

## JPN 730: CLAN Workshop

### I. OVERVIEW

#### 1. Terminology

- **CHILDES** (Child Language Data Exchange System): The name of the project.
- **CHAT** (Codes for the Human Analysis of Transcripts): The format for transcribing speech. You need to use the CHAT format in order to analyze your transcripts with CLAN.
- **CLAN** (Computerized Language Analysis): The program for analyzing transcripts written in the CHAT format.

#### 2. CHILDES website <<http://childes.psy.cmu.edu>>

- You can download the CLAN program for free.
  - Windows: CLANWinU
  - Mac: CLANXu
  - You need a unicode font installed on your computer. Officially recommended fonts are:
    - ▶ Arial Unicode MS (not just Arial): Included in MS Office Professional, Windows Vista, and Mac OS X 10.5 (Leopard).
    - ▶ Charis SIL: Downloadable from <[http://scripts.sil.org/cms/scripts/page.php?site\\_id=nrsi&id=CharisSIL\\_download](http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&id=CharisSIL_download)>.
- Data (Link “Database”)
  - To download transcripts (compressed in the zip format), go to Database > Downloadable Transcripts.
  - Transcripts from approximately fifty languages, as well as bilingual, clinical, and narrative data, are available.
  - Some transcripts come with audio and video files.
- Japanese data

Database	Sex	Age range	Recording frequency	Dialectal background	Citation information	Note
Hamasaki	M	2;2.3 - 3;4.22	2-3/m	Nagoya	Hamasaki, Naomi. (2002). The Timing Shift of Two-Year-Olds' Responses to Caretakers' Yes/No Questions. In: Shirai, Y., Kobayashi, H., Miyata, S., Nakamura, K., Ogura, T. & Sirai, H. (Eds.). Studies in Language Sciences (2) - Papers from the Second Annual Conference of the Japanese Society for Language Sciences. p.193-206.	
Ishii	M	0;6.1 - 3;8.16	4/m	Kyoto	Ishii, Takeo 1999, The JUN Corpus, unpublished.	Audio and video available
Miyata-Aki	M	1;5.7 - 3;0	4/m	Nagoya	Miyata, S. (1995). The Aki corpus — Longitudinal speech data of a Japanese boy aged 1.6-2.12 -, Bulletin of Aichi Shukutoku Junior College, 34,183–191.	
Miyata-Ryo	M	1;4.3 - 3;0	4/m	Nagoya	Miyata, S. (1992) Wh-Questions of the Third Kind: The Strange Use of Wa-Questions in Japanese Children, Bulletin of Aichi Shukutoku Junior College, 31, 151–155	
Miyata-Tai	M	1;5.20 - 3;1.29	4/m	Nagoya	Miyata, Susanne (2000). The TAI Corpus: Longitudinal Speech Data of a Japanese Boy aged 1;5.20 - 3;1.1 Bulletin of Shukutoku Junior College 39, 77-85.	Audio available
Noji	M	0-7	?	Hiroshima	Noji, Junya. (1973-77). Yooji no gengo seikatsu no jittai I -IV. Bunka Hyoron Shuppan.	Diary study in 1948

- Citation
  - When you present or publish your study that used CHILDES, you are required to cite: MacWhinney, Brian. (2000). The CHILDES project: Tools for analyzing talk. Third Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

- In addition, you need to cite the work that is associated with the corpora that you used (see the above table).
- Info-CHILDES <<http://groups.google.com/group/info-childes>>
  - Info-CHILDES is a mailing list for CHILDES users.
- J-CHAT website <<http://www.cyber.sist.chukyo-u.ac.jp/JCHAT/index-j.html>>
  - Japanese manuals and information about workshops held in Japan are available.
  - No transcripts or programs are available there.

### 3. CHAT format

- Example transcript

```

@Begin
@Languages:   jp
@Participants: CHI Akifumi Target_Child, AMO Okaasan Mother, SUZ Suuze
               Investigator, REE Ree Brother
@ID:          jplMiyata-Aki|CHI|2;3.26|Target_Child|
@ID:          jplMiyata-Aki|AMO|Mother|
@ID:          jplMiyata-Aki|SUZ|Investigator|
@ID:          jplMiyata-Aki|REE|Brother|
@Date:        22-JAN-1990
@Warning:     recorded time: 1:00:00
@Comment:     using chigau also for dame or iya
@Situation:   looking at video camera
*CHI:         nani ,, koko ?
%cod:         $Q
%gpx:         pointing at Suuze's camera
%act:         looks through the camera
*AMO:         baa@o .
*AMO:         Reechan mo haitteru .
*CHI:         Reechan &=laugh .
@Situation:   reading books
*CHI:         ku(ru)m:a .
*CHI:         koko ne # hoshisan .
%gpx:         pointing at stars
*SUZ:         hoshisan .
*CHI:         &kumaSan [: kumasan] .
*SUZ:         kumasan mo .
*CHI:         shup(patsu) [/] shuppatsu .
*CHI:         <kore ne> [/] kore ne bubu da .
*SUZ:         fuun .
*AMO:         kore shiranai .
%cod:         $NEG
*CHI:         <kore ne> [/] kore ne ame !
%sit:         next picture shows many candies
.....
@End

```

- Headers

- Headers (which always begin with an @) show basic information about the transcript.
  - @Begin (obligatory)



frequency of each word that appears in the file named “aki20.cha” in the working folder.

- Various “switches” (which usually begin with “+” or “-”) are placed between the command name and the file name in order to limit or adjust the output.

freq +t\*CHI +f aki20.cha  
 Command      Switches      Filename

- This command looks only at the \*CHI tiers in the file “aki20.cha,” and saves the output as a file in the working folder, instead of displaying it in the output window.
- Frequently used commands
  - `combo` search for a particular word or combination of words that are either adjacent or non-adjacent.
  - `kw1` search for a particular word. Unlike `combo`, you can specify multiple words to do an “or” search.
  - `freq` returns the frequency of each word that occurs in the transcript(s).
  - `mlu` returns the mean length of utterance.
- Frequently used switches
  - `+f` saves the output in the output folder, instead of displaying it in the output window.
  - `+o` makes `freq` return the list of words in the order of frequency (i.e., not in the alphabetical order, which is the default setting).
  - `+u` combines the results from multiple transcripts. This is useful when you want to see the frequency of words over all target transcripts.
  - `+s` is used to specify a search string (word or string of words) in the `combo`, `freq`, and `kw1` commands. For example,  
`combo +snani aki20.cha`  
 returns utterances containing the word *nani* (but not *nanika*).
  - `+w` and `-w` are used to extract utterances along with their context. For example,  
`combo +sdare +t*CHI +w2 -w2 aki21.cha`  
 returns the utterances containing *dare* by the child, along with the preceding and the following two lines.
- \*: Wildcard
  - An asterisk “\*” is used in commands as a wildcard meaning “any string of letters.” For example,  
`mlu +t*CHI *.cha`  
 returns the child’s MLUs for all the CHAT files in the working folder (note that the asterisk following the `+t` switch is not a wildcard but the sign for independent tiers).  
 Another example:  
`combo +t*CHI +stabe* *.cha`  
 searches for the words beginning with *tabe* by Aki in all the CHAT files in the working folder.
- ^: “Immediately followed by”
  - You may want to search for not a single word but a string of words. In such a case, you use the “^” sign, which means “immediately followed by,” in a `combo` command. For example,  
`combo +t*CHI +sdore^tabe* *.cha`  
 means “Search all the CHAT files for utterances containing *dore* immediately followed

by a word beginning with *tabe-*. It should return utterances such as *dore tabeta*, *dore taberu* etc.

- You can use “^” signs together with a wildcard to search for a combination of words that are not adjacent. For example,
   
combo +t\*CHI +skore^^nani \*.cha
   
searches for *kore* followed by anything, followed by *nani*. It should return utterances such as *kore nani*, *kore wa nani* etc.

## II. EXERCISES

- Specify the working folder first!
- (1) What are three-letter codes for the target child and his mother in the Aki corpus?
  - (2) List all the inflectional forms of *taberu* produced by Aki in the order of appearance. Do the same with *nomu* and *kaku*.
  - (3) When does Aki utter his first *wh*-word? What is it? What if imitations are excluded? (Hint: You can use +s multiple times in a *kw* command.)
  - (4) Examine all the utterances by Aki containing *chitchai* immediately followed by *no*. Do you see any error(s)?
  - (5) Examine all the utterances by Aki containing *dore* and *ii* in this order, optionally with other words intervening. Do you see any ungrammatical utterance(s)?

- (6) What are the ten most frequent words in Aki's speech over all the files? How about his mother's speech? (Hint: This will produce a very long list, but you need only the top ten. In order to stop the analysis before it is complete, press [Ctrl]+[.]).
  
- (7) Compare the frequency of the following words in Aki's speech: *ga*, *o*, and *wa*. Do you need to produce the list of all the words in the transcripts?
  
- (8) When does Aki first answer (in an intelligible way) his mother's question containing *nani*?
  
- (9) [Advanced] Create MLU graphs for Aki and his mother using CLAN and Excel. (Hint: Use the +d switch to remove the redundant text, such as "Number of utterances" and "Number of morphemes," from the output.)