/]■+/.
*ELN:        你 坐 喺 度 畫 aa1 .
*CHI:        +,■你 畫 狗狗 [: 狗] aa1 .
```

How should one mark a speaker's direct quote or content from a book?

- In our project, do not mark quotes in the adult input, because whether an utterance is a quote or not, it is part of the input available to the children.
- However, for quotes in the children’s utterances, use +" to begin the quote, e.g.
    - *CHI: +"■锄禾日当午.

How to correctly use [] and <>?

- Put functional symbols inside square brackets [] []
    - Used for functional symbols
- Put the speech content and utterances inside angle brackets <> <>
    - Used to explain which segments of language in a sentence the functional symbols affect.

How to deal with stuttering or repetitive speech?

- Repeated only once
    - *CHI: what■[/]■what is that ?
- Repeated twice
    - *CHI: <what what>■[/]■what is that ?
- Repeated ≥ 3 times:
    - List all, or use“[x■n]” • (Note: "n" represents the total number of times a word is spoken, not the number of repetitions.)

*CHI: <what what what>■[/]■what is that ? *CHI: what■[x■4]■is that ?

How to deal with the phenomenon of self-correction (repeating) while speaking?

Transcribe all that was said, then use the■[//]■where you see fit.

Transcribe the entire content accurately. When the speaker returns to repeat a part that is more than one word, mark it with "<>" followed by the repeated content.

Example:

*MOT: where is it ?

*CHI: the tortoise■[//]■turtle is in the ocean .

*MOT: <what does it>■[//]■what does the turtle eat ?

What should I do if the speaker needs to be counted or listed?

Counting

Within 10: write down each number in separate lines;

More than 10: use “ %exp: ”.

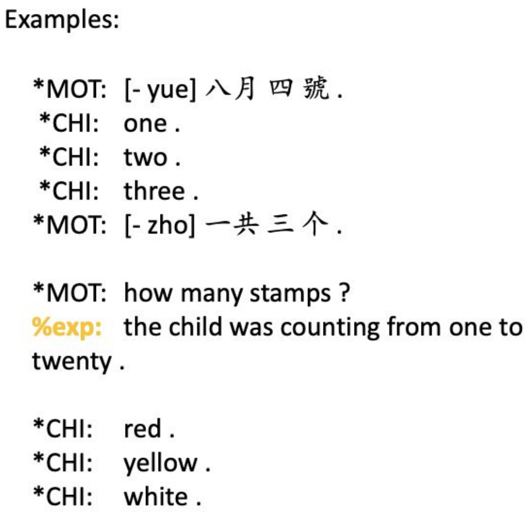
数数时:

10以内的数数: 每一个数字都单独算一个语句;

超过10的数数: 用“%exp:”解释从几数到几，不需要全部单行列出。

Enumerating

Write down each words in separate lines.



Examples:

```
*MOT:  [- yue] 八月 四 號 .
*CHI:  one .
*CHI:  two .
*CHI:  three .
*MOT:  [- zho] 一共 三 个 .

*MOT:  how many stamps ?
%exp:  the child was counting from one to
twenty .

*CHI:  red .
*CHI:  yellow .
*CHI:  white .
```

What should you do if there is a long pause in a complete sentence?

Example:

*CHI: 冇 得 食 (.) chocolate@s 呀 .

(.)

What should you do if there are errors in the speaker's sentence?

- Do not correct the speaker's mistakes.
- Transcribe accurately.
- Only use [] to mark the correct form when the incorrect form cannot be recognized by MOR.

What if speakers are speaking simultaneously, or their speech partially overlaps?

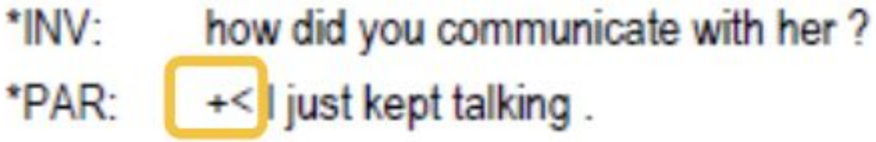
- “+<”

Example:

*MOT: 橙 , 係 喎 , 橙 .

*ALI: +<■係 喎 , 橙 嚟 㗎 喎 .

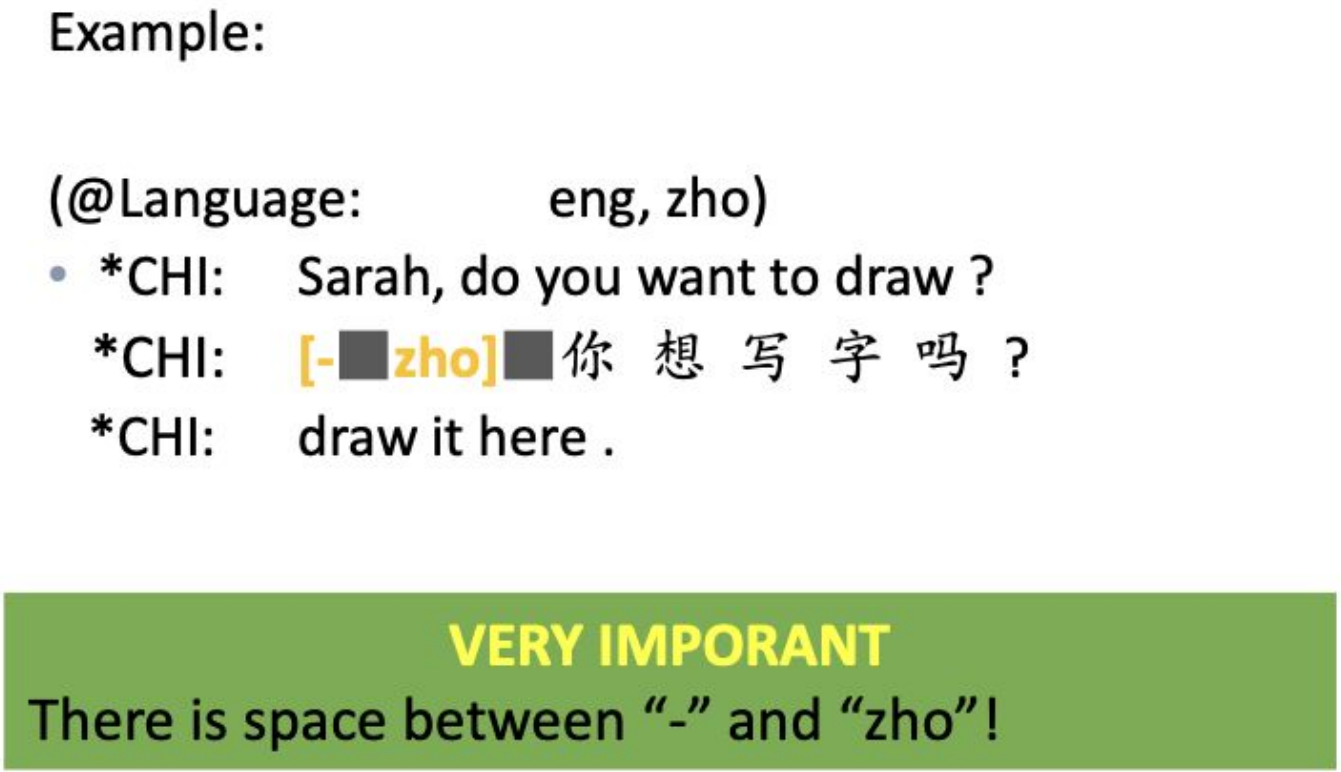
Example:



*INV: how did you communicate with her ?

*PAR: +< I just kept talking .

What if the speaker switches languages?



Example:

(@Language: eng, zho)

- *CHI: Sarah, do you want to draw ?

*CHI: [-■zho]■你 想 写 字 吗 ?

*CHI: draw it here .

VERY IMPORANT

There is space between “-” and “zho”!

How to record actions and events in a video?

- Explanation tier: **%exp:**

- Example:

```
*CHI:     give me !
%exp:     the mother gave the toy car to the child .
```

Appendix 1

- CHILDES: https://childes.talkbank.org/
- CHAT transcription manual: https://talkbank.org/manuals/CHAT.pdf
- CLAN analysis manual: https://talkbank.org/manuals/CLAN.pdf
- MOR analysis manual: https://talkbank.org/manuals/MOR.pdf
- MOR grammars: https://talkbank.org/morgrams/
- Guide to CHAT and CLAN written in Chinese: https://talkbank.org/manuals/Clin-CLAN-zho.pdf