Introduction to *CHILDES* Corpora: English and Russian

Irina A. Sekerina

College of Staten Island and The Graduate Center (CUNY)
Центр языка и мозга (Высшая школа экономики)

23 June, 2020

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

Relevant Research Milestones

- **1997-1999:** First experiments (with children and Russian participants) in the *Visual World Paradigm* (UPenn)
- **2002-2010:** VWP experiments with monolingual English- and Russian-speaking adults and children (CUNY)
- **2011-2019:** VWP experiments with bilingual heritage Russian adults (CUNY)
- **2017-present:**
  - Eye movements in reading and literacy acquisition in children and heritage Russian adults (CUNY, Высшая школа экономики)
  - Experiments, assessment, and corpus data with bilingual heritage Russian children (English, Norwegian, German)

Irina Sekerina's web site on ResearchGate.

Methods and Modalities

- Behavioral:
  - Preferential looking paradigm, looking-while-listening
  - Act-out
  - Sentence-picture (matching, selection, verification)
  - Cross-modal priming
  - The Visual World eye-tracking paradigm
- Corpus-based:
  - **CHILDES corpora**
- Neuroimaging:
  - ERP, fMRI, NIRS, MEG, etc.
- Modality:
  - **Naturalistic production**
  - Comprehension

"*Theory Building Requires More Data.*" (Michael Frank)

Input predictors (20 min annotated recordings) →

Scientific hypothesis about learning →

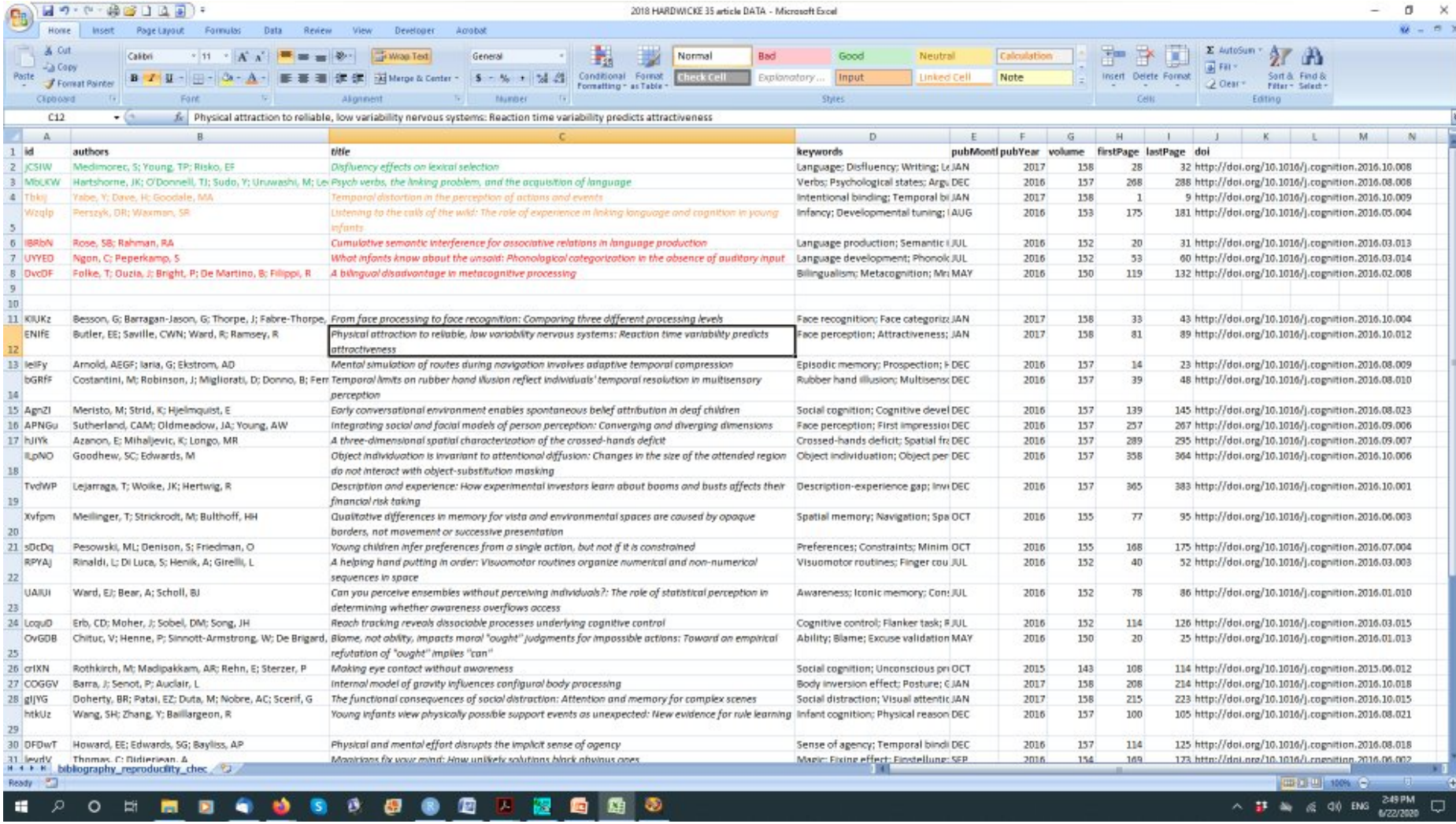
Outcomes (a handful of experimental results)

Crisis in reproducible research:

- Small scale experiments may not be replicable (*Open Science Framework*, 2015)
- 35 articles from *Cognition*: Only 13 were reproduced (Hardwicke et al., 2018)

Frank, M. (2019). *The role of data sharing in studying language learning: WordBank and childes-db.* A talk at the Symposium to honor Brian Macwhinney.

35 *Cognition* articles (Hardwicke et al., 2018)



id	authors	title	keywords	pubMonth	pubYear	volume	firstPage	lastPage	doi
jCSIW	Medimorec, S; Young, TP; Risko, EF	*Disfluency effects on lexical selection*	Language; Disfluency; Writing; Le	JAN	2017	158	28	32	http://doi.org/10.1016/j.cognition.2016.10.008
MbLKW	Hartshorne, JK; O'Donnell, TJ; Sudo, Y; Uruwashi, M; Le	*Psych verbs, the linking problem, and the acquisition of language*	Verbs; Psychological states; Arg	DEC	2016	157	268	288	http://doi.org/10.1016/j.cognition.2016.08.008
Tbkij	Yabe, Y; Dove, H; Goodale, MA	*Temporal distortion in the perception of actions and events*	Intentional binding; Temporal bi	JAN	2017	158	1	9	http://doi.org/10.1016/j.cognition.2016.10.009
Wzqlp	Perszyk, DR; Waxman, SR	*Listening to the calls of the wild: The role of experience in linking language and cognition in young infants*	Infancy; Developmental tuning; I	AUG	2016	153	175	181	http://doi.org/10.1016/j.cognition.2016.05.004
IBRbN	Rose, SB; Rahman, RA	*Cumulative semantic interference for associative relations in language production*	Language production; Semantic i	JUL	2016	152	20	31	http://doi.org/10.1016/j.cognition.2016.03.013
UYYEO	Ngon, C; Peperkamp, S	*What infants know about the unsaid: Phonological categorization in the absence of auditory input*	Language development; Phonolo	JUL	2016	152	53	60	http://doi.org/10.1016/j.cognition.2016.03.014
DvcDF	Folke, T; Ouzia, J; Bright, P; De Martino, B; Filippi, R	*A bilingual disadvantage in metacognitive processing*	Bilingualism; Metacognition; Mr	MAY	2016	150	119	132	http://doi.org/10.1016/j.cognition.2016.02.008
KIUKz	Besson, G; Barragan-Jason, G; Thorpe, J; Fabre-Thorpe,	*From face processing to face recognition: Comparing three different processing levels*	Face recognition; Face categoriz	JAN	2017	158	33	43	http://doi.org/10.1016/j.cognition.2016.10.004
ENIfE	Butler, EE; Saville, CWN; Ward, R; Ramsey, R	*Physical attraction to reliable, low variability nervous systems: Reaction time variability predicts attractiveness*	Face perception; Attractiveness;	JAN	2017	158	81	89	http://doi.org/10.1016/j.cognition.2016.10.012
IeIFy	Arnold, AEGF; Iaria, G; Ekstrom, AD	*Mental simulation of routes during navigation involves adaptive temporal compression*	Episodic memory; Prospection; F	DEC	2016	157	14	23	http://doi.org/10.1016/j.cognition.2016.08.009
bGRfF	Costantini, M; Robinson, J; Migliorati, D; Donno, B; Ferr	*Temporal limits on rubber hand illusion reflect individuals' temporal resolution in multisensory perception*	Rubber hand illusion; Multisens	DEC	2016	157	39	48	http://doi.org/10.1016/j.cognition.2016.08.010
AgnZI	Meristo, M; Strid, K; Hjelmquist, E	*Early conversational environment enables spontaneous belief attribution in deaf children*	Social cognition; Cognitive devel	DEC	2016	157	139	145	http://doi.org/10.1016/j.cognition.2016.08.023
APNGu	Sutherland, CAM; Oldmeadow, JA; Young, AW	*Integrating social and facial models of person perception: Converging and diverging dimensions*	Face perception; First impressio	DEC	2016	157	257	267	http://doi.org/10.1016/j.cognition.2016.09.006
hJIYk	Azanon, E; Mihaljevic, K; Longo, MR	*A three-dimensional spatial characterization of the crossed-hands deficit*	Crossed-hands deficit; Spatial fra	DEC	2016	157	289	295	http://doi.org/10.1016/j.cognition.2016.09.007
ILpNO	Goodhew, SC; Edwards, M	*Object individuation is invariant to attentional diffusion: Changes in the size of the attended region do not interact with object-substitution masking*	Object individuation; Object per	DEC	2016	157	358	364	http://doi.org/10.1016/j.cognition.2016.10.006
TvdWP	Lejarraga, T; Wolke, JK; Hertwig, R	*Description and experience: How experimental investors learn about booms and busts affects their financial risk taking*	Description-experience gap; Inv	DEC	2016	157	365	383	http://doi.org/10.1016/j.cognition.2016.10.001
Xvfpm	Meilinger, T; Strickrodt, M; Bulthoff, HH	*Qualitative differences in memory for vista and environmental spaces are caused by opaque borders, not movement or successive presentation*	Spatial memory; Navigation; Spa	OCT	2016	155	77	95	http://doi.org/10.1016/j.cognition.2016.06.003
sDcDq	Pesowski, ML; Denison, S; Friedman, O	*Young children infer preferences from a single action, but not if it is constrained*	Preferences; Constraints; Minim	OCT	2016	155	168	175	http://doi.org/10.1016/j.cognition.2016.07.004
RPYAj	Rinaldi, L; Di Luca, S; Henik, A; Girelli, L	*A helping hand putting in order: Visuomotor routines organize numerical and non-numerical sequences in space*	Visuomotor routines; Finger cou	JUL	2016	152	40	52	http://doi.org/10.1016/j.cognition.2016.03.003
UAIUI	Ward, EJ; Bear, A; Scholl, BJ	*Can you perceive ensembles without perceiving individuals?: The role of statistical perception in determining whether awareness overflows access*	Awareness; Iconic memory; Con	JUL	2016	152	78	86	http://doi.org/10.1016/j.cognition.2016.01.010
LoquD	Erb, CD; Moher, J; Sobel, DM; Song, JH	*Reach tracking reveals dissociable processes underlying cognitive control*	Cognitive control; Flanker task; R	JUL	2016	152	114	126	http://doi.org/10.1016/j.cognition.2016.03.015
OvGDB	Chituc, V; Henne, P; Sinnott-Armstrong, W; De Brigard,	*Blame, not ability, impacts moral "ought" judgments for impossible actions: Toward an empirical refutation of "ought" implies "can"*	Ability; Blame; Excuse validation	MAY	2016	150	20	25	http://doi.org/10.1016/j.cognition.2016.01.013
crIXN	Rothkirch, M; Madipakkam, AR; Rehn, E; Sterzer, P	*Making eye contact without awareness*	Social cognition; Unconscious pr	OCT	2015	143	108	114	http://doi.org/10.1016/j.cognition.2015.06.012
COGGV	Barra, J; Senot, P; Auclair, L	*Internal model of gravity influences configural body processing*	Body inversion effect; Posture; C	JAN	2017	158	208	214	http://doi.org/10.1016/j.cognition.2016.10.018
gIjYG	Doherty, BR; Patai, EZ; Duta, M; Nobre, AC; Scerif, G	*The functional consequences of social distraction: Attention and memory for complex scenes*	Social distraction; Visual attentic	JAN	2017	158	215	223	http://doi.org/10.1016/j.cognition.2016.10.015
htkUz	Wang, SH; Zhang, Y; Baillargeon, R	*Young infants view physically possible support events as unexpected: New evidence for rule learning*	Infant cognition; Physical reason	DEC	2016	157	100	105	http://doi.org/10.1016/j.cognition.2016.08.021
DFDwT	Howard, EE; Edwards, SG; Bayliss, AP	*Physical and mental effort disrupts the implicit sense of agency*	Sense of agency; Temporal bindi	DEC	2016	157	114	125	http://doi.org/10.1016/j.cognition.2016.08.018
levrV	Thomas, C; Didierjean, A	*Magicians fix your mind: How unlikely solutions block obvious ones*	Magic; Fixing effect; Fixatellung;	SEP	2016	156	169	173	http://doi.org/10.1016/j.cognition.2016.06.002

Experiments vs. Corpora

	Experiments	**Naturalistic data**
Advantages	-Can target particular forms -Can pin down development	-Only way to capture what children hear -Possible breadth of coverage and context
Disadvantages	-Can be very artificial	-Difficulty of sampling dense enough data -How naturalistic is it?

Solution: Use one (i.e., *CHILDES*) as a control for the other

1. Introduction to *CHILDES*

TalkBank (http://talkbank.org)

TalkBank is a database of 14 language databases:

- AphasiaBank for language in aphasia
- ASD Bank for language in autism
- BilingBank for the study of bilingualism and code-switching
- CHILDES for child language acquisition
- HomeBank for daylong recordings in the home

Before *CHILDES*

1. **Paper diaries of children's speech:**
   - Darwin (1877), Leopold (1949)
   - А. Н. Гвоздев (1949; 2019). *От первых слов до первого класса: Дневник научных наблюдений.*
2. **Transcripts:** From 1950s
   - Brown (1973): Typed transcripts: *Adam, Eva, and Sarah*

Not Just *CHILDES*

Proprietary databases:

1. **English:** Deb Roy (MIT): A longitudinal dense collection of video recordings of Roy's son's language acquisition.
   Watch Roy's clip from his TED talk in 2011 on the word *water*.
2. **Russian:**
   - Sabine Stall (U of Zurich): *ACQDIV* - L1 acquisition data from 10 languages, including a corpus of 5 Russian children
   - Natalia Gagarina (Leibniz-ZAS, Berlin) – 4 monolingual, 2 bilingual corpora compatible with *CHILDES*
   - Sophia Malamud (Brandeis U): *BiRCh* – a number of monolingual and bilingual Russian children's corpora

Gagarina (2008)

Researchers often take a long time to prepare their collected data and transcribe them in the .cha format. In addition to transcribing, morphosyntactic tagging is also a time-consuming step.

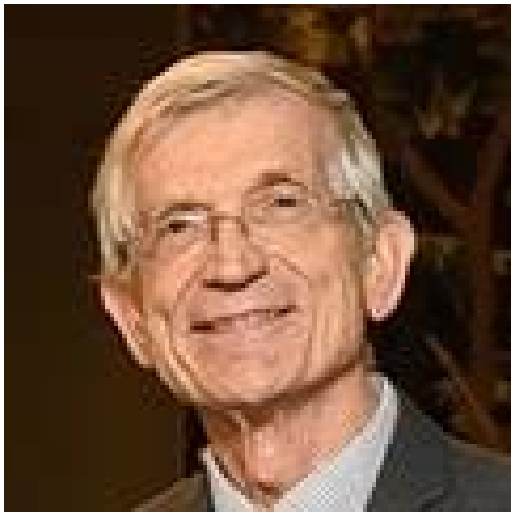
- Morphosyntactic parser for Russian: MyStem

Natalia Gagarina (ZAS-Leibniz, Berlin):

1. 4 monolingual children: *Liza, Roma, Vanja, Vitja,*
2. 2 Russian-German children: *Katja, Maya*

Гагарина, Н.В. (2008). *Становление грамматических категорий русского глагола в детской речи.* СПб.: Наука. ISBN 978-5-02-025279-0

CHILDES: 1984-



Beyond a simple data repository:

- The MacArthur Foundation grant to Brian MacWhinney (Carnegie Mellon U) and Catherine Snow
- **3 tools:**
  1. **Databases**
  2. CHAT: transcriptions
  3. CLAN: analysis

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

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

CHILDES Index of Corpora by Language

- Bilingual
- Clinical-MOR
- Narrative
- *Frog* story narratives
- Eng-NA and -UK,Chinese, East Asian, French, German, Scandinavian, Spanish, Celtic, Romance, Other 1-3
- Slavic: Croatian, Czech, Polish, Russian, Serbian, and Slovenian

A Sample CHAT File: English – *Adam* (Brown Corpus)

```
@Begin
@Languages: eng
@Participants: CHI Adam Target Child, MOT Mother
@ID: eng|Brown|CHI|2;03.04|male|typical|MC|Target Child|||
@ID: eng|Brown|MOT||female|||Mother|||
@Date: 08-OCT-1962
@Comment: Birth of CHI is 4-JUL-1960
@Time Duration: 10:00-11:00
@Types: long, toyplay, TD
*CHI: play checkers .
%mor: n|play n|checker-PL .
%gra: 1|2|MOD 2|0|INCROOT 3|2|PUNCT
%xpho: <1> pe
*CHI: big drum .
%mor: adj|big n|drum .
%gra: 1|2|MOD 2|0|INCROOT 3|2|PUNCT
*MOT: big drum ?
%mor: adj|big n|drum ?
%gra: 1|2|MOD 2|0|INCROOT 3|2|PUNCT
```

A Sample CHAT Russian file: *T_2018_04_19_0.cha*

@Begin
@Languages: rus
@Participants: РЕБ Name, Target Child, МАМ Mother
@ID:
@Birth of CHI:
@Location:
@Date:
*МАМ: 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?
*МАМ: 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 .
*МАМ: 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 .
*МАМ: Укрыть ? %mor: V|укрыть&инф:сов ?

2. Using *CHILDES* in language acquisition research

Kinds of Questions to Address with CHILDES

1. How language is used in naturalistic interactions
2. Data already exist so they could be used to answer new research questions
3. Child-directed speech (=input) and child output
   - Rare constructions: Passives
   - Statistical regularities: co-occurrences between pronouns and verbs

---

Corrigan, R. (2012). Using the CHILDES database. *Research Methods in Child Language: A Practical Guide.* (pp. 271-284). Blackwell.

Classical Central Themes

1. Longitudinal case studies (Roger Brown; Deb Roy)
2. Prelinguistic development: birth to 1-word, sound making (cooing and babbling)
3. Lexical and semantic development: over- and under-extensions, the *Semantic Feature hypothesis* (Eve Clark)
4. Morphology: the *Wug* experiment (Jean Berko-Gleason)
5. Phonology and intonation: 'mispronunciations' (Cruttenden)
6. Grammar: passives, verb classes (Carol Chomsky)
7. Metalinguistic awareness
8. Child-directed speech (CDS) (Catherine Snow)

More Recent Computational Themes

1. Learning where the stress is in words. Metrical systems group syllables into metrical feet differently in different languages.
2. Transitional probability learning
3. Learning parts of speech (grammatical categorization of words): frequent frames are useful for languages with fixed word order
4. Learning morphology: the English past tense

Pearl, L. (2010). Chapter 8. Using computational modeling in language acquisition research. *Experimental Methods in Language Acquisition Research.* (pp.163-184). John Benjamins.

German Illustration (Behrens, 2006): UG vs. Usage-Based

Leo: 1;11-4;11

1. # of recordings: 383
2. # of utterances: Input 258,592 - Leo's output: 158,336
3. # of words: Input 1,363,955 - Leo's output: 495,681

Analysis:

1. Production measures: MLU, PoS, N (300,000) and V (200,000) morphosyntax; speed of production

Leo showed a steady approximation towards the adult distribution.

Behrens, H. (2006). The input-output relationship in first language acquisition. *Language and Cognitive Processes, 21*, 2-24.

3. *CHILDES*: English and Russian

CHILDES: Eng-NA

- **50+ databases with North American English child language:**
  - Age: 0;6 months-8 years
  - 4 databases with video recordings
  - 21 databases with audio recordings
  - all include morphological analysis (%mor and %gra tiers)
  - McMillan (2004)

CHILDES Data and their analysis with *CLAN* (p. 49)

There are seven types of CLAN commands:

1. **Analysis commands are the basic commands for searching and corpus analysis: FREQ, KWAL, COMBO, MLU, TTR, etc.**
2. Profiling commands put a large number of analysis and profiling commands into a single command package, often comparing a file against a database standard.
3. Format conversion commands convert from other formats to CHAT and from CHAT to other formats.
4. Reformatting commands are used to add features to transcripts which have passed CHECK and are in good CHAT format.
5. Format repair commands are used to rework files into CHAT format.
6. Supplementary commands are for file operations such as renaming or deleting files.
7. Morphosyntactic analysis commands serve to create a %mor tier for part-of-speech tagging and a %gra tier for grammatical relations tagging.

English Illustration 1: Early Words

Research question: What is the order in which words are acquired across different languages?

1. Input: Child-directed speech (CDS) from *CHILDES*
2. Output: 400 words from 32,000 children, 10 languages
3. 9 predictors: both production and comprehension
   - form (# of phonemes)
   - frequency
   - MLU-w
   - meaning (concreteness, arousal, babiness)
   - input

Braginsky, M., at al. (2019). Consistency and variability in children's word learning across languages. *Open Mind. Discoveries in Cognitive Science, 3*, 52-67.

Braginsky et al., 2019: *CHILDES* and *WordBank*

Table 1. Statistics for data from Wordbank and CHILDES. N indicates number of children.

		Production		Comprehension		CHILDES	
Language	CDI items	N	Ages	N	Ages	Types	Tokens
Croatian	388	627	8–30	250	8–16	12,064	218,775
Danish	381	6,112	8–36	2,398	8–20	4,956	195,658
English (American)	393	7,312	8–30	1,792	8–18	45,597	7,679,042
French (Quebec)	396	1,364	8–30	537	8–16	28,819	2,551,113
Italian	392	1,400	7–36	648	7–24	7,544	188,879
Norwegian	380	7,466	8–36	2,374	8–20	10,670	231,763
Russian	410	1,805	8–36	768	8–18	5,191	32,398
Spanish (Mexican)	399	1,891	8–30	788	8–18	33,529	1,609,614
Swedish	371	1,367	8–28	467	8–16	8,815	359,155
Turkish	395	3,537	8–36	1,115	8–16	6,503	44,347

Note. CDI = MacArthur-Bates Communicative Development Inventory.

CDI for monolingual Russian children was developed in S.-Petersburgh by Stella Ceytlin and her colleagues

- I. up to 16 months: *Слова и жесты*
- II. up to 16 months: *Слова и предложения*

Predicting When Words are Learned: *dog* and *jump*

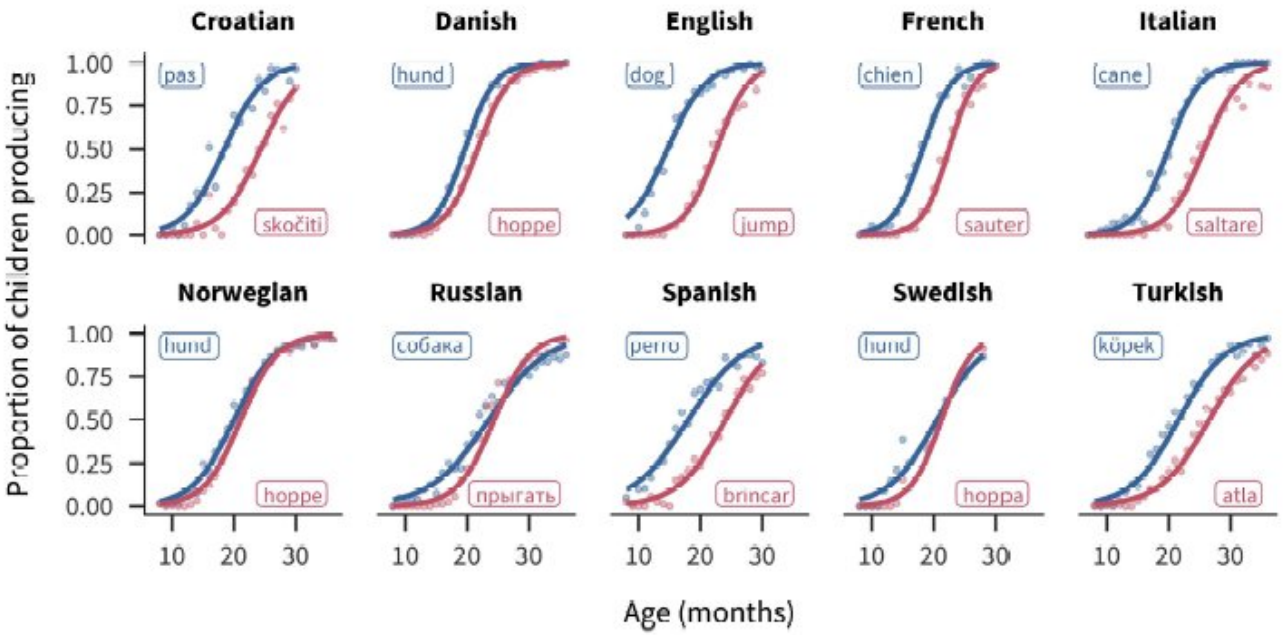


Figure 1. Example production trajectories for the words "dog" and "jump" across languages. Points show the proportion of children producing each word for each one-month age group. Lines show the best-fitting logistic curve. Labels show the forms of the words in each language.

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

Theories of Acquisition of *Wh*-questions

1. What is the pattern of correct use and error in English children's early *wh*-questions?
2. Inversion is present in Universal Grammar (UG):
   1. Children learn that inversion is obligatory *wh*-word by *wh*-word (Valian et al., 1992)
   2. Specific auxiliary (*DO*) and copula (*BE*) are difficult (Stromswold, 1990)

Rowland, C.F. et al. (2005). The incidence of error in young children's *wh*-questions. *Journal of Speech, Language, and Hearing Research, 48*. 384-404.

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?*

Rowland et al. (2005)

Manchester Corpus (Eng-UK)

- 12 children, age range: 1;8-3;0; MLU range: 1.58-3.49
- 2-hrs audio recordings, every 3 weeks for a year
- 34 1-hour transcripts were available for each child
- Mean # of *wh*-questions: 443

*We will have a closer look at the acquisition of *wh*-questions tomorrow, during our tutorial on *CHILDES*

Rowland et al. (2005): Participants' characteristics

Table 1. Participant information.

Child	CDI score	MLU from screening tape	Age range	MLU range	Total no. *wh*-questions
Anne	180	1.47	1;10.7–2;9.10	1.61–3.46	619
Aran	153	1.47	1;11.12–2;10.28	1.41–3.84	395
Becky	138	1.24	2;0.7–2;11.15	1.46–3.24	1,040
Carl	187	2.50	1;8.22–2;8.15	2.17–3.93	770
Dominic	153	1.25	1;10.24–2;10.16	1.20–2.85	203
Gail	262	1.48	1;11.27–2;11.12	1.76–3.42	495
Joel	122	1.13	1;11.1–2;10.11	1.33–3.32	351
John	191	2.12	1;11.15–2;10.24	2.22–2.93	177
Liz	359	Recording failed	1;11.9–2;10.18	1.35–4.12	447
Nicole	102	1.14	2;0.25–3;0.10	1.06–3.26	304
Ruth	44	1.43	1;11.15–2;11.21	1.41–3.35	201
Warren	124	1.62	1;10.06–2;9.20	2.01–4.12	316
M	167.92	1.53	—	1.58–3.49	443.17
Lara	—	—	2;7.21–2;11.14	MLU at start = 3.39	3,062

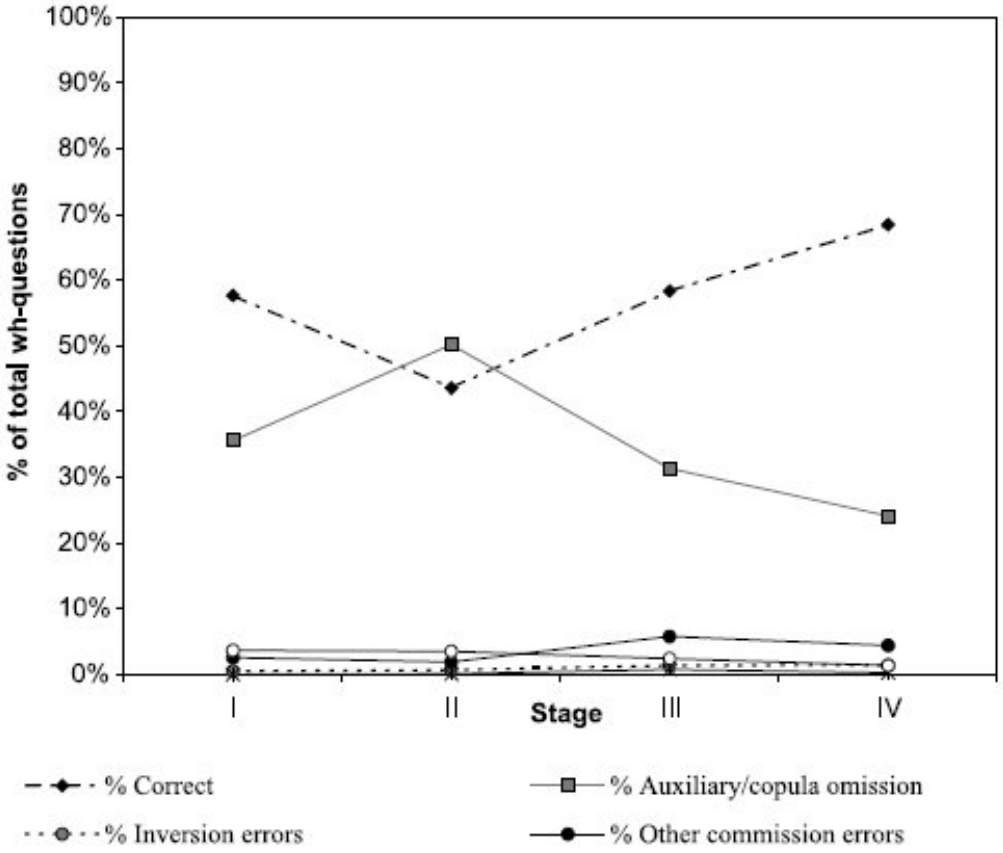
Note. CDI = MacArthur Communicative Development Inventory; MLU = mean length of utterance.

Rowland et al. (2005): Correct questions and omission errors

Brown's
MLU (mean length
of utterances) stages:
Stage I: MLU 1.00-1.99
Stage II: MLU 2.00-2.49
Stage III: MLU 2.50-2.99
Stage IV: MLU > 3

Results:

- Copula *is* > cop. *are*
- Aux *is* > aux *are*
- Aux *has* > aux *have*



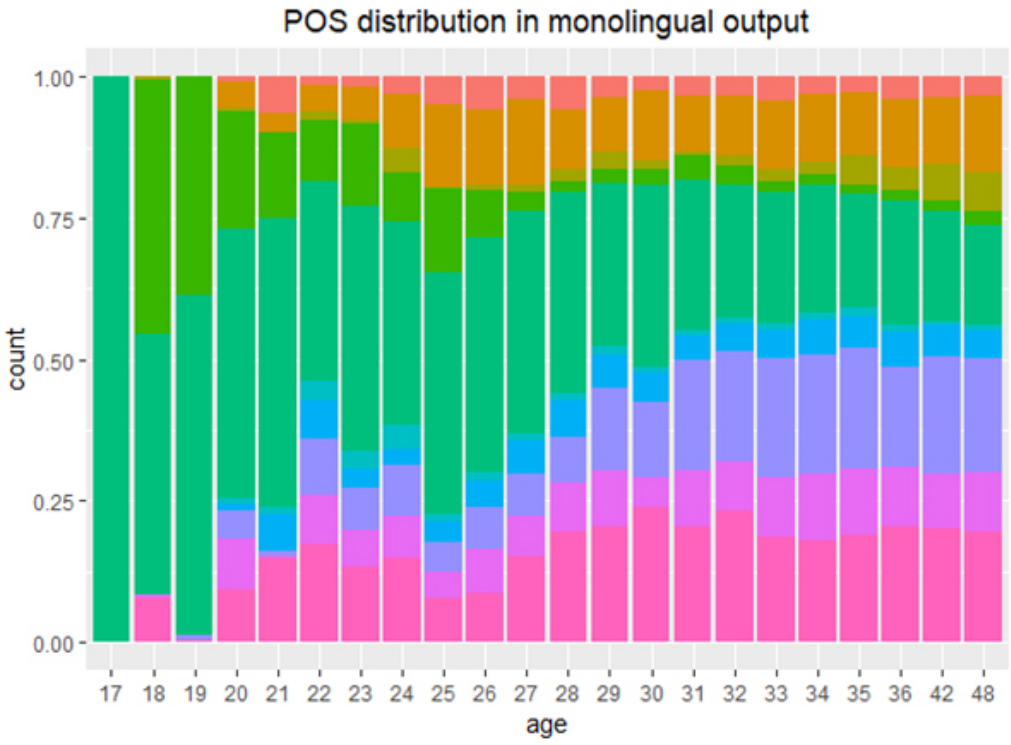
Monolingual Russian (Gagarina, 2008) and *Uliyana*

Corpus	**# transcripts**	**# child utterances**	**Mean MLU**
Liza	21	7032	2.5
Roma	12	1903	2.3
Vanja	22	19131	2.1
Vitja	15	5798	3
Bilingual: Uliyana	18	4552	2

Kobzeva, A. (2019). Distributional properties of input and output in the acquisition of Russian as a heritage language. *EMCL Master's Thesis*. University of Groningen.

Russian Illustration 1: Early noun bias

- Early noun advantage:
  - Noun-friendly languages (English, French)
  - Verb-friendly languages (Mandarin, Korean, Japanese)
- English: 10% are adjectives
- *Liza, Roma, Vanya, Vitya*[1]:
  - Age range: 1;5-4;0
  - # of transcripts: 12-22
  - # of child utterances: 2,000-20,000
- Early noun advantage in Russian[2]
- Few adjectives produced: 2.5%[2,3]

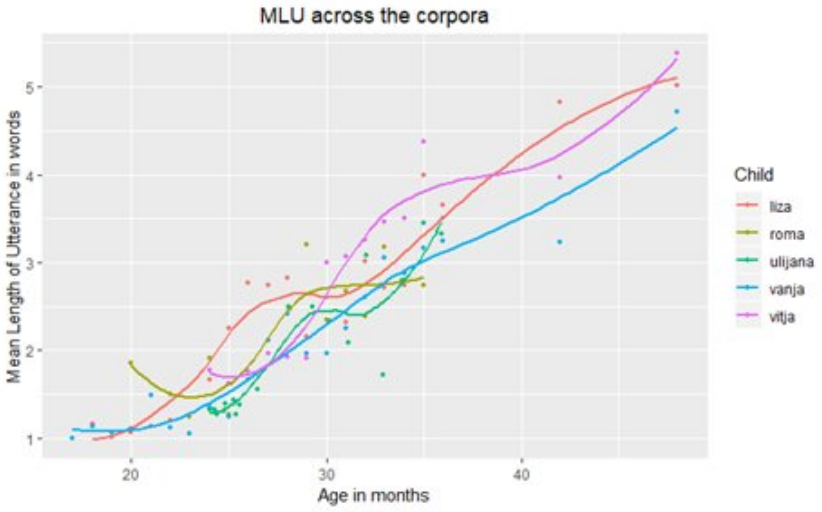


[1]Gagarina (2008); [2]Kobzeva (2019); [3]Tribushinina et al. (2018)

Russian Illustration 2: Russian MLU and TTR (Brown, 1973)

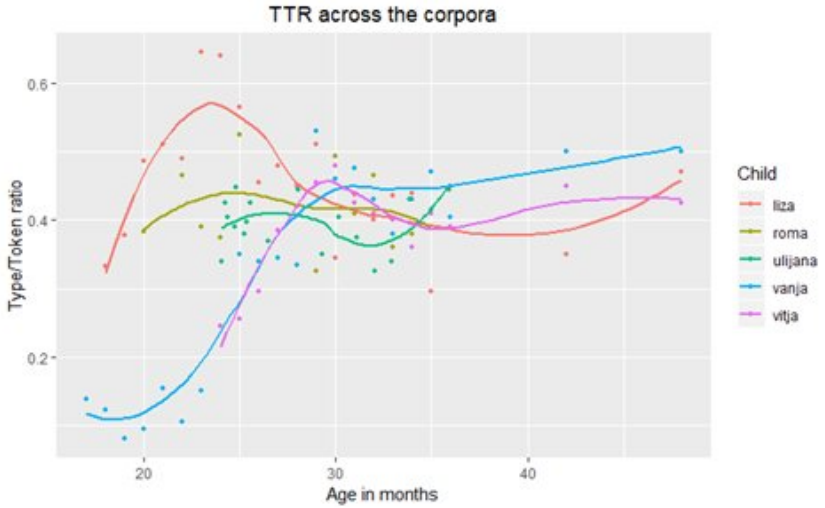
Mean length of utterance (MLU):

- # of morphemes/100 utterances
- Measure of general language development
- Tricky in Slavic languages



Type-token ratio (TTR):

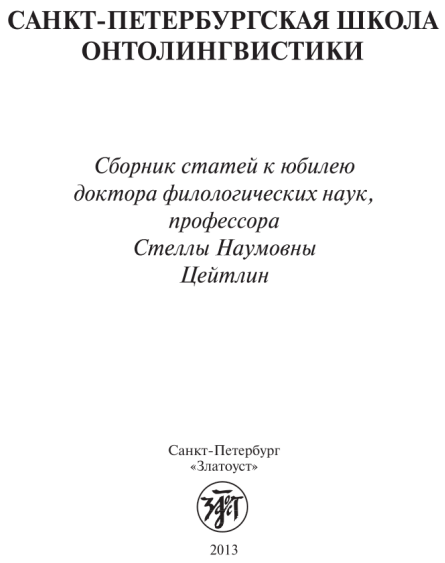
- # of different words/total # of words
- is heavily affected by the size of the text sample
- decreases with ages



On *CHILDES* in Russian

Using CHILDES on Russian and in Russian:

- Зыранова, Е. В. (2008). *Система CHILDES как метод сбора материалов и изучения детской речи.*
- Stella N. Ceytlin's school of *ontolinguistics* (St.-Petersburg)


САНКТ-ПЕТЕРБУРГСКАЯ ШКОЛА
ОНТОЛИНГВИСТИКИ

Сборник статей к юбилею
доктора филологических наук,
профессора
Стеллы Наумовны
Цейтлин

Санкт-Петербург
«Златоуст»
2013

4. *CHILDES*: Bilingual

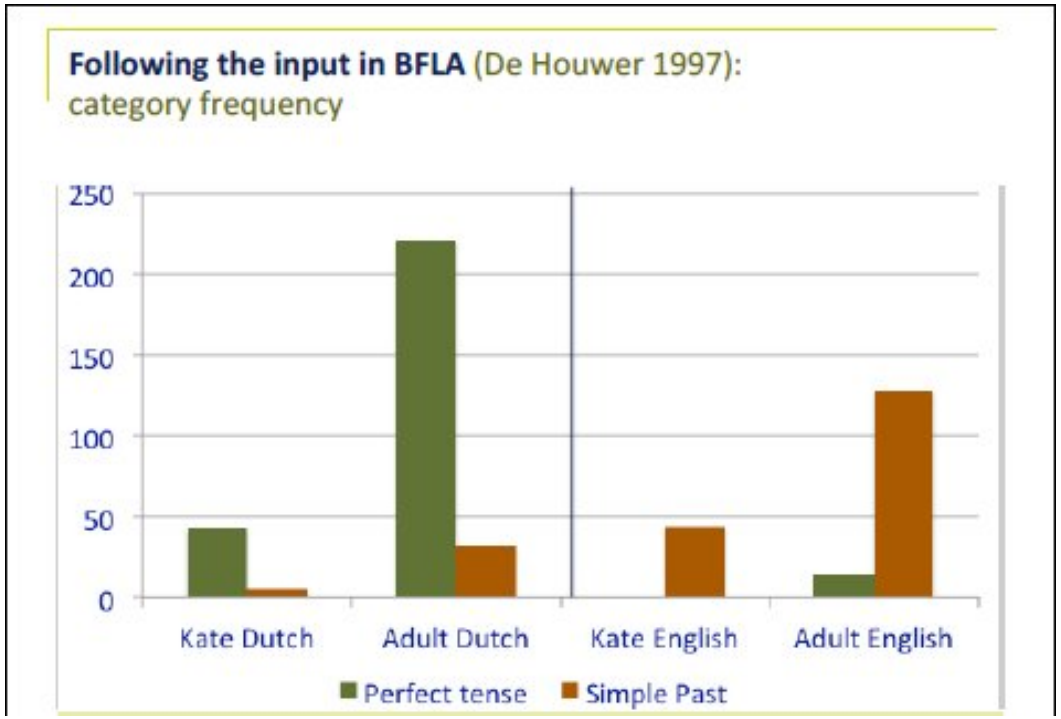
Bilingual L1 Acquisition: Dutch-English (De Houwer, 1990)

- *Kate*, bilingual English-Dutch child, audio recordings, 2;7-3;4
- 1979-1980: 19 files
- Manual extraction of utterances, hand counting
- 1985-86: Typed into the computer
- 1989-1990: transferred into *CHILDES*

- De Houwer, A. (1990). *The acquisition of two languages from birth: A case study.* Cambridge University Press.
- De Houwer, A. (2019). *Corpus-based work with CHAT/CLAN: A small catalog.* A talk at the Symposium to honor Brian MacWhinney.

De Houwer (1990): Past Tense Verbs

The Separate Development Hypothesis: Past tense verbs in the input and output



Yip, Mai, & Matthews (2018): *CHILDES* in Bilingualism

1. 2016: 32 bilingual corpora in *CHILDES*, 10 case studies. English: 12, Spanish: 7, Italian: 5, Dutch, French: 4, German: 3, Catalan, Portuguese: 2, Polish, Gaelic, Swedish: 1
2. 2020: 41 bilingual corpora, 18 case studies. New languages: Mandarin, Cantonese, Japanese, Arabic, Marathi, Farsi, Hungarian.

Yip et al. (2018). *CHILDES* in bilingualism. In *Bilingual Cognition and Language*. (pp. 183-202).

Yip, Mai, & Matthews (2018): Topics in Bilingualism

Comparing English and Cantonese:

1. Grammatical constructions:
   - Order of locative PP and V
   - Null subjects
   - Right dislocation (in English, the rate: 0.016-0.083% of utterances)
2. Language dominance assessment
   - MLU and MLU differentials
   - Upper bound: the longest utterance
   - VOCD
3. Code-switching

Russian in Bilingualism in *CHILDES*

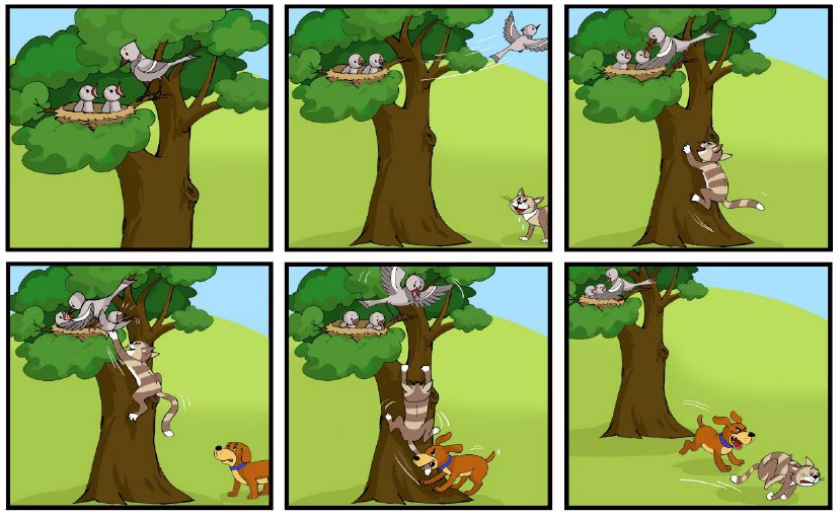
- -English: *Ionin* - 22 children (age: 2;4-12;5)
- -Dutch: *BiSLI* - 1059 transcripts (age: 3-9)
- -German: *ZAS Transcripts* - 80 children (age: 3;11-7;0)
- -French: *Bailleul* - 1 (age: 2;4-3;8)

BiSLI: MAIN Narratives of Russian-Dutch Bilingual Children

Russian-Dutch child (3;06.13):

@G: 1
*CHI: Cyplenok.
*EX1: Tak.
*CHI: I mama priletaet ego [?] cyplenok [*] xxx.
%com: ne sovsem ponjatno bez poslednego slova
- e to "svoego cyplenka xx"?
*EX1: Tak, xorosho, molodec.
*EX1: A zdes' chto?
@G: 2
*CHI: I potom ona uletel [*].
%com: uletela.
*CHI: i potom kotik prishla [*].
%com: prishel.
*EX1: Da, molodec, a zdes' chto?
@G: 3
*CHI: I potom netu <ptichki esli> [?].
%com: "ptichki" proiznosit kak "tichki".

Multilingual Assessment for Narratives:



- Special issue of *Applied Psycholinguistics* (2016), 37.
- MAIN materials on the ZAS-Leibniz web site

Russian Illustration 3: Case and Aspect

***S.*, Russian-Turkish bilingual child:** 2;11-4;0

1. # of recordings: 25, every 2 weeks for 30 min
2. Case: High accuracy
3. Aspect: At ceiling for both:
   - Imperfective: 1,015
   - Perfective: 838

Table 4. The use of Russian cases in S.'s data

Cases	Total use	Correct use	Percentage
Nominative	1930	1928	100
Genitive	312	256	82
Dative	163	155	95
Accusative	802	746	93
Instrumental	161	151	94
Prepositional	120	112	93

- Antonova-Ünlü, E., & Wei, L. (2016). Aspect acquisition in Russian as the weaker language. Evidence from a Turkish-Russian child. *International Journal of Bilingualism, 20*(2), 210-228.
- Antonova-Ünlü, E., & Wei, L. (2018). The acquisition of the weaker language. Evidence from the acquisition of Russian cases by a Turkish-Russian child. *Linguistic Approaches to Bilingualism, 8*(5), 637-663.

How to find out how the CHILDES Data are Used

The easiest way is to go to the particular corpus web page in *CHILDES*:

BiSLI Bilingual Corpus

Elena Tribushinina
Utrecht Institute of Linguistics
Utrecht University
e.tribushinina@uu.nl
website

Participants:	~1000
Type of Study:	narrative
Location:	Netherlands
Media type:	audio not open
DOI:	doi:10.21415/T5NG62

Browsable transcripts

Download transcripts

Citation information

Publications using these data should cite:

Gülzow, I. and Gagarina, N. (2007). Noun phrases, pronouns and anaphoric reference in young children narratives. In D. Bittner & N. Gagarina (eds.), *Intersentential pronominal reference in child and adult language*, pp. 203–223. Berlin: ZAS Papers in Linguistics.

Gagarina, N. (2012) Elicited narratives of monolingual Russian-speaking preschoolers: A comparison of typically developing children and children with language disorders. In L. Szucsich, N. Gagarina, E. Gorishneva & J. Leszkowicz (eds.) *Linguistische Beiträge zur Slavistik. XIX. Jungslavist/innen Treffen in Berlin, 16.-18. Dezember 2010* (= Specimina Philologiae Slavicae 171), pp. 71–90. München: Otto Sagner.

Tribushinina, E., Dubinkina, E. and Sanders, T. (2015) Can connective use differentiate between children with and without specific language impairment? *First Language* 35(1): 3–26.

Tribushinina, E., Mak, W.M., Andreiushina, E., Dubinkina, E. and Sanders, T. (2015). Connective use by bilinguals and monolinguals with SLI. *Bilingualism: Language and Cognition*. Online first, doi:10.1017/S1366728915000577.

Tribushinina, E., Mak, W.M., Andreiushina, E., Dubinkina, E. and Sanders, T. (2017). Connective use by bilinguals and monolinguals with SLI. *Bilingualism: Language and Cognition* 20(1), 98–113.

In accordance with TalkBank rules, any use of data from this corpus must be accompanied by at least one of the above references.

Project Description

This corpus contains 1058 transcriptions of narratives collected within the framework of the European project "Discourse Coherence in Bilingualism and SLI" coordinated by Elena Tribushinina (Utrecht University, The Netherlands), Natalia Gagarina (ZAS Berlin, Germany) and Ekaterina Abrosova (Herzen State Pedagogical University of Russia, St. Petersburg, Russia). The project aimed to disentangle discourse profiles of simultaneous bilinguals and their peers with a language impairment (SLI) and has been supported by a Marie Curie International Research Staff Exchange Scheme Fellowship within the 7th European Community Framework Programme (grant number 269173). Data collection by ZAS has also been partly supported by the German Federal Ministry for Education and Research (BMBF) and Deutsche Forschungsgemeinschaft (DFG). This database is related to an earlier corpus of narratives collected from Russian-German bilinguals (see Gagarina corpus in Bilingual Corpora).

Bilingual Russian Corpora in the Works

In addition to Natalia Gagarina's corpora at ZAS-Leibniz (Berlin):

1. BiRCh by Sophia Malamud (Brandeis University): both monolingual and bilingual (English, German)
2. Dense bilingual LENA corpora:
   - 2 heritage Russian-American English corpora: *Jenna* and *Sasha* (my lab, CUNY, New York)
   - 1 Heritage Russian-Norwegian corpus: *Nina* (Yulia Rodina, University of Tromsø, Norway)
3. Dense monolingual corpora: 2 at the Center for Language and Brain (HSE, Moscow)

5. *CHILDES*: Practical Applications

Ratner & MacWhinney (2016): *Language Sample Analysis*

Language Sample Analysis (LSA) supplements standardized assessment

- A speech-language pathologist collects short language samples from children for clinical practice, transcribes and analyzes child output.
- 2 projects:
  1. KIDEVAL: n=125 diads, evaluated at 7, 10, 11, 18, and 24 months: 1,250, with 15-30 min transcripts (kideval +leng +t*CHI)
  2. *CHILDES* and *CLAN* in support for clinical practice:
     - Increase the number of languages and adapt automatic *mor/gra* utilities
     - Use archived *CHILDES* data to improve LSA outcome measures.

Ratner, N. B., & MacWhinney, B. (2016). Your laptop to the rescue: Using the *CHILDES* archive and CLAN utilities to improve child Language Sample Analysis. *Seminars in Speech and Language, 37*(2), 74-84.

Resources

1. *CLAN* Manual
2. Brian MacWhinney's 2019 Symposium: June 6-8, 2019, at Carnegie Mellon University (Pittsburgh, PA)
3. Screencast tutorials for *CHILDES*
4. References