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

Irina A. Sekerina* and Sophia Malamud**

*College of Staten Island (CUNY) and Высшая школа экономики
**Brandeis University

24 June, 2020

7th Summer Neurolinguistics School, Center for Language and Brain, HSE

1. Working with *CHILDES* using CLAN

2. Working with *CHILDES*: English

3. *BiRCh:* Russian: V-final Word Order

4. *CHILDES:* Russian: Morphosyntax

5. *childes-db*

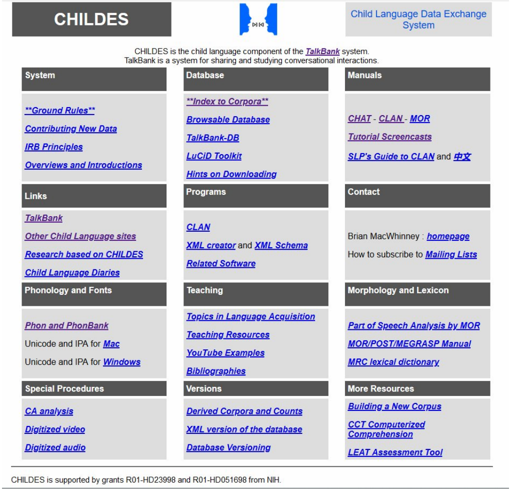
CHILDES and *BiRCh*

In this short tutorial aimed at beginners, we hope to demonstrate how L1 child corpora can be used to answer theoretical (and applied) questions in language acquisition research:

- *CHILDES*: Irina Sekerina
- *BiRCh*: Sophia Malamud

1. Working with *CHILDES* using CLAN

CHILDES entry screen



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	Database	Manuals
Ground Rules Contributing New Data IRB Principles Overviews and Introductions	**Index to Corpora** Browsable Database TalkBank-DB LuCiD Toolkit Hints on Downloading	CHAT - CLAN - MOR Tutorial Screencasts SLP's Guide to CLAN and 中文
Links	**Programs**	**Contact**
TalkBank Other Child Language sites Research based on CHILDES Child Language Diaries	CLAN XML creator and XML Schema Related Software	Brian MacWhinney : homepage How to subscribe to Mailing Lists
Phonology and Fonts	**Teaching**	**Morphology and Lexicon**
Phon and PhonBank Unicode and IPA for Mac Unicode and IPA for Windows	Topics in Language Acquisition Teaching Resources YouTube Examples Bibliographies	Part of Speech Analysis by MOR MOR/POST/MEGRASP Manual MRC lexical dictionary
Special Procedures	**Versions**	**More Resources**
CA analysis Digitized video Digitized audio	Derived Corpora and Counts XML version of the database Database Versioning	Building a New Corpus CCT Computerized Comprehension LEAT Assessment Tool

CHILDES is supported by grants R01-HD23998 and R01-HD051698 from NIH.

To learn the *CHILDES* System More

1. Understand the *.cha* format used for *CHILDES* data files
2. Download and install the CLAN program
3. Go through the CLAN tutorial: Ch. 3 of the Manual (pp. 12-29)
4. Do the exercises from the Manual
5. Explore the specific language corpora by using the analysis and profile CLAN commands

Understanding .cha Format

CHAT: Codes for the Human Analysis of Transcripts

1. A standardized format for producing computerized transcripts of face-to-face conversational interactions

2. Track a wide variety of structures, compute automatic indices, and analyze morphosyntax

3. Compatible with such programs as ELAN, Praat, EXMARaLDA, Phon, Transcriber, etc.

4. *Sonic and video CHAT:* CLAN is used to link transcripts to audio or video recordings

1. Working with *CHILDES* using CLAN

CLAN: Computerized Language ANalysis

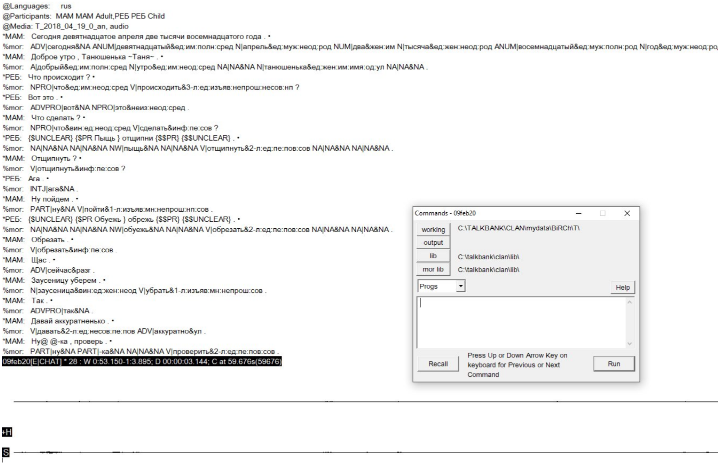
1. To follow this tutorial or to practice working with CLAN later, you need to dowload:
   1.1 The CLAN program
   1.2 The CLAN manual
   1.3 *CHILDES* data files: *Liza14.cha, T1.cha, T_2018_04_19_0.cha*
2. You could also access *CHILDES* files via the *CHILDES* browsable (CBD) interface directly at the TalkBank web site.

**Liza14.cha* and *T1.cha* used in this tutorial are not yet available in *CHILDES*.

What is *CLAN*?

1. The CLAN program:
   - is a collection of several different instructions that you can use to analyze your data.
   - Command + parameters + transcripts: mlu +t*CHI *.cha
2. There are seven types of CLAN commands:
   - Analysis commands are the basic commands for searching and corpus analysis: FREQ, KWAL, COMBO, MLU, TTR, etc.
   - Profiling commands put a large number of analysis and profiling commands into a single command package, often comparing a file against a database standard.

CLAN Windows



```
@Languages:	rus
@Participants:	MAM MAM Adult,РЕБ РЕБ Child
@Media: T_2018_04_19_0_an, audio
*MAM:	Сегодня девятнадцатое апреля две тысячи восемнадцатого года . •
%mor:	ADV|сегодня&NA ANUM|девятнадцатый&ед:им:полн:сред N|апрель&ед:муж:неод:род NUM|два&жен:им N|тысяча&ед:жен:неод:род ANUM|восемнадцатый&ед:муж:полн:род N|год&ед:муж:неод:ро,
*MAM:	Доброе утро , Танюшенька ~Таня~ . •
%mor:	A|добрый&ед:им:полн:сред N|утро&ед:им:неод:сред NA|NA&NA N|танюшенька&ед:жен:им:имя:од:ул NA|NA&NA .
*РЕБ:	Что происходит ? •
%mor:	NPRO|что&ед:им:неод:сред V|происходить&3-л:ед:изъяв:непрош:несов:нп ?
*РЕБ:	Вот это . •
%mor:	ADVPRO|вот&NA NPRO|это&неиз:неод:сред .
*MAM:	Что сделать ? •
%mor:	NPRO|что&вин:ед:неод:сред V|сделать&инф:пе:сов ?
*РЕБ:	{$UNCLEAR} {$PR Пыщь } отщипни {$$PR} {$$UNCLEAR} . •
%mor:	NA|NA&NA NA|NA&NA NW|пыщь&NA NA|NA&NA V|отщипнуть&2-л:ед:пе:пов:сов NA|NA&NA NA|NA&NA .
*MAM:	Отщипнуть ? •
%mor:	V|отщипнуть&инф:пе:сов ?
*РЕБ:	Ага . •
%mor:	INTJ|ага&NA .
*MAM:	Ну пойдем . •
%mor:	PART|ну&NA V|пойти&1-л:изъяв:мн:непрош:нп:сов .
*РЕБ:	{$UNCLEAR} {$PR Обуежь } обрежь {$$PR} {$$UNCLEAR} . •
%mor:	NA|NA&NA NA|NA&NA NW|обуежь&NA NA|NA&NA V|обрезать&2-л:ед:пе:пов:сов NA|NA&NA NA|NA&NA .
*MAM:	Обрезать . •
%mor:	V|обрезать&инф:пе:сов .
*MAM:	Щас . •
%mor:	ADV|сейчас&разг .
*MAM:	Заусеницу уберем . •
%mor:	N|заусеница&вин:ед:жен:неод V|убрать&1-л:изъяв:мн:непрош:сов .
*MAM:	Так . •
%mor:	ADVPRO|так&NA .
*MAM:	Давай аккуратненько . •
%mor:	V|давать&2-л:ед:несов:пе:пов ADV|аккуратно&ул .
*MAM:	Ну@ @-ка , проверь . •
%mor:	PART|ну&NA PART|-ка&NA NA|NA&NA V|проверить&2-л:ед:пе:пов:сов .
09feb20[E|CHAT] * 28 : W 0:53.150-1:3.895; D 00:00:03.144; C at 59.676s(59676)
```

Commands - 09feb20

working C:\TALKBANK\CLAN\mydata\BiRCh\T\
output
lib C:\talkbank\clan\lib\
mor lib C:\talkbank\clan\lib\

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

Sonic CLAN: A Sample English FIle

Some of the data files have audio or video recordings attached to them.

- Example: Clinical-MOR/Ambrose/TD/36 at TalkBank web site.
- A 3-year-old typically developing child:

```
12 *MOT: oh where's the chicken ?
13 *CHI: why we have two sheeps [: sheep] [*] ?
14 *MOT: we just have one sheep .
15 *MOT: and there's a chicken and a farmer .
16 *CHI: why there two sheeps [: sheep] [*] ?
17 *MOT: where's the other sheep ?
18 *CHI: I don't know I xxx go find it .
19 *MOT: okay, xxx .
20 *MOT: a cow .
21 *CHI: cowie, cow [/] cow .
22 *MOT: here's where the horse goes so the cow must go there .
23 *MOT: there xxx cow .
24 *MOT: do those ?
25 *CHI: why is there two +...
```

A Sample CHAT Russian file: *T_2018_04_19_0.cha*

```
@Begin
@Languages: rus
@Participants: РЕБ Name, Target Child, MAM Mother
@ID:
@Birth of CHI:
@Location:
@Date:
*MAM: Ну как ты спала ?
%mor: PART|ну&NA ADVPRO|как&NA NPRO|ты&2-л:ед:им
V|спать&ед:жен:изъяв:несов:нп:прош?
*MAM: Что тебе приснилось ?
%mor: NPRO|что&ед:им:неод:сред NPRO|ты&2-л:дат:ед
V|присниться&ед:изъяв:нп:прош:сов:сред?
*РЕБ: Не расскажу .
%mor: PART|не&отрп V|рассказать&1-л:ед:изъяв:непрош:сов .
*MAM: Не расскажешь ?
%mor: PART|не&отрп V|рассказать&2-л:ед:изъяв:непрош:сов ?
*РЕБ: Мам .
%mor: N|мама&ед:жен:зват:од .
*РЕБ: Мне холодно .
%mor: NPRO|я&1-л:дат:ед ADV|холодно&прдк .
*MAM: Укрыть ? %mor: V|укрыть&инф:сов ?
```

Tiers in .cha Files for CLAN Analysis

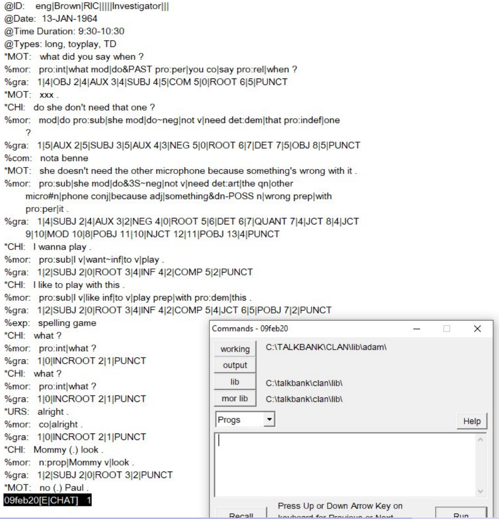
1. Participant tiers: =*lexical*
   - Eng: CHI, MOT, etc.
   - Rus: REB, MAM, etc.
2. Analysis tiers: *%mor, %gra* and many others
3. We will focus on the Participant tiers and *%mor* tier.

2. Working with *CHILDES*: English

Adam's transcripts

1. Adam's transcripts:
   - From the school's *tutorial* web site
   - Or directly: *CHILDES – Eng-NA – Brown – download transcripts*.
   - We will need: *Adam01, 08, 10, 12, 14, 20, 22, 24, 28, 30, 32, 34*.
2. Start CLAN
   - Set CLAN's working and output directory to the folder "tutorial"→ "Adam"

Adam32.cha in CLAN



```
@ID:    eng|Brown|RIC|||||Investigator|||
@Date:  13-JAN-1964
@Time Duration: 9:30-10:30
@Types: long, toyplay, TD
*MOT:   what did you say when ?
%mor:   pro:int|what mod|do&PAST pro:per|you co|say pro:rel|when ?
%gra:   1|4|OBJ 2|4|AUX 3|4|SUBJ 4|5|COM 5|0|ROOT 6|5|PUNCT
*MOT:   xxx .
*CHI:   do she don't need that one ?
%mor:   mod|do pro:sub|she mod|do~neg|not v|need det:dem|that pro:indef|one
        ?
%gra:   1|5|AUX 2|5|SUBJ 3|5|AUX 4|3|NEG 5|0|ROOT 6|7|DET 7|5|OBJ 8|5|PUNCT
%com:   nota benne
*MOT:   she doesn't need the other microphone because something's wrong with it .
%mor:   pro:sub|she mod|do&3S~neg|not v|need det:art|the qn|other
        micro#n|phone conj|because adj|something&dn-POSS n|wrong prep|with
        pro:per|it .
%gra:   1|4|SUBJ 2|4|AUX 3|2|NEG 4|0|ROOT 5|6|DET 6|7|QUANT 7|4|JCT 8|4|JCT
        9|10|MOD 10|8|POBJ 11|10|NJCT 12|11|POBJ 13|4|PUNCT
*CHI:   I wanna play .
%mor:   pro:sub|I v|want~inf|to v|play .
%gra:   1|2|SUBJ 2|0|ROOT 3|4|INF 4|2|COMP 5|2|PUNCT
*CHI:   I like to play with this .
%mor:   pro:sub|I v|like inf|to v|play prep|with pro:dem|this .
%gra:   1|2|SUBJ 2|0|ROOT 3|4|INF 4|2|COMP 5|4|JCT 6|5|POBJ 7|2|PUNCT
%exp:   spelling game
*CHI:   what ?
%mor:   pro:int|what ?
%gra:   1|0|INCROOT 2|1|PUNCT
*CHI:   what ?
%mor:   pro:int|what ?
%gra:   1|0|INCROOT 2|1|PUNCT
*URS:   alright .
%mor:   co|alright .
%gra:   1|0|INCROOT 2|1|PUNCT
*CHI:   Mommy (.) look .
%mor:   n:prop|Mommy v|look .
%gra:   1|2|SUBJ 2|0|ROOT 3|2|PUNCT
*MOT:   no (.) Paul .
09feb20[E|CHAT]   1
```

Commands - 09feb20

working C:\TALKBANK\CLAN\lib\adam\
output
lib C:\talkbank\clan\lib\
mor lib C:\talkbank\clan\lib\

Trying basic commands for *Adam32.cha*: FREQ, MLU, KWAL

1. **FREQ:** Click *File in* – Add Adam32: freq @ +t*CHI
   - 471 Total # of different item types used
   - 2504 Total# of item (tokens)
   - 0.188 Type/Token ration
2. **MLU:** File in all files: mlu @ +t*CHI > adam-mlu
   - Adam01: Age - 2;3.4 MLU - 2.202
   - Adam22: Age - 3;1.9 MLU - 4.05
3. **KWAL:** Adam32:
   - kwal @ +swhat +t*CHI - 37 instances
   - kwal @ +swho +t*CHI - 4 instances

*You could also use the RECALL button to re-use the same commands without re-typing them.

English Illustration 2: Acquisition of *Wh*-questions

Is acquisition of *wh*-questions lexically specific?

1. *I saw what – What did I see?*
2. A child must identify the lexical properties of *wh*-words
3. Order: *what, where*; *how; when, why, which, whose*
4. Landing site (*Where go?*), subject-auxiliary inversion, *do*-support

Roeper, T., & de Villiers, J. (2010). The acquisition path for *wh*-questions. *Handbook of Generative Approaches to Language Acquisition*. (pp. 189-246). Springer.

Acquisition of *wh*-Questions by Adam (Brown, 1973)

- Inversions appear at different points for *what, how, when*
- Formulaic questions with contractions: *what's, where's*
- *Why* and *why not* questions that were appended to declaratives his mother had just uttered:
  MOT: *You can't dance.*
  CHI: *Why not me can't dance?*

Acquisition of *wh*-Questions by Adam (Adam32.cha)

1. Some of the *what*-questions Adam produced:
   - *what the string is for ?*
   - *what is the string for ?*
   - *what dat [: that] behind me ?*
   - *what Paul driving ?*
   - *what he trying to get ?*
   - *what he went to play with ?*
2. We will investigate how Adam is doing with Subj-Aux inversion

*Based on laboratory examples by Paul Hogstrom (Boston University).

Acquisition of *wh*-Questions by Adam (Adam32.cha)

1. Find all *wh*–questions Adam produced and send the results to the file "adam_allqs"using the combo @ command:
   strings matched 126 times
2. Limit our search to questions that include auxiliaries by using *%mor* tier:
   - 440: *why he's stopping ?*
   - 464: *it has batteries ?*
   - 954: *is it a nap time [= For Paul] ?*
   - 1279: *is it Boston ?*
3. Questions with *aux* or copula *be*

Sub-Aux Inversion by Adam

Total # of questions in *Adam32.cha*: 33

question	**type**	**inverted**	**error**
wh-question	affirmative	*where is a door?*	**what the string is for?*
wh-question	negative	–	–
yes-no question	affirmative	*is it Boston?*	**it has batteries?*
yes-no question	negative	*isn't it cute?*	–

- We can calculate % of inversion in various questions and describe patterns of acquisition of *wh*-questions, *yes-no*-questions, negative questions, etc.
- If we have time, in Section 5, we will talk about childed-db that allows for simultaneous analysis of multiple corpora.

3. *BiRCh:* Russian: V-final Word Order

Russian word order

Typological consensus:
Russian is an SVO language with reordering possibilities usually attributed to specific information structure.

Ты	всех	четырех	лягушек	покормила
Ty	vseh	četyreh	ljagušek	pokormila
you.SG.NOM	all.ACC	four.ACC	frogs.PL.ACC	fed.PFV.F

'You fed all four frogs' (file O_2017_07_03_0)

Bailyn, J. F. (2001). Inversion, dislocation and optionality in Russian. Current issues in formal Slavic linguistics, 3, 280-293.
Makarchuk, I., & Slioussar, N.(2020). *SOV in Russian: Using large corpora to solve the enigma.* A talk presented at the FASL-29 Workshop, University of Washington. 8-11 May, 2020.
Yokoyama, O. T. (1986). Discourse and Word Order. Vol. 6. John Benjamins Publishing: p.234, p.326.

German and English word order

Typological consensus:

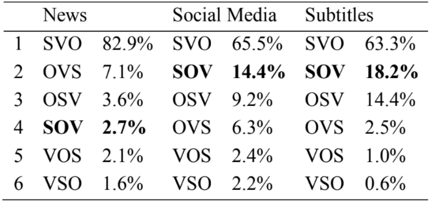
- German is an SOV language (arguably), V2 in main clauses, with reordering possibilities for arguments
  - Most frequent surface order in main clauses: VO
    but OV pretty frequent, too
  - Surface order in subordinate clauses: OV
- English is a strict SVO language with little reordering

Dryer, M. S. (2007). *Word order*. In T. Shopen (Ed.) Language typology and syntactic description, Vol.1, Ch.2, 61-131.

Louden, M. L. (1992). German as an object-verb language: A unification of generative and typological approaches. In I. Rauch, G. F. Carr, R. L. Kyes (Eds.) On Germanic Linguistics. Issues and Methods, 217-231.

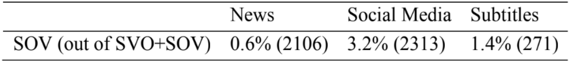
V-final order is rare in Russian (Makarchuk & Slioussar)

Taiga corpus: texts from open web, 6 billion words by now, POS-tagged & syntactically tagged in Universal Dependencies
⇒ studied data: 875k clauses including V_{trans}, S, & O in 3 genres.



	News		Social Media		Subtitles	
1	SVO	82.9%	SVO	65.5%	SVO	63.3%
2	OVS	7.1%	**SOV**	**14.4%**	**SOV**	**18.2%**
3	OSV	3.6%	OSV	9.2%	OSV	14.4%
4	**SOV**	**2.7%**	OVS	6.3%	OVS	2.5%
5	VOS	2.1%	VOS	2.4%	VOS	1.0%
6	VSO	1.6%	VSO	2.2%	VSO	0.6%

Table 1. The distribution of different word orders in our sample.



	News	Social Media	Subtitles
SOV (out of SVO+SOV)	0.6% (2106)	3.2% (2313)	1.4% (271)

Table 2. The share of SOV sentences among SOV and SVO sentences with non-pronominal objects.

V-final order is more frequent in Russian as a heritage language (RHL) speakers in Germany (Gagarina)

RUEG corpus: RHL speakers in Germany & the US; monolinguals

- **OV vs. VO**: evidence for a change in progress in monolingual Russian from rare OV to more OV.

  ⇒ Is OV underrepresented in Taiga?

  - No evidence of transfer given increased OV frequencies in Russian making it more like German.
- **V-final in subordinate clauses**: evidence of transfer from German: V-final more frequent in RHL speakers in Germany.
- **V-final in main declarative clauses**: no specific claim.

Gagarina, N. (2019). *Syntax: Word Order.* Lectures presented at the Higher School of Economics. December, 2019.

BiRCh: Is there evidence of transfer for V-final orders?

- How frequent are V-final orders in declarative clauses in monolingual Russian adults and children?
  - Once the data is parsed, simple syntactic search
  - For now, segmented and morphologically tagged data
    - A segment = main clause with all its subordinate clauses
    - Tagging includes POS, transitivity, case information
  - So, we will look for segments containing an accusative noun or pronoun followed by a transitive verb in an indicative form, immediately followed by ‘,’ or ‘.’
    - doesn’t separate main vs subordinate clauses

Pitfalls of searching in unparsed data

- Without syntactic annotation, the relationship between each verb and its object is not marked.
- This results in both the mistake of ‘noise’ = incorrect inclusion of an example in search results

В	него	ставят	свечки	и	дуют
V	nego	stavjat	svečki	i	dujut
in	it	put.3PL	candles.ACC	and	blow.3PL

‘They put candles in it and blow [them out]’

(file P_2018_02_18_1)

Pitfalls of searching in unparsed data

example of incorrect omission

- Without syntactic annotation, the relationship between each verb and its object is not marked.
- This results in both the mistake of 'noise' = incorrect inclusion of an example in search results
- and the mistake of 'silence' = incorrect omission of an example from search results

Мишка	любит	малину	кушать
Miška	ljubit	malinu	kušat
bear.SG.NOM	likes	raspberry.SG.ACC	eat.INF

'A/The bear likes to eat raspberries' (file O_2017_07_03_0)

Is there transfer in V-final orders in RHL?

- How frequent are these orders in RHL adults and children in Germany and the US?
  - If V-final in Germany » V-final in Russia, US
    - ⇒ evidence of transfer
  - If V-final in parents in Germany, US « V-final in Russia
    - ⇒ evidence of ongoing change in Russia not shared by immigrant populations

Is there transfer in V-final orders in RHL?

- Is there a difference in the information structure for these orders in RHL?
  - Like *Taiga*, *BiRCh* doesn't mark information structure
  - But pronouns tend to be given information.
  - Compare different populations with respect to the rates of V-final orders in which the object is a pronoun (NPRO) as opposed to a noun phrase based on a common noun (N)

Pitfalls of using pronominal status as proxy for information status

- Not all pronouns represent given information
  - e.g., всех / vseh / ‘all.ACC’, никого / nikogo / ‘nobody.ACC’ - no referent, either old or new
- We can make this more precise if we look for NPRO vs N as a measure of weight, and for a list of definite (personal) pronoun lemmas as a measure of givenness.

Посмотри	кого	я	нашла
Posmotri	kogo	ja	našla
Look.IMP.2SG	who.ACC	I.NOM	found.PFV.PAST.SG.F

‘Look whom I found’ (file S_2018_09_29_2)

Descriptive Stats for Morphologically Tagged Part of *BiRCh*

	US	**Russia**	**Germany**	**Total**
# of recordings	*76*	*53*	*13*	***142***
Duration (hour)	*25.55*	*22.15*	*4.43*	***52.13***
# of utterances	*31,325*	*25,842*	*6,586*	***63,753***
• child	*17,998*	*14,245*	*3,470*	***35,713***
• non-child	*13,327*	*11,597*	*3,116*	***28,040***
# of Russian words	*99,985*	*85,925*	*25,661*	***211,571***
• child	*34,690*	*33,280*	*10,069*	***78,039***
• non-child	*65,295*	*52,645*	*15,592*	***133,532***

V-final orders in the *BiRCh* Corpus

Here is a Google Sheets file containing

- the Python script we (= Alex Lưu) used to search the morphologically tagged transcripts
- the spreadsheet with the search results
- the tables with rates of search hits per 1000 utterances in the data from
  1. (1) adults, children in Germany
  2. (2) adults, children in the US
  3. (3) adults, children in Russia

  for accusative pronouns, accusative nouns, and overall.

Steps for creating the counts and graphs in Excel

- perform conditional counts of V-final constructions (e.g. COUNTIFS in Microsoft Excel/Goggle Sheets) based on:
  - regions (e.g. US vs Russia vs Germany)
  - speaker types (e.g. child vs adult)
  - part-of-speech types (e.g. noun vs pronoun)

Steps for creating the counts and graphs in Excel

- calculate rates of V-final constructions, for example, per 1000 utterances, constrained by regions and speaker types:
  - $rate = count \times 1000 \div \#_{utts}$

Steps for creating the counts and graphs in Excel

- create 2-D graphs (e.g. insert charts in Microsoft Excel/Google Sheets) from the target rate tables, in which the X-axis shows the grouping categories (regions and speaker types) and the Y-axis is the rate of V-final constructions

What's next

- We are planning to publish this subset of morphologically tagged part of *BiRCh* in *CHILDES* this year, so you can do all this stuff yourself using CLAN.
- If you want to use our data before that, email us, we want to work with you.
- Eventually, we will publish the full 1-million-word audio-aligned, segmented, and morphologically tagged *BiRCh* corpus in *CHILDES*
- We will also publish the full syntactically parsed *BiRCh* corpus on a separate website, equipped with a search function that does not require any computational savvy to use.

4. *CHILDES:* Russian: Morphosyntax

Trying basic commands for *Liza14.cha* and *T1.cha*

Monolingual Russian children:

1. **Liza14.chat:** transliterated (Gagarina, 2008)
   - Target_Child: CHI, age: 2;07
   - Run FREQ, MLU, KWAL (+schto)
2. **T1:** in Cyrillic (*BiRCh*, Malamud)
   - Target_Child: РЕБ, age: 3;09
   - Run FREQ, KWAL (+t*MAM +sчто)

*You could also use the RECALL button to re-use the same commands without re-typing them.

Illustration 1: Acquisition of Case in Russian

We can capitalize on the morphosyntactic features available at the *%mor* tier.

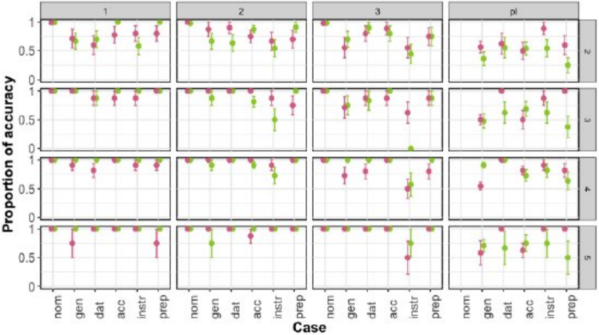
1. Parts-of-Speech: FREQ to get all words classified for PoS
2. Find all instances of a particular case (=INSTR):
   - Lisa: freq @ +t*CHI +t%mor +s“*:INSTR”
   - T1: freq @ +t*РЕБ +t%mor +s“N|*:муж*:твор”

Corpus: Acquisition of Case (Gagarina and colleagues)

- Order of appearance of case forms: 20-27 months
  - NOM and ACC-DAT-lexical ACC, GEN, PREP; no special GEN, INSTR
  - *Liza* (1;9): NOM-ACC-LOC-GEN-DAT-INSTR
- But: takes place later (36+ months) when modified by:
  - Pl number
  - 3rd declension class: FEM nouns ending in soft sign

Gagarina, N., & Voiekova, M. (2009). Acquisition of case and number in Russian. In *Development of Nominal Inflection in First Language Acquisition: A Cross-Linguistic Perspective.* (pp. 179-2106). Mouton de Gruyter.

Experimental Research: Acquisition of Case in Russian



Ladinskaya, N. et al. (2019). Acquisition of Russian nominal case inflections by monolingual children: A psycholinguistic approach. *Basic Research Program. Working Papers. Series Linguistics.81/LNG*

Illustration 2: Acquisition of Aspect in Russian

Some Russian data argue against the *Aspect First* hypothesis:

1;08-2;03 (Gagarina & Voiekova, 2009)

	IMPERF	PERF
Liza	55%	45%
Katya	58%	42%

1;06-2;11, 4 corpora (Bar-Shalom, 2020):

- Past: both telic and atelic
- Future: PERF > IMPERF
- more PERF (69-96%) and earlier

2;0-2;07

	IMPERF	PERF
Vanya	49%	51%
Vitya	45%	55%

2;11-4;0: Russian-Turkish Bilingual boy *S.* (Antonova-Ünlü & Wei, 2016)

- At ceiling for both IMPERF and PERF verb forms

Bar-Shalom, E. (2002). Tense and aspect in early child Russian. *Language Acquisition, 10*(4), 321-337.

Exercise: Acquisition of Aspect by T1

As a quick check:

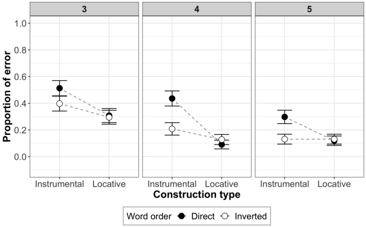
1. use *T1.cha* and find all verbs produced by T (at age 3;09)
   - in the IMPERF form
   - in the PERF form
2. Which CHILDES analysis command will you use?
3. Write down the #:
   - IMPERF =
   - PERF =

Complex Phenomena: Verbs with double objects

Sentences with iconic (=direct) and non-iconic (inverted) order of two objects:

1. Direct ACC-INSTR: *Накрой тарелочку платочком.*
   'Cover [a] plate-ACC [a] handkerchief-INSTR'
2. Inverted INSTR-ACC: *Накрой платочком тарелочку.*

- Correlation between input and language development
- Do Russian children actually hear sentences with ACC-INSTR?
- Why is it a difficult question?
- *BiRCh* to the rescue: 530 utterances
  - *Ну он может лоб потереть рукой*
  - *Наполняю кружку водой*



Крабис, А. et al. (2017). Роль моторного стереотипа в понимании лингвистических пространственных конструкций детьми дошкольного возраста. *Вестник ВГУ. Серия: Лингвистика и Межкультурная коммуникация, 1*, 82-87.

childes-db (Sanchez et al., 2019)

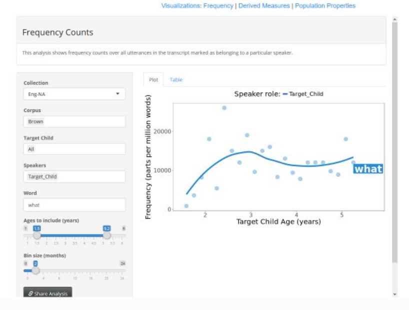
1. *CHILDES* uses specialized .cha format searchable in CLAN
   - Too difficult for novices and classroom use
   - Not very flexible for a processing pipeline in R or Python
2. childes-db: access to *CHILDES* through an application programming interface (API)
   - Visualizations
   - API in R

Sanchez, A., et al. (2019). childes-db: A flexible and reproducible interface to the *child language data exchange system*. *Behavior Research Methods, 51*, 1928-1941.

Acquisition of *wh*-questions in *childes-db*

Visualization: Frequency of the *wh*-word *what*:

Eng-NA, Brown



Visualizations: Frequency | Derived Measures | Population Properties

Frequency Counts

This analysis shows frequency counts over all utterances in the transcript marked as belonging to a particular speaker.

Collection: Eng-NA

Corpus: Brown

Target Child: All

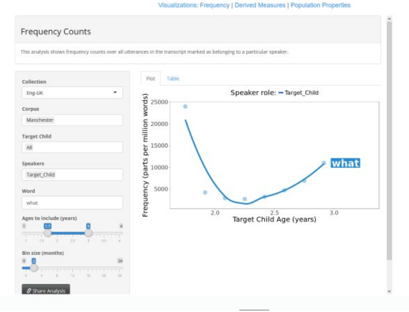
Speakers: Target_Child

Word: what

Ages to include (years): 1.5 – 5.2

Bin size (months): 2

Eng-UK, Manchester



Visualizations: Frequency | Derived Measures | Population Properties

Frequency Counts

This analysis shows frequency counts over all utterances in the transcript marked as belonging to a particular speaker.

Collection: Eng-UK

Corpus: Manchester

Target Child: All

Speakers: Target_Child

Word: what

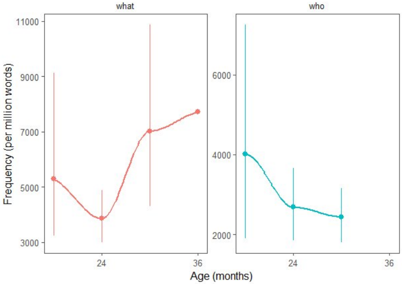
Ages to include (years): 1.7 – 3

Bin size (months): 2

Acquisition of *wh*-questions in *childes-db* (Manchester)

API: Acquisition of the *wh*-words *what* and *who*:

1. Install the package and load library *childesr*
2. Use R to work with data with the API *get* data functions:
   - Functions, e.g., *get_ transcripts, participants, utterances, types, tokens*, speaker stats, etc., allow you to extract data from the aggregated corpora
   - d_manchester <- get_transcripts(corpus = "Manchester"): 804 rows



Contact Information

- Irina Sekerina (CUNY and Высшая школа экономики): irina.sekerina@csi.cuny.edu
- Sophia Malamud (Brandeis): smalamud@brandeis.edu