[><] UVALAL
Language
Acquisition
Laboratory
University of Valladolid

Universidad de Valladolid

https://uvalal.uva.es/

October 1, 2014
Revised: July 11, 2022

TRANSCRIBING IN CHAT

(soraUVALAL corpus)

INDEX

1. INTRODUCTION

- SLABank database (MacWhinney 2019): the language component of the TalkBank system (for sharing and studying conversational interactions) https://www.talkbank.org/ [CHAT & CLAN, updated frequently!]

- Basic requirement: download **CLAN** at https://dali.talkbank.org/clan/
    - Watch the tutorial at http://talkbank.org/screencasts/install.mp4

- Main tools:
    - **CHAT** (*Codes for the Human Analysis of Transcripts*)
        - transcription system [http://talkbank.org/manuals/CHAT.pdf]
    - **CLAN** (*Computerized Language ANalysis*)
        - codification & analysis system [http://talkbank.org/manuals/CLAN.pdf]
    - **MOR** (*MORphosyntactic analysis*)
        - morphosyntactic analysis [http://talkbank.org/manuals/MOR.pdf]

2. HOW A CHAT TRANSCRIPTION LOOKS LIKE

<table>
<tr><th>Example 1(a)</th><th>Example 1(b)</th></tr>
<tr><td>@Begin
@Languages: eng
@Participants: CHI Naomi Target_Child, FAT Father, MOT Mother
@ID: eng|sachs|CHI|1;10.20||||Target_Child||
@ID: eng|sachs|FAT|||||Father||
@ID: eng|sachs|MOT|||||Mother||
@Tape Location: Tape 14
*CHI: empty .
%mor: adj|empty .
%gra: 1|0|ROOT 2|1|PUNCT
*CHI: empty .
%mor: adj|empty .
%gra: 1|0|ROOT 2|1|PUNCT
*CHI: hot bottle .
%mor: adj|hot n|bottle .
%gra: 1|2|MOD 2|0|ROOT 3|2|PUNCT
*MOT: empty hot bottle .
%mor: adj|empty adj|hot n|bottle .
%gra: 1|3|MOD 2|3|MOD 3|0|ROOT 4|3|PUNCT
*CHI: what's this ?
%mor: pro:wh|what~v:cop|be&3S pro:dem|this ?
(…)
@End</td><td>@Begin
@Languages: bos, eng
@Participants: CHI BLBO4.22 Target_Child, SON Sonja Investigator
@ID: bos,eng|soraUVALAL|CHI|13;00.|male|||Target_Child|||
@ID: bos, eng |soraUVALAL|SON|||||Investigator|||
@L1 of CHI: Bosnian
@Date: 21-OCT-2014
@Time Duration: 00:08:17
@Location: Banja Luka, Bosnia and Hercegovina
@Situation: semiguided interview
@Transcriber: Sonja
*SON: so hello www.
*SON: how are you?
*CHI: I'm good.
*CHI: how are you?
*SON: I'm good.
*SON: what have you been doing today?
*CHI: puff my friends.
*CHI: I was study.
(…)
@End</td></tr>
<tr><td><b>Sachs (NA EN, Monolingual)</b></td><td><b>soraUVALAL (L2 EN, Bilingual)</b></td></tr>
</table>

3. THE CHAT TRANSCRIPTION SYSTEM

- CHAT manual [updated frequently!]: http://talkbank.org/manuals/CHAT.pdf
  - Tutorial http://talkbank.org/screencasts/template.mp4

- The components:
  - Headers
  - Tiers
  - Codes

3.1. HEADERS

Headers provide metalinguistic information about the transcript file, and they are always preceded by an @ sign, and followed by a colon and a tab. Some of them are **obligatory** (in bold) while some other are optional.

For the soraUVALAL corpus, see example 2(b).

Example 2(a): spontaneous data

```
@Begin
@Languages:    eng , spa
@Participants:  CHI1 Leo Target_Child , CHI2 Simon Target_Child , MEL Melanie Mother , EST
Esther Investigator , IVO Ivo Father
@ID:       eng , spa|FerFuLice|CHI1|3;01.20|male|||Target_Child|||
@ID:       eng , spa|FerFuLice|CHI2|3;01.20|male|||Target_Child|||
@ID:       eng , spa|FerFuLice|MEL|||||Mother|||
@ID:       eng , spa|FerFuLice|EST|||||Investigator|||
@ID:       eng , spa|FerFuLice|IVO|||||Father|||
@Date:     03-FEB-2002
@Time Duration: 00:00:00-00:30:04
@Location:        Salamanca , Spain
@Situation:       Playing at home
@Comment:         First fragment of Session 30
(…)
@End
```

Example 2(b): experimental data

```
@Begin
@Languages:    bos , eng
@Participants:   CHI BLBOV2.01 Target_Child , SON Sonja Investigator
@ID:       bos , eng |soraUVALAL|CHI1|10;00.|male|||Target_Child|||
@ID:       bos , eng |soraUVALAL|SON|||||Investigator|||
@Date:     20-OCT-2014
@Time Duration: 00:13:44
@Location:        Banja Luka, Bosnia and Hercegovina
@Situation:       semiguided interview
@Transcriber:     Sonja
(…)
@End
```

3.1.1. Obligatory headers [See *Chapter 5. minCHAT*, pp. 22-23; 29-… in the CHAT manual]

- The **@Begin** and **@End** headers indicate the beginning and end of the CHAT file respectively.

- The **@Languages** header indicates which language the participants speak in a given session. A three-letter ISO 639-3 code must be used (p. 31 in the CHAT manual or check the tutorial (4.35) http://talkbank.org/screencasts/template.mp4)

- The **@Participants** header contains all the participants for a given session and each speaker is assigned a 3-letter speaker ID based on a name (e.g., SON for Sonja) and a role in the setting (e.g., Investigator)

- The **@ID** header is required for each participant, and you can add each ID either manually (http://talkbank.org/screencasts/addID.mp4) or automatically (http://talkbank.org/screencasts/autoinsertID.mp4)

Example 3

@ID: language|corpus|code|age|sex|group|SES|role|education|
@ID: dan , eng|soraUVALAL|CHI1|3;01.20|male|||Target_Child||

Note: SES refers to Social Economic Status (UC= upper class; WC= working class; MC= middleclass) or Ethnicity (White, Black, Latino, Asian, Pacific).

If not all of this information is encoded, then some pipe characters (|) may appear together, meaning that some of these fields have been left empty.

3.1.2. Optional headers

The other headers in examples in 4 are optional and they may include information about the name of the file, comments about the recording situation, and more. The headers in example 4(a) are the ones used for the soraUVALAL corpus.

Example 4(a)

@Date:	14-OCT-2014	date of the recording
@Time Duration:	00:10:04	duration of the interaction
@Location:	Banja Luka, Bosnia	where the interaction took place
@Situation:	semiguided interview	brief description of the setting
@Transcriber:	Sonja	person who transcribed the data

Other optional headers include the following:

Example 4(b)

@Birth of CHI:
@Media:

After including all the headers, you should check for **errors** in the transcription: **press Escape** [release] **and press "L"**. Errors will be notified in the CHAT document.

3.2. TIERS

There are 2 types of tiers, independent tiers or main lines, and dependent or secondary tiers.

3.2.1. Independent tiers

Independent tiers indicate who is speaking at a given moment and they always contain an asterisk (*), three capital letters corresponding to the participant code assigned in the @Participants header, a colon and a tab.

***CHI**: (…) .
***SON**: (…) .

- Each utterance in the independent tiers must end in an **utterance terminator** (not necessarily preceded by a space): full stop (.), question (?) or exclamation mark (!)

  Example 5

  *SON: why ?
  *CHI: because you run way more indoor than outdoor .

- Each utterance begins with **small caps**, except if they start with the 1st person pronoun "**I**" or a proper name.

- **Proper names**: only the names of the investigators may be transcribed in full. As to the other participants, only the first initial is transcribed if pseudonyms are used.

  Example 6

  *SON: where do you live?
  *CHI: in B.

3.2.2. Dependent tiers

Dependent tiers provide supplementary information about the preceding utterance, and they always contain a percentage sign (%) followed by 3 lowercase letters (e.g., com, pho, mor, etc.) indicating the type of additional information provided about the preceding utterance, then a colon and then a tab.

Example 7(a)

%com: somebody interrupted the interview

Example 7(b) and example 1(a) above

%mor: morphological information
%pho: phonetic information
%gram: grammatical information

No utterance terminators are required at the end of dependent tiers.

For the soraUVALAL corpus, the only dependent tier that is used is %com.

4. CODES IN THE INDEPENDENT TIERS

Some basic types of codes are used in independent tiers in order to indicate a variety of form markers, e.g. a formulaic use of words, unidentifiable material, incomplete or omitted words, tone markers, and language switches [Other markers at http://talkbank.org/manuals/CHAT.pdf].

4.1. Codes in angle < > and square [ ] brackets

Repetition: [/]

Example 8(a)

*CHI: and it was [/] was not saving because it's just froze .

If a word is repeated more than once, [x number] is used.

Example 8(b)

*CHI: I like [x 3] (.) Messi .

If the repetition applies to more than one word, use angle brackets < >.

Example 8(c)

*SON: <can you> [/] can you please repeat ?

Retracing: [//]

Example 9

*SON <where do> [//] what did you say you went ?

If the repetition with self-repair applies to more than one word, use angle brackets < >.

Reformulation: [///]

Example 10

*SON: been [///] have you visited France ?

Overlap: Overlap follows [>]Overlap precedes [<]

Text overlapping (one word or more) must be enclosed in angle brackets < >.

Example 11

*CHI: yes <I like> [>] .
*SON: <&eh> [<] sorry .

Replacement: [: text]

Example 12

*CHI: he go [: goes] in third &eh grade.

Do not correct word order or other errors which are not clear, as in examples in 13.

Example 13

Quiero yo esto	(word order)
Go cow	(word order)
He go	(*goes* or *went*?)
La amigas	(*la amiga* or *las amigas*?)
Van papá	(*va papá* or *van papás* or *va con papá*?)

Best guess: [?]

Use with words you are certain to have heard but that do not appear to fit in an utterance.

Example 14

*CHI: because my father has like Arsenal since [?] he was very small.

It does not require angle brackets, unless a whole string of words is stressed.

Alternative transcription: [=? text]

It is used when it is difficult to choose between two possible transcriptions for a word or group of words. Hence both are included.

Example 15

*CHI: we want <one or two> [=? one too] .

4.2. Codes with the + sign

Trailing off terminator: +...

Example 16

*CHI: the one that run on the +...

Interruption: +/.

Example 17

*SON: it's was +/.
*CHI: a book ?

Self-interruption: +//.

Example 18

*SON: I don't think +//.
*SON: maybe you can tell me something .

Interruption of question: +/?

Example 19

*SON: do you like +/?
*CHI: football ?

Self-completion: +.

Example 20

*CHI: I +.
*SON: sorry?
*TOD: +, I like English .

When a segment with +/. is followed by +, on the next utterance MLU program treats it as a single utterance.

Quotations: +"/. +"

The code +"/. indicates that the quotation follows and +" indicates the quoted utterance.

Example 21

*CHI: and then she said +"/.
*CHI: +" I don´t like him anymore.

4.3. Codes with the @ sign

Interjections and filled pauses: @i or &-

The ampersand mark (&) allows this material to be ignored as words, while words with @i will be computed as such.

Example 22(a)

*CHI: in &eh Atlanta .

Example 22(b)

*SON: oh@i that is very nice.

Onomatopoeia: @o

Example 23

*CHI: bang@o bang@o .

4.4. Other codes

Unintelligible speech: xxx

It is used for utterances or parts of an utterance that you cannot make out because participants are whispering, talking at the same time, other participants are drowning the target children out, or the sound quality of the recording is bad.

Example 24

*CHI: I do my homework and then I [/] then I go to xxx to play games and +...

Untranscribed material: www

When the participants' real names or very specific places are mentioned, these will be considered untranscribed material so that the participants cannot be identified and can keep their anonymity (see section 5.2).

Example 25

*SON: good morning , www .
%exp: the child´s real name has not been transcribed

Missing sounds or non-completion of a word: () [See also "shortenings", p. 51, CHAT manual]

Parentheses are used with missing sounds and uncompleted words.

Example 26

*CHI: I (a)m good .

Contractions like "gonna" or "gimme" can only be used with participants other than the target child(ren)/participant because they count only as one word for the MLU program calculations. When these contractions appear, use the replacement notation as in 27.

Example 27

*CHI: lemme [: let me]

Pause: (.) (..) (…)

Use (.), (..) or (…) depending on the duration of the pause within a sentence. If within a word, use ^ instead:

Example 28(a)

*CHI: &eh I go to school and (.) I go to home.

Example 28(b)

*TAM: and okey (..) so you like rugby .

Example 28(c)

*SON: you like (…) Messi ?

Lengthened vowel or syllable: **:**

Example 29

*CHI: we do the: (.) kako@s:bos su@s:bos lekcije@s:bos lessons.

5. FINAL CONSIDERATIONS

5.1. Use of the **plus-sign +** instead of the hyphen (to avoid confusion with the suffixation in the %mor line):

Example 30

*CHI: twenty+one (.) twenty+two +...

5.2. **Confidentiality:** If participants other than the investigators have pseudonyms to protect their privacy and their names appear in the recordings, only the first initial or first and second initials are transcribed:

- In the unlikely event that they reveal other confidential information such as their last name, address, phone number or email, that part of the recording must remain untranscribed.

Example 31

*SON: so hello www.
*SON: how are you today?
*CHI: I am very good!
[…]
*SON: do you live close to school?
*CHI: I live in www .

5.4. The **CLAN program CHECK** must be run **to verify** that the transcription contains no mistakes. When running CHECK, the depfile.cut file must be located in the lib directory (on Windows, this should be c:\TalkBank\CLAN\lib). If you run CHECK and errors are detected, you need to solve them and keep on running CHECK until you get no error messages.