



Introduction to corpus research in language acquisition

Sophia A. Malamud

Brandeis University

23 June, 2020

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

Some relevant highlights from my research

1999-2006 (UPenn) - Using plain text written (parallel) corpora; annotation

- *Russian word order; passives & impersonals in Russian, English, Italian, German*

2008-now (Brandeis) - Using grammatically annotated / spoken corpora

- *Meaning & use in Russian, Ancient Greek (with Irina Osadcha & Jana Beck), parallel spoken data in English, Mandarin (with Tatjana Scheffler)*

2009-2010 (Brandeis) - First major foray into acquisition

- *Combining Irina Dubinina's experimental study with corpus controls in investigating requests in adult heritage speakers of Russian*

2013-now (Brandeis) - Building a large grammatically annotated corpus

- *Bilingual Russian child and child-directed speech (BiRCh)*





What is a corpus?



What is a corpus?

- corpus - Latin (“body”), plural corpora
- In linguistics/language studies, a structured collection of texts

“A corpus is a collection of

1. machine-readable
2. authentic texts (including transcripts of spoken data) which is
3. sampled to be
4. representative of a particular language or language variety.”

McEnery et al. (2006): 5



What is a corpus?

“A corpus is a collection of

- machine-readable
 - key to reliable searching, amenable to annotation to create more structure, statistics, sharing
- authentic texts (including transcripts of spoken data)
 - key to obtaining true observational data, finding unexpected examples and data for new hypotheses
- *sampled to be representative of a particular language or language variety.*
 - *designed from a linguistic perspective*



What is a corpus?

“A corpus is a collection of

- machine-readable
- authentic texts (including transcripts of spoken data) which is
- *sampled to be representative of a particular language or language variety.*”


- Russian National Corpus (НКРЯ): sampled to be representative
- The Wampanoag Corpus: all documents
 - 2 King James Bible translations, hundreds of personal letters, wills, deeds, and land transactions written in Wôpanâôt8âôt.
- Brown 1973 corpus:
 - 3 children, one in each subcorpus, organised by age



What may not be a corpus

- Observational data that is not machine-readable
 - Sampling reliability can vary: can't go back and check
 - Not machine-readable, but can be digitised
 - Context and structure key to usefulness
- A text archive
 - may not be representative, but could be (e.g. entire oeuvre of a writer, all documents available for a sleeping language, translational corpora)
- The internet is not a corpus
 - But there are corpora sampled from the internet
 - E.g., Taiga https://tatianashavrina.github.io/taiga_site/





Observational data in the history of child language acquisition research

Language Acquisition research

Two basic complementary methodologies

- Experimental methods
 - Controlled contexts: more certainty about causation
 - Used to test hypotheses
 - Mostly need to be conducted in-person
- Corpus methods
 - Large range of linguistic behaviour
 - Natural contexts for spontaneous production
 - Used to study the development of behaviour, discover new patterns
 - More leeway for remote study



Observational data / corpora in Language Acquisition research

- Longitudinal: (usually) fewer people, observed over more time
 - Systematic diary studies
 - Transcripts of audio data
 - Multimedia corpora
- Cross-sectional studies: more people, few observations



Longitudinal observations in Language Acquisition research

- Systematic diary studies
 - Comprehensive diaries: great up to 2-word stage
 - Jaeger 1985
 - Clara & William Stern 1907
 - Leopold 1939-1940
 - Topical diaries
 - Useful for lower-frequency phenomena,
 - e.g. passives
 - CDI: MacArthur-Bates Communicative Development Inventories (Fenson et al.1993)
 - vocabulary, early grammar
 - longitudinal or cross-sectional



Longitudinal observations in Language Acquisition research

- Multimedia corpora: audio/video and transcripts
 - Brown 1973
 - recorded data
 - reliability checks
 - different socioeconomic backgrounds
 - qualitative and quantitative data from 3 children
 - MLU developed as assessment method
 - Morpheme order/productivity



Longitudinal observations in Language Acquisition research

- Multimedia corpora: audio/video and transcripts
 - Eleanor Ochs 1979: need transcription conventions
 - linguistic information
 - e.g. spelling of fillers
 - contextual information
 - e.g., turn-taking



Longitudinal observations in Language Acquisition research

- Multimedia corpora: audio/video and transcripts
 - 1983 Catherine Snow and Brian MacWhinney started to implement many of these suggestions in CHAT (Codes for Human Analysis of Transcript)
 - CHILDES database - Child Language Data Exchange System (cf. MacWhinney and Snow 1985).
 - Over 200 corpora
 - Over 30 languages
 - CLAN - computer programs for analysis



A note on cross-sectional observations in Language Acquisition research

- Behaviorist studies in 1930s-1950s
 - 70 to 430 children
 - 50 sentences to 6 hours per child
- CDI
- Cross-sectional corpora
 - Transcripts of narratives, e.g. frog stories

<https://chilides.talkbank.org/browser/index.php?url=Frogs/>



Goals of language acquisition research:

- what is being learned
 - corpus: direct evidence of the input = parents' speech,
 - indirect evidence for underlying grammar
- when it is being learned
 - developmental trajectory: corpus gives direct evidence of production,
 - only hints for comprehension
- how it is being learned
 - corpus provides indirect evidence based on properties of input and
 - output over time





Example of a language acquisition corpus:

Bilingual Russian Child and Child-directed Speech Corpus
BiRCh

Introducing BiRCh: what is a native speaker?

- No unusual accent
 - dialectal variation
- Extensive vocabulary
 - education levels
- Background cultural knowledge
- Socio-cultural norms
- Idiomatic & formulaic expressions
- Sense of “ownership” of the language



Introducing BiRCh: what is a native speaker?

- Unconscious knowledge of grammar
 - “little words”, endings, etc.
- Sensitivity toward ambiguities & multiple meanings
 - *I saw an owl with a telescope*
- Grammaticality judgements
 - *Colorless green ideas sleep furiously*
- Ability to construct complex sentences with multiple embeddings
 - *The dinosaur that was drawn by my friend that I had a playdate with yesterday was a T-Rex*



Introducing BiRCh: bilingualism

- Simultaneous vs. consecutive bilingual
- Balanced vs. unbalanced & stable vs. unstable bilingual
- Dominant vs. weak language (highly complex issue!)
- Majority vs. minority language



Introducing BiRCh: immigrant bilingualism

- Immigrant parents in our corpus are bilinguals: Russian is their chronologically first and dominant language
- Language past the critical period for acquisition is considered stable, but...
 - influence of L2 can change verbal behaviour even if it doesn't change underlying knowledge
 - attrition and restructuring of the underlying knowledge is also possible



Introducing BiRCh: heritage bilingualism

- Input to immigrants' children is different from monolingual peers
- Community at large is missing
- Schooling = a system of checks and balances
 - Drastic reduction in input \Rightarrow form-meaning connection suffers
 - Children learn about the world through the majority language \Rightarrow impoverished vocabulary
- What happens if childhood bilingualism involving a minority and a majority language stops developing and social factors start privileging the majority language?



Introducing BiRCh: heritage bilingualism

- Heritage speakers (HS)
 - those who have been exposed to a particular language in childhood,
 - and are to some degree bilingual in it,
 - but did not learn it to full capacity –
 - because another language became dominant



Introducing BiRCh: heritage bilingualism

Quantitatively

- Monolingual child: about 5000 utterances per day, over a million in a year
- College language class: about 400 per day?
- Heritage child:
 - initial exposure similar to monolingual
 - diminishing after “point of interruption”, sometimes to near-nothing
 - Total hours of HL exposure are highly variable from person to person



Introducing BiRCh: heritage bilingualism

Qualitatively:

- Is the input at home the same for bilingual and monolingual families?
 - Presence of majority language in the home
 - Attrition and transfer from the majority language in immigrant parents' speech
 - Changes in the homeland that immigrants do not participate in
- How many different genres, registers, topics, situations are part of the input?



Heritage English

Tamarine Tamasugarn

Okay, everybody always thought like I grown up in States, but actually no. I was born in States, and when I was four I moved back to Thailand with parents and I grown up in Thailand. So I definitely Thai. Everything, the culture, everything Thai. But I also know also American culture also because part of my family also in L.A.

So I learn language and, you know, how, maybe you can tell from my speak. But I think it's great to know both of culture and, you know, adjust in your life and bring all the good stuff on each culture to improve your life and make your life happy. So I think that's a very good to learn for both culture, yeah. *(2009 YouTube interview, from M.Polinsky 2018)*



Heritage English: some observations



- High fluency
- Problems with morphology
- Missing (& a few extra) “grammatical” words
- Redundancies & repetition
- Short utterances, no embeddings
- Word order different from baseline
- *I grown up in States*
- *with parents*
- *I definitely Thai.*
- *I also know also American culture also because part of my family also in L.A.*
- *maybe you can tell from my speak.*
- *it's great to know both of culture*
- *that's a very good to learn for both culture*

Introducing BiRCh: causes of observed divergences

- Incomplete acquisition:
 - Fossilization of child language (HS fail to learn some structures)
- Attrition over time
 - Forgetting (older HS children know less than younger ones)
- Transfer:
 - Influence of L2 on the structures of L1 (HS with different majority languages have different grammars)
- Language-internal restructuring
 - General linguistic or cognitive mechanisms cause incompletely acquired grammar to take a different path
- Divergent input
 - Bilingual parents' speech is different from monolingual ones

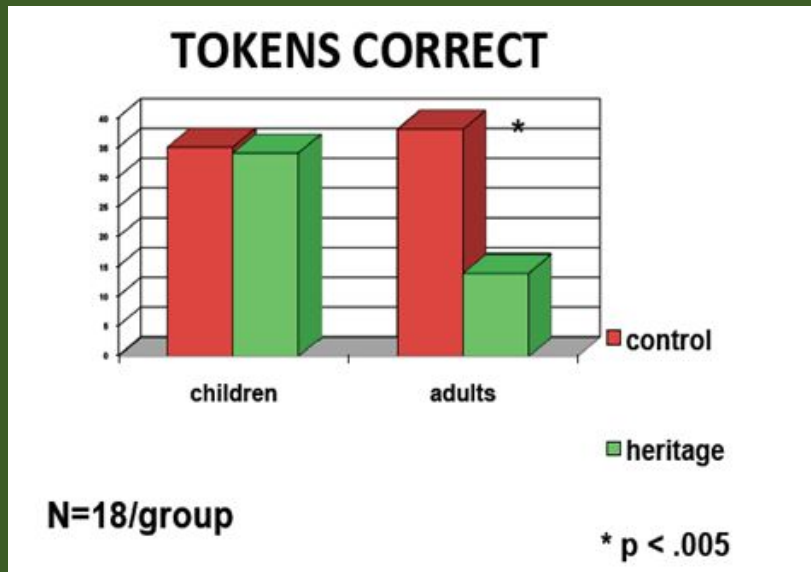


Introducing BiRCh: previous experimental results

- Adult HS are not fossilized children
- Adult HS are not children who simply forgot their HL

So, we conclude that adult HS' divergent attainment results from

- divergent input and/or incomplete acquisition and/or attrition
- followed by transfer and/or internal restructuring



Maria Polinsky (2011) "REANALYSIS IN ADULT HERITAGE LANGUAGE: New Evidence in Support of Attrition" *Studies in Second Language Acquisition*: 33, 305– 328.

Introducing BiRCh: why study heritage speakers?

“Bilingualism is for me the fundamental problem of linguistics” (Jacobson 1953)

Insight into the questions:

- What defines a native speaker?
- What does it take to acquire a (specific aspect of a) language?
- Is all linguistic knowledge systematic?
- How robust is a speaker’s knowledge of various components of L1 grammar?
- What grammatical options emerge in the absence of input or with impoverished input?

Heritage learners “provide a crucial missing link between competent L1 learners, balanced bilinguals, and possibly L2 learners” (Polinsky 2008g)

Heritage bilingualism: non-immigrant linguistic minorities

The number of native speakers of most languages is declining

Language documentation perspective:

- understanding features shared by heritage languages,
- and recognizing heritage language acquisition situations
- is crucial for scholars working on understudied and endangered languages.

The data on endangered languages often comes from bilingual consultants who may be heritage speakers of the language of interest. “How much does [the speech of the last speakers of a language] reveal of the original structure [of the native grammar]?”

(Sasse, 1992, p. 76).



Heritage bilingualism: non-immigrant linguistic minorities

- Input to indigenous linguistic minority children is different from monolingual peers
- Community at large is missing
- Schooling = a system of checks and balances
 - Drastic reduction in input \Rightarrow form-meaning connection suffers
 - Children learn about the world through the majority language \Rightarrow impoverished vocabulary
- What happens if childhood bilingualism involving a minority and a majority language stops developing and social factors start privileging the majority language?



Heritage bilingualism: non-immigrant linguistic minorities

- There are some important (and understudied) differences between immigrant bilinguals and indigenous linguistic minority bilinguals
- One is a social factor favouring minority language maintenance: its cultural importance
 - Whether an immigrant Russian child keeps speaking Russian has no effect whatsoever on the development and continued existence of Russian language and culture
 - In contrast, language maintenance has great importance for indigenous communities



Practical necessities in studying heritage speakers

Scontras et al. 2015

- establishment of a clear native baseline
 - to compare acquisition of, e.g., Russian HL with acquisition of Russian
- determination of the input to heritage language acquisition by documenting the language of the émigré parents
 - to locate the potential source of reanalysis and differences from the language in the homeland
- determination of bilingual child language behavior
 - to test for variation, change, and attrition over the lifespan



Corpus of Bilingual Russian Child and child-directed speech (BiRCh)



"Audio-aligned Longitudinal Corpus of Bilingual Russian Child and Child-directed Speech (BiRCh Longitudinal)"
Dubinina, Malamud & Denisova-Schmidt



Brandeis



University of St.Gallen



Corpus of Bilingual Russian Child and child-directed speech (BiRCh)



"Audio-aligned Longitudinal Corpus of Bilingual Russian Child and Child-directed Speech (BiRCh Longitudinal)"
Dubinina, Malamud & Denisova-Schmidt

Design and goals:

- Parents record interactions: data collection over a decade
- Families in Russia & Ukraine; USA and Canada; Germany
 - Monolingual families as approximate dialectal matches for bilingual families
 - Bilingual parents immigrated after age 13
 - Children before age of 4 at start of study
- Sociolinguistic information
 - Utrecht Bilingual Language Exposure Calculator (Unsworth 2015)



Corpus of Bilingual Russian Child and child-directed speech (BiRCh)



Estimating input: some results (Utrecht BiLEC Questionnaire)

- All parents are extremely motivated for language maintenance
- Circumstances differ:
 - How much daycare / language(s) spoken in daycare
 - Extended family, family size
 - Friends, games, cartoons, books, activities

Child	Age	Russian input quantity	Russian input quality
B (USA)	4yr 9mo	2yr 9mo	2.5/5
I (USA)	4yr 9mo	4yr 5mo	2.98/5
M (Germany)	4yr 8mo	4yr 2 mo (4.22)	5/5
N (USA)	3yr 10mo	3yr 10mo	5/5



Corpus of Bilingual Russian Child and child-directed speech (BiRCh)



"Audio-aligned Longitudinal Corpus of Bilingual Russian Child and Child-directed Speech (BiRCh Longitudinal)"
Dubinina, Malamud & Denisova-Schmidt

- Bilingual groups: “best-case” scenario for language maintenance
 - Parents (and older siblings, if any) use Russian
 - Parents highly motivated
 - Educated middle-class
 - repeating biases of many CHILDES corpora
 - minimizing differences between monolingual & bilingual families
- Resulting differences in acquisition trajectory, input, and output will be easier to attribute to presence of majority language



Corpus of Bilingual Russian Child and child-directed speech (BiRCh)



"Parsed and Audio-aligned Corpus of Bilingual Russian Child and Child-directed Speech (Parsed BiRCh)" (NSF #1651083)

Sophia A. Malamud, Irina Dubinina, Pavel Koval, Nianwen Xue & Alex Luu



Brandeis



UConn



Brandeis



Brandeis



Consulting on disfluency annotation, segmentation, parsing
Beatrice Santorini, UPenn, Parser-in-Chief of Penn Parsed Corpora



Corpus of Bilingual Russian Child and child-directed speech (BiRCh)



"Parsed and Audio-aligned Corpus of Bilingual Russian Child and Child-directed Speech (Parsed BiRCh)" (NSF #1651083) Malamud, Dubinina, Koval, Xue & LU'u

1-million word "pilot" corpus of Russian in open web access

1. Pseudonymized recordings
 - Names, addresses, etc. replaced by silence to protect privacy
2. Text-searchable transcripts time-aligned with the speech signal
 - Names etc. replaced by pseudonyms to protect privacy
 - Searching the text of the transcripts allows jumping to the corresponding portion of the audio recording
 - For researchers, educators, & parents who are interested in lexicon, prosody, etc.
3. A morphologically tagged and parsed version of the transcripts



Corpus of Bilingual Russian Child and child-directed speech (BiRCh)



"Parsed and Audio-aligned Corpus of Bilingual Russian Child and Child-directed Speech (Parsed BiRCh)" (NSF #1651083) Malamud, Dubinina, Koval, Xue & Lʹʹu

1-million word “pilot” corpus of Russian in open web access

1. Pseudonymized recordings
2. Text-searchable transcripts time-aligned with the speech signal
3. A morphologically tagged and parsed version of the transcripts
 - Also linked to audio
 - Unprecedented access to the grammatical properties of the children’s and parents’ speech



Corpus of Bilingual Russian Child and child-directed speech (BiRCh)



- Why corpus? (and not experiment)
 - Multi-purpose, multi-use, gives a fuller picture of language use
- Why parsed? (and not just morphologically tagged)
 - Detailed information about linguistic structure allows us to study grammatical development and the role of parents' speech.
 - Statistical, distributional information on frequencies of constructions allows us to study speech varieties with a lot of grammatical variation and change.
 - Parsing the data allows us to retrieve all instances of a given construction in the entire corpus.



Corpus of Bilingual Russian Child and child-directed speech (BiRCh)



Where we are: <http://birch.ling.brandeis.edu/>

- Data collection since 2013
 - 3-4 families in each group
 - Children aged 1;6 - 4;6 at start of recording (average age 35 mo)
1. 1,677,000 words collected (466.4 hours of audio)
 2. 726,060 words transcribed, annotated for disfluencies, and checked
 3. 201,508 words morphologically annotated



Corpus of Bilingual Russian Child and child-directed speech (BiRCh)



Where we are: <http://birch.ling.brandeis.edu/>

- Annotation guidelines are all available on the project website
- We are developing parsing annotation guidelines, scripts
- We are planning to publish a morphologically annotated subcorpus in CHILDES later this year
 - eventually the full 1-million word corpus, including audio, transcripts, and morphological tags (but not the syntactic annotations) will be added to CHILDES
- The entire Parsed BiRCh will be available in open access



Corpus of Bilingual Russian Child and child-directed speech (BiRCh)



Data: audio-aligned disfluency-marked transcripts

- BiRCh Corpus is part of the umbrella group of Penn-style Audio-Aligned Corpora, itself part of Penn-style Parsed Corpora
 - Corpus of Appalachian English (AAPCApE)
 - Corpus of NYC English
 - Corpus of AAVE
- Comparable data
- No need to reinvent the wheel, just adopt for current needs
 - Disfluency annotation: adopted from AAPCApE
 - Segmentation (sentence tokenization): adopted from Penn-Helsinki Historical Corpora (Penn Parsed Corpora)
 - Morphological annotation: adopted from Mystem/НКРЯ, and UD



Corpus of Bilingual Russian Child and child-directed speech (BiRCh)



Data: audio-aligned disfluency-marked transcripts

- Disfluency annotation:
 - Initially performed by transcriber
 - Corrected by annotation/transcription checker
 - Corrected by segmentation checker
- Segmentation (sentence tokenization)
 - Initially performed by transcriber
 - Corrected by segmentation checker



Corpus of Bilingual Russian Child and child-directed speech (BiRCh)



Data: audio-aligned disfluency-marked transcripts

- Disfluency and bilingualism-specific annotation:
 - False-starts
 - Elaborations and parenthetical clauses
 - Self-interruptions
 - Noise, coughing, laughing, singing
 - Code-switching and borrowings
- Segmentation (sentence tokenization)
 - 1 segment = 1 main clause and all of its subordinate clauses
 - fragments



ELAN: transcription, disfluency annotation, segmentation into sentences

The screenshot displays the ELAN 4.9.4 software interface. At the top, the title bar reads "ELAN 4.9.4 - L_2016_05_01_2_ed.eaf". Below this is a menu bar with options: File, Edit, Annotation, Tier, Type, Search, View, Options, Window, Help. A secondary menu bar contains tabs for Grid, Text, Subtitles, Lexicon, Comments, Recognizers, Metadata, and Controls. The main area is divided into two sections: "Volume:" and "Rate:". Each section has a numerical input field (both set to 100) and a horizontal slider. The volume section also includes a "Mute" checkbox and a "Solo" radio button. Below these controls is a playback control bar with a time display of "00:01:35.030" and a "Selection" range of "00:01:23.337 - 00:01:24.630 1293". The playback bar includes standard navigation buttons (stop, previous, play/pause, next, fast back, fast forward) and checkboxes for "Selection Mode" and "Loop Mode". The bottom portion of the interface shows a timeline with a red cursor at approximately 00:01:24.5. Below the timeline are three transcription tiers: "default" (red text), "Ребенок [170]" (blue text), and "Мама [181]" (green text). The "Папа [0]" tier is also visible but empty. The transcription text includes: "Очень хороший цвето", "Но <PAREN> ты знаешь <\$\$PAR", "Он пушисте", "Он скоро о", "Мне кажется лучше на него подуть чтобы у нас выросло по", and "Нет.".

Corpus of Bilingual Russian Child and child-directed speech (BiRCh)



In progress: morphological annotation

1. Annotation guidelines (Dubinina & Malamud 2019)
2. Preparing ELAN files for automatic tagging (Alex Lʙu)
3. Automatic tagging with Mystem (Alex Lʙu)
4. Post-processing to get files closer to BiRCh guidelines (Alex Lʙu)
5. Initial correction by an annotator (Olga Shtan')
6. Checking by second annotator (Olga Ivchenko)



Corpus of Bilingual Russian Child and child-directed speech (BiRCh)



In progress: morphological annotation

Mother to 2-year-old participant:

Скоро будем кушать, подожди

Skoro budem kušat', podoždi

Soon will.be.1PL eat.INF wait.2SG.PFV.IMPER

“We will eat soon, wait”



Corpus of Bilingual Russian Child and child-directed speech (BiRCh)



In progress: morphological annotation

Soon Adverb

LEMMA= скоро / skoro / “soon”

will.be Verb intransitive imperfective decl nonpast plural 1st

LEMMA= быть / byt' / “be”

eat Verb transitive imperfective infinitive

LEMMA= кушать / kušat' / “eat”

wait Verb transitive perfective imperative singular 2nd

LEMMA= подождать / podoždat' / “wait”



FoLiA Linguistic Annotation Tool (FLAT): morphological correction

МАМА: {CG} .
МАМА: А мы какой
РЕБЕНОК: <\$PAREN
РЕБЕНОК: <\$PAREN
РЕБЕНОК: Смотри э
РЕБЕНОК: Амм вот
МАМА: А как он наз
МАМА: Просто назв
РЕБЕНОК: Да .
МАМА: Да ?
РЕБЕНОК: Да .
РЕБЕНОК: Амм назв
МАМА: Винтовой са
РЕБЕНОК: Да .
МАМА: Здорово .
РЕБЕНОК: Простой
МАМА: Беня , фу .
МАМА: Слезь .
РЕБЕНОК: Что ?
МАМА: Да Беня , ви
МАМА: Она думает
РЕБЕНОК: {CR} .
МАМА: Ей интерес
РЕБЕНОК: Беня , мы не един

Annotation Editor

a2.w.3

Text Select span>

<https://raw.githubusercontent.com>

Lemma +↓

<https://raw.githubusercontent.com> *confidence: (not set)*

Part-of-Speech +↓

<https://raw.githubusercontent.com> *confidence: (not set)*

Description:

Queue for later submission

Repeat this annotation for the next target

Open console window after submission

Ok

ртинка .

Corpus of Bilingual Russian Child and child-directed speech (BiRCh)



In progress: parsing guidelines and scripts

Parsing team (besides myself)

- Pavel Koval (UConn, team lead)
- Alex LUu (Brandeis)
- Benjamin Rozonoyer (Brandeis)
- Ruth Rosenblum (Brandeis)
- Nianwen Xue (Brandeis, consulting faculty)





Introduction to corpus-building



To build a corpus: data collection depends on questions you are trying to answer

- **Sampling context, rate, schedule, and duration:**
 - Type of data:
 - tense, event sequencing (\Rightarrow narratives) vs tag questions, imperatives (\Rightarrow interactive conversation)
 - audio vs video
 - additional information (meta-data, interviews)
- **Multimedia corpus:**
 - balancing privacy concerns with research goals



To build a corpus: data collection depends on questions you are trying to answer

- **Density:**
 - how frequent is the phenomenon you're trying to study?
 - Verb acquisition vs passives
 - **Sampling regime**
 - typical = 20-60 minutes a week
 - dense = 5-10 hours a week
- **Timing/duration:**
 - Is it tied to a specific developmental stage?
 - morphological overgeneralisation errors - "I see-ed two sheeps"
 - point of interruption
- **Total amount:**
 - How much data do you need for statistical analysis?



To build a corpus: data collection depends on logistical conditions you are dealing with

- Recording quality
 - if at all possible, use wav rather than mp3 recording (mp3 distorts data for phonetic analysis)
 - but wav recorders tend to be more expensive
- Access to speakers - community-based research
 - Texas Spanish corpus: bilingual students record their interviews with members of their community
 - BiRCh: parents record their interaction with children
 - This in part dictates contexts for sampling - whatever is easiest for parents
- Dense recordings
 - LENA is special-purpose system for collecting dense data
 - But we were collecting dense data from 1 child each in 3 countries, so not practical or cost-effective to use LENA



To build a corpus: data collection, annotation, and sharing to maximise usefulness & reusability

- Our initial research goals: syntax-discourse interface (the use of пожалуйста “please”; passives & impersonals)
 - But we have large data collection of audio recordings
 - so quality and format of audio is important to enable future phonological and phonetic studies (e.g. prosody, vowel quality)
 - once we realised the quality issue, we moved to wav
- But acquisition of Russian, a morphologically rich language, involves many questions about morphology
- so we are conducting a more in-depth morphological annotation than just the minimum necessary for syntactic parsing



To build a corpus: data collection, annotation, and sharing to maximise usefulness & reusability

- If at all possible, publish data in established databases
 - For publication in CHILDES, data format and annotation has to be CHAT-compatible
- If you can share audio/video, do so



To build a corpus: data annotation

- Annotation = judgements
 - your annotation is only as good as the annotators' judgements.
 - Reliability check: inter-annotator agreement
- Approaches to ensuring reliability:
 - each item is annotated by a single annotator, with random checks by a second annotator
 - some of the items are annotated by two or more annotators
 - each item is annotated by two or more annotators - followed by reconciliation
 - each item is annotated by two or more annotators - followed by final decision by superannotator (expert)
- In all cases, measure of reliability: coefficients of agreement
 - e.g. Cohen's Kappa (Cohen, 1960)



To build a corpus: data annotation


- Turning annotation into a small experiment (cf. Pustejovsky & Stubbs 2013):
 - start with a small amount of data
 - model: create annotation guidelines that describe the phenomena you're going to study
 - annotate - this also helps refine tools and procedures for annotation
 - evaluate: check inter-annotator agreement, quality of annotation
 - revise annotation guidelines, repeat the model-annotates-evaluate process
 - expand to more data once the annotation is reliable.
- Turning annotators into experts
 - ability to detect patterns, intuitions sharpen with practice
 - continual training and discussion essential



To build a corpus: data annotation

- Annotation standards:
 - CHAT: transcription and annotation in e.g., CHILDES
 - ISO and other international standards for annotating specific types of information
 - Universal Dependencies (UD, <https://universaldependencies.org/>)
- What free tools are available for transcription and annotation?
 - ELAN for multi-media corpora
 - FLAT/FoLiA
 - See annotation and analysis tools at Universal Dependencies
- Tools for automating some or all of the annotation tasks
 - Russian morphology - e.g., MyStem
 - Syntax and morphology in UD, more computationally savvy: <http://ufal.mff.cuni.cz/udpipe>





Example of corpus use to study language acquisition

Properties of input to acquisition:
fillers in the speech of Russian parents

How to use a corpus: fillers in émigré Russian

Data in this study: a subset of audio-aligned disfluency-marked transcripts

- Disfluency and bilingualism-specific annotation:
 - False-starts
 - Elaborations and parenthetical clauses
 - Self-interruptions
 - Noise, coughing, laughing, singing
 - Code-switching and borrowings
- Segmentation (sentence tokenization)
 - 1 segment = 1 main clause and all of its subordinate clauses
 - fragments



How to use a corpus: fillers in émigré Russian

- Pause fillers in Russian:
 - *aa, mm, èè* - most frequent non-lexemic fillers
 - *èto, èto samoe, vot, nu, kak by* - frequent lexemic fillers
- Differences between Russian, English, and German (Candea et al. 2005, Fox et al. 2010, a.o.)
 - *èto* “this” is a pause filler in Russian, but not in English
 - *um* is a pause filler in English, but not in Russian
 - German *äh* is similar to Russian *èè*; but *ähm* & *also* are not



How to use a corpus: fillers in émigré Russian

Background:

- Original view of pause fillers and disfluencies
 - performance errors (Chomsky 1965, 1972; Seidenberg 1997)
- More recent work
 - pause fillers “facilitate both the production and perception” (Erker & Brusco 2017),
 - distribution of disfluencies is subject to some of the same constraints as other linguistic phenomena (Bell et al. 2003; Erker & Brusco 2017; Ginzburg et al. 2014; Swerts 1998, inter alia).



How to use a corpus: fillers in émigré Russian

- Are pause fillers a symptom or a signal of speaker's mental state?
 - Lexemic fillers are generally thought to be signals
 - pause filler use - just one of several word meanings/discourse functions
- Are non-lexemic fillers different?
 - We don't say that *èto* is not a word when it's used to fill pauses
 - Is *aa* a word?



How to use a corpus: fillers in émigré Russian

- Starting point: non-lexical pause fillers *aa* & *mm*
- Functionally somewhat similar to English *um*, *uh*

(1) Èto nužno položít' v aa v musor

This need put.INF in aa in garbage

'This goes into uh into the garbage'

(2) Dva dva mm na dva kružka prygnul

Two two mm over two circles.DIM jumped

'Two two um he jumped over two circles'



How to use a corpus: fillers in émigré Russian

- Other uses: nothing to do with pause-filling

(3) Ma, ty golodnaja? Aa? - Mm? - Ty golodnaja? Call for addressee
Ma, you hungry? Aa? Mm? you hungry? engagement
'Ma, are you hungry? Eh?' - 'Eh?' - 'Are you hungry?'

(4) Mm! Belka idët - Aa, ponjatno!
Signal of

mm squirrel goes.3SG Aa clear information
receipt

'Oh, a squirrel!' - 'Oh, I got it now!' or commitment



How to use a corpus: fillers in émigré Russian

- What are *aa* & *mm*?
 - Symptoms of processing difficulties?
 - Words & part of core native speaker knowledge?
- Do bilinguals, and specifically those in US and Germany, use *aa* & *mm* in the same way as monolinguals?
 - Effect of bilingualism generally?
 - Effect of English and German, respectively?



How to use a corpus: fillers in émigré Russian

- What are *aa* & *mm*? BOTH
 - Symptoms of processing difficulties? YES
 - Words & part of core native speaker knowledge? YES
- Do bilinguals, and specifically those in US and Germany, use *aa* & *mm* in the same way as monolinguals? NO
 - Effect of bilingualism generally? NOT CLEAR
 - Effect of English and German, respectively? NOT CLEAR



How to use a corpus: fillers in émigré Russian

- BiRCh data used:
 - Russia - 47,413 words
 - Germany - 22,493 words
 - US - 19,865 words
 - automatically extracted all examples from child-directed speech
- Informal spoken subcorpus of the Russian National Corpus
 - randomly selected 100 examples each of *aa* & *mm*



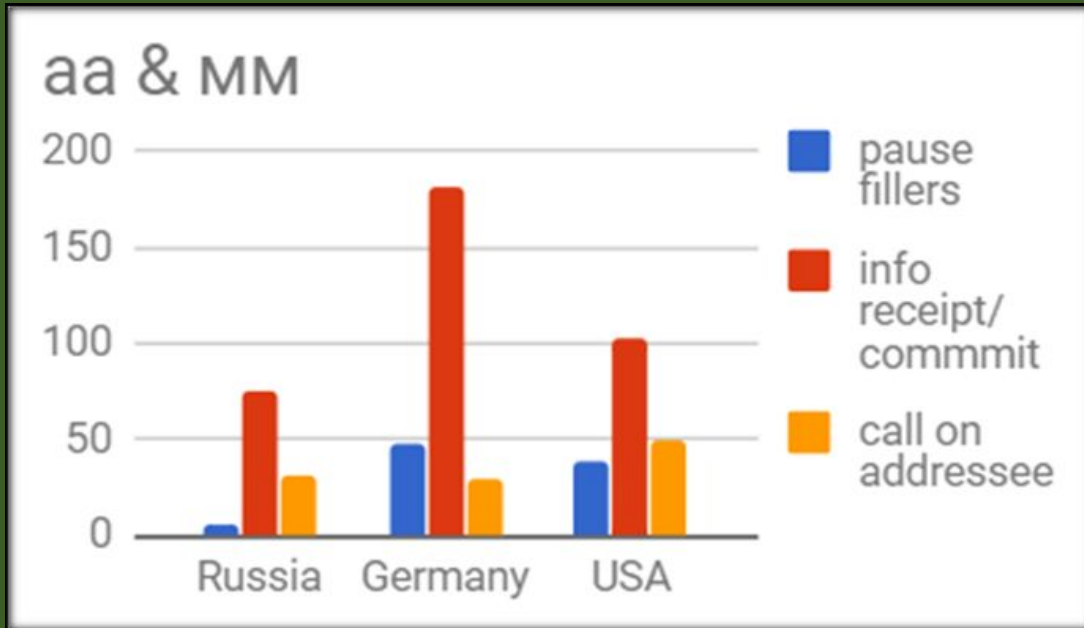
How to use a corpus: fillers in émigré Russian

Classifying examples: further annotation in Excel

- 2 annotators
- 4-person team adjudicating
- Various discourse functions grouped into three broad categories:
 1. pause filler
 2. signal of information receipt or commitment
 3. call for addressee engagement

How to use a corpus: fillers in émigré Russian

Rates of use for *aa* & *mm* by function and population



How to use a corpus: fillers in émigré Russian

Mixed effects regression

- Using speakers as random effect
(to factor out individual variation)

Pause fillers vs. not pause fillers

- Country is the significant predictor ($p \ll 0.01$)
 - Bilingual parents use significantly more pause fillers than monolinguals
- Filler type (*aa* vs. *mm*) is not ($p=0.07$)
 - when speakers are filling pauses, it doesn't matter what sound they use



How to use a corpus: fillers in émigré Russian

Mixed effects regression

- Using speakers as random effect
(to factor out individual variation)

Non-pause-fillers: call on addressee vs. information signal

- filler type (*aa* vs *mm*) is the significant predictor ($p \ll 0.01$)
 - After excluding pause-filler uses from consideration:
 - Among non-pause-filler *aa* tokens, 58% are information signals
 - Among non-pause-filler *mm* tokens, 42% are information signals
 - *aa* & *mm* behave like words
- country is not a predictor ($p=0.05$)
 - Bilingual and monolingual parents use these words in similar ways



How to use a corpus: fillers in émigré Russian

Mixed effects regression

- Using speakers as random effect
(to factor out individual variation)

Comparing bilingual parents in US & Germany:

- Pause fillers vs non-fillers:
 - country is not a predictor ($p=0.95$)
 - filler type is borderline predictor ($p=0.038$)
- call on addressee vs information signal:
 - no significance: $p=0.13$ for filler type; $p=0.155$ for country
- More individual variation makes everything non-significant:
 - need more speakers, more data



How to use a corpus: fillers in émigré Russian

Conclusions: As pause fillers, *aa* & *mm* signal processing difficulties

- Russian National Corpus (НКРЯ) data (monolingual):
predominant function of *aa* & *mm* is pause filling
 - Unlike monolingual parents in BiRCh
 - НКРЯ data is not child-directed
 - plausibly more cognitive load/ required vocabulary/more face-work
- Bilingual parents plausibly have higher processing load
 - due to activation of other language
 - possibly lower activation levels making lexical retrieval more costly
 - possibly lower active vocabulary
 - possibly influence of pause filler use in majority language

How to use a corpus: fillers in émigré Russian

Conclusions:

As non-pause fillers *aa* & *mm* behave like different words

- Bilinguals maintain their use in monolingual-like ways
- If *aa* & *mm* are words when not filling pauses, arguably they are words even on their pause-filling uses
 - their meanings as signals of processing difficulty are more similar to each other than their meanings in other uses



How to use a corpus: fillers in émigré Russian

- Analysing data in Excel or other spreadsheet program: pivot tables, counts, graphs, and statistical tests such as chi-square
- If you can, analysing data with more sophisticated methods to ensure reliability of conclusions: mixed-effects regression (takes care of individual variation between speakers)





Спасибо!

Thank you!

