

## Practical Introduction to Corpus Analysis: L1 Acquisition of English and Russian

### Childes Tutorial

Irina Sekerina ([irina.sekerina@csi.cuny.edu](mailto:irina.sekerina@csi.cuny.edu))

---

#### 1. #9 -- BEFORE THE TUTORIAL:

- 1.1. Download and install CLAN from <https://dali.talkbank.org/clan/>
- 1.2. In the TalkBank/CLAN directory on your , create a new folder and call it "tutorial":
  - 1.2.1. Create a subfolder "Adam" in the "tutorial" folder.
  - 1.2.2. Download 12 Adam\*.cha files (1,8,10,12,14,20,22,24,28,30,32,34) to 1.2.1.
- 1.3. Download the following data files from the School web site to the folder "tutorial":
  - 1.3.1. T\_2018\_04\_19\_0.cha and the corresponding audio file T\_2018\_04\_19\_0\_an
  - 1.3.2. T1.cha
  - 1.3.3. Liza14.cha

#### 2. EXPLORING CLAN:

- 2.1. #11 -- Open CLAN
  - 2.1.1. Set your WORKING DIRECTORY to "tutorial" where the CHILDES data files from 1.3 are.
- 2.2. #12-13 -- Open the file "T\_2018\_04\_09\_0.cha" using FILE --> OPEN.
  - 2.2.1. Using MODE --> enable SONIC MODE and CONTINUOUS PLAYBACK (=ESC-8). Now we can listen to the audio recording to which this transcript is linked.
  - 2.2.2. Making a tier to stand out with color: VIEW --> COLOR KEYWORDS --> \*PEB --> ADD TO LIST, EDIT COLOR (red), COLOR ENTIRE TIER --> CLOSE DIALOG
  - 2.2.3. Close "T\_2018\_04\_09\_0.cha".

#### 3. WORKING WITH ENGLISH: ADAM32.cha:

- 3.1. #16 -- Set your working directory to "Adam"
- 3.2. #17 -- Open Adam32.cha to view it.
- 3.3. #18 -- Using basic analysis commands:
  - 3.3.1. In Command window, type `FREQ` --> click on FILE IN --> ADD Adam32.cha --> DONE
  - 3.3.2. After `"freq @"`, type `"t*CHI"` --> ENTER
  - 3.3.3. In Commands window, type `MLU` --> FILE IN --> ADD ALL --> DONE
  - 3.3.4. After `"mlu @"`, type `"t*CHI > adam_mlu"` --> ENTER
  - 3.3.5. In the folder "Adam", find the file "adam\_mlu" and click on it to view it in the text editor. You will see 12 MLU scores for Adam.
  - 3.3.6. Type `"kwal"` --> FILE IN --> CLEAR --> ADD Adam32.cha --> DONE
  - 3.3.7. After `"kwal @"`, type `"t*CHI +swhat"` --> ENTER
  - 3.3.8. Repeat for the wh-word "who"
  - 3.3.9. Repeat 3.3.7 but sending the results to the file "adam\_what"

### 3.4. #22-23 -- Acquisition of wh-questions:

- 3.4.1. combo @ +t\*CHI +s"^?" > adam32\_allqs --> 126 questions
- 3.4.2. combo @ +t\*CHI +t%mor +s"%mor:^^\*aux\*^^?" > adam32\_qswithaux
- 3.4.3. combo @ +t\*CHI +t%mor +s"%mor:^^(\*aux\*+\*be\*)^^?" > adam32\_qsauxorbe

## 4. WORKING WITH RUSSIAN

### 4.1. Set your working directory to "tutorial"

### 4.2. #42 -- View the files "Liza14.cha" and "T1.cha" in CLAN

- 4.2.1. Liza: freq @ +t\*CHI; MLU @ +t\*CHI; kwal @ +scto
- 4.2.2. T1: freq @ +t\*PEB; kwal @ +t\*MAM +scto

### 4.3. Parts-of-Speech:

- 4.3.1. Liza: freq @ +t\*CHI +t%mor +sm;\*
- 4.3.2. freq @ +t\*PEB +t%mor +sm;\*
- 4.3.3. Prepositions: freq @ +t\*PEB +t%mor +sm|PR

### 4.4. #44 -- CASE: INSTR

- 4.4.1. INSTR Case on nouns: freq @ +t\*CHI +t%mor +s"N|\*:INSTR"
- 4.4.2. INSTR on MASC: freq @ +t\*CHI +t%mor +s"N|\*:MASC\*:INSTR"
- 4.4.3. in Liza14.cha, with the line: freq @ +t\*CHI +t%mor +d +s"N|\*:MASC\*:INSTR"

### 4.5. #46-47 -- ASPECT:

- 4.5.1. Use T1.cha and find all the verbs that T produced at 3;09 (T1.cha) in:
  - 4.5.1.1. IMPERF:
  - 4.5.1.2. PERF:

## 5. childes-db: #49-51

### 5.1. Go to: <http://childes-db.stanford.edu/analyses.html>

- 5.1.1. Try all three analyses Frequency, Derived Measures, Population Properties using *Eng-UK* -> *Manchester* corpus. Visualize frequencies of different lexical items, e.g., wh-words *what* and *who*, *dog* and *jump* (Braginsky et., 2019), *water* (Deb ROy), etc.
- 5.1.2. Choose a different collection and a corpus and play with something new.

### 5.2. For API: <http://childes-db.stanford.edu/api.html>. *\*Note: the tutorial on the web site does not require any special programming skills as long as you are familiar with the basic R commands and structure.*

- 5.2.1. Install *RStudio*
- 5.2.2. Install *childesr* package
- 5.2.3. Go through the API tutorial.
- 5.2.4. Go through the exercises in Sanchez et al.'s (2019) article. *\*Note: These exercises require an intermediate-level familiarity with R.*

### 5.3. Work with the Slavic collection, 2 Russian corpora *Tanja* and *Protassova*:

#### 5.3.1. Library(childesr)

#### 5.3.2. Get transcripts:

```
>d_russian <- get_transcripts(corpus = c("Tanja", "Protassova"))
> nrow(d_russian)
> head(d_russian)
```

#### 5.3.3. Get tokens (*Protassova*):

```
d_varja_prod <- get_tokens(corpus = "Protassova", role = "target_child",
target_child = "Varvara", token = "chto")
```

#### 5.3.4. Get speaker stats (*Tanja*):

```
d_tanja_stats$num_utterances
```

```
>d_tanja_stats$num_utterances
> d_tanja_stats$num_types
> d_tanja_stats$num_tokens
> d_tanja_stats$num_morphemes
> d_tanja_stats$mlu_w
> mean(d_tanja_stats$mlu_w)
```