D.O.I. http://dx.doi.org/10.1590/2176-4573e55831

ARTICLES

Creating a Large-Scale Audio-Aligned Parsed Corpus of Bilingual Russian Child and Child-Directed Speech (BiRCh): Challenges, Solutions, and Implications for Research / *A construção de* corpus *de larga escala da fala bilíngue de crianças e da fala bilíngue dirigida à criança, anotado e alinhado aos arquivos de áudio: desafios, soluções e implicações para a pesquisa*

Alex Luu [*]
Pasha Koval [**]
Sophia A. Malamud [***]
Irina Y. Dubinina [****]

ABSTRACT
The BiRCh Project (The Corpus of Bilingual Russian Child Speech) involves collecting a longitudinal audio *corpus* of Russian spoken by children and their families in Russia, Ukraine, Germany, the U.S., and Canada. We are building a large-scale corpus based on a subset of this data, the "Parsed and Audio-aligned Corpus of Bilingual Russian Child and Child-directed Speech (BiRCh)" with two basic components: (1) 1-million-word transcripts which are time-aligned with the audio speech signal and fully text-searchable, and (2) a 500K-word morphologically annotated and parsed portion of the transcripts, also audio-aligned. We are using this corpus to investigate various phenomena in the linguistic input and the developmental trajectory of heritage bilinguals, e.g., case, gender, passives, impersonals, politeness markers, disfluencies, and discourse markers. This article focuses on the challenges and solutions of the BiRCh development and the implications for research on the richly annotated data provided by the corpus.
KEYWORDS: Spoken Russian corpus; Disfluency annotation; Morphological tagging; Syntactic parsing; Bilingual and heritage speakers

RESUMO
*O projeto BiRCh (*The Corpus of Bilingual Russian Child Speech, Corpus *de fala de crianças bilíngues em russo) envolve a construção de um* corpus *longitudinal composto de gravações de fala em russo produzida por crianças e suas famílias na Rússia, Ucrânia, Alemanha, EUA e Canadá. Estamos construindo um* corpus *de larga escala com base no conjunto dessas gravações, o* 'Parsed and Audio-aligned Corpus of

[*] Brandeis University, Michtom School of Computer Science, Waltham, Massachusetts, USA; https://orcid.org/0000-0003-1393-6791; alexluu@brandeis.edu
[**] New York University Abu Dhabi, Program in Psychology, Abu Dhabi, United Arab Emirates; https://orcid.org/0000-0002-5597-0587; pasha.koval@nyu.edu
[***] Brandeis University, Michtom School of Computer Science, the Linguistics Program, Waltham, Massachusetts, USA; https://orcid.org/0000-0002-1321-7685; smalamud@brandeis.edu
[****] Brandeis University, Department of German, Russian and Asian Languages and Literature, Waltham, Massachusetts, USA; https://orcid.org/0000-0001-9960-3271; idubinin@brandeis.edu

Bilingual Russian Child and Child-directed Speech (BiRCh)', *com os dois componentes básicos: (1) as transcrições de um milhão de palavras alinhadas com os arquivos de áudio, em que pode ser realizada a busca textual, e (2) as transcrições de 500 mil palavras anotadas morfologicamente e analisadas sintaticamente, também alinhadas com os arquivos de áudio. Estamos utilizando o* corpus *para investigar os diversos fenômenos no input linguístico e na trajetória do desenvolvimento de falantes de herança, tais como o uso de caso, gênero, construções passivas e impessoais, marcadores de polidez, disfluências e marcadores discursivos. Este artigo enfoca os desafios e soluções no processo da construção do BiRCh e as implicações para a pesquisa com base nos dados detalhadamente anotados fornecidos pelo* corpus.
PALAVRAS-CHAVE: Corpus *de fala em russo; Anotação de disfluências; Marcação morfológica; Análise sintática; Falantes bilíngues; Falantes de herança*

Introduction

This article describes the *corpus* of Bilingual Russian Child Speech (BiRCh, http://birch.ling.brandeis.edu), being developed at Brandeis University (Waltham, MA, USA). The project involves 10 bilingual children (between the ages of 2 and 9) from 9 Russian-speaking families in the U.S., Canada, and Germany, representing two language contact situations with two different majority languages (English and German), and 5 monolingual families (4 from Russia and 1 from Ukraine) with age-matched children who serve as a control group. BiRCh consists of audio-recordings of naturalistic interactions between children and their caregivers (usually parents) in familial contexts. The audio recordings are roughly balanced across the three majority language groups. The corpus includes 1-million-word transcriptions of the audio recordings with information about speech disfluencies (mainly false starts) and discourse phenomena (such as intra-sentential elaborations and repetitions), 500K-word portion of which is morphologically annotated and syntactically parsed. It also provides sociolinguistic information for every participating family: e.g., the amount and type of language contact the participating children have with the home and the majority language, educational levels of the parents and their proficiency in the majority language, etc. All transcripts are aligned with the audio signal, and the annotated data is connected to both the audio and the transcript, which makes it possible to use either text or grammatical searches to jump to the relevant point in the audio.

A unique and crucially important feature of BiRCh is the detailed morphological and syntactic annotation. In a morphologically rich language like Russian, many linguistic phenomena are impossible to study without detailed morphological information, extending beyond the part-of-speech (further, POS) tagging, and without syntactic annotation. For example, a study of Russian passives includes investigating three types of constructions: those which are formed with the help of a passive participle verb form, those that include the multifunctional suffix *-sja*, and impersonal constructions which often have a passive meaning, but have an active verb form with a null subject. Searching for null subjects and participial passives requires both morphological and syntactic annotation.

BiRCh is the first project of its kind,[1] and it is uniquely positioned for an investigation of factors affecting the development and change of grammatical competence in bilingual children because it is based on naturally occurring longitudinal data starting from early childhood. The corpus traces the acquisitional paths of bilingual and monolingual children before the time when the asymmetry of input and language use begins to grow in bilingual contexts with the onset of schooling (Benmamoun; Montrul; Polinsky, 2010). It documents a broad range of linguistic phenomena in multiple usage instances for both child participants and their parents, and therefore facilitates statistically significant generalizations, viable comparisons, and reliable correlations when comparing bilingual and monolingual parents, bilingual and monolingual children, and bilingual parents and their children. Additionally, the speech of the BiRCh parents presents important data not only for the study of input properties, but also for the investigation of language changes taking place across the lifespan of adult bilingual parents and for comparisons between different types of bilinguals.

For language acquisition researchers BiRCh will be interesting and important because it offers a closer look at deviations from the monolingual language acquisition trajectory as they accumulate over time and lead to a (potentially) heritage grammar. Adult heritage speakers[2] (HSs) are often compared to child language learners, and

[1] We would like to acknowledge the important RUEG project (Emerging Grammars in Language-Contact Situations: A Comparative View) being conducted in Germany; unlike BiRCh, however, it is based on elicited responses produced in experimental contexts by teenagers (14-18 years of age) and young adults.

[2] Heritage language speakers (HSs) are defined as "simultaneous or sequential (successive) bilingual[s] whose weaker language corresponds to the minority language of their society and whose stronger language is the dominant language of that society" (Polinsky, 2018, p.9).

indeed linguists have identified multiple areas of grammar in which both groups seem to pattern together and to differ from adult L1 speakers (Benmamoun *et al.*, 2014; Arslan, 2015; Sekerina; Sauermann, 2015; Arslan; Bastiaanse, 2020). However, these converging characteristics should not be taken as evidence that the heritage grammar "froze" mid-way to the adult L1 grammar (see e.g., Polinsky, 2011 for evidence of reanalysis).

There are at least four processes that can result in an adult heritage grammar that differs from the adult L1 baseline. Innovation in the heritage grammar may be caused by features of the dominant language (language transfer). Alternatively, the heritage L1 grammar initially converges with the adult L1 grammar and later either loses a feature or changes it after a period of disuse (language attrition). A third possibility is that the adult heritage grammar offers a different solution than the adult L1 grammar to the language input both receive (divergent attainment). Finally, the input to the acquisition process—bilingual and monolingual parents' speech—may differ as a result of changes to the bilingual parents' linguistic behavior (different input). Understanding the trajectory of the heritage language acquisition is a requisite component in diagnosing and disentangling these processes, and in turn gaining further insight into the nature of language acquisition and language knowledge in general. A grammatically annotated corpus of longitudinal records is a crucial tool in this research.

In the next sections, we detail the methodology for the construction of the BiRCh corpus, including an overall sketch of corpus construction workflow, and describe data collection, transcription, initial annotation (including bilingual, disfluency, and discourse phenomena), morphological annotation, and syntactic parsing. In each section we address the difficulties associated with building a deeply annotated corpus and describe our solutions to the challenge of finding a balance between building a rich and reliable resource and the need to limit the finite expense and effort. Finally, we provide examples of the BiRCh corpus use, both current and suggested.

1 Overall Sketch of the Corpus Development Pipeline

Figure 1 shows an overall development pipeline of the BiRCh corpus and the corresponding deliverables at each pipeline stage (cf. Põldvere *et al.*, 2021 for a recent attempt to create an audio-aligned spoken corpus in British English).

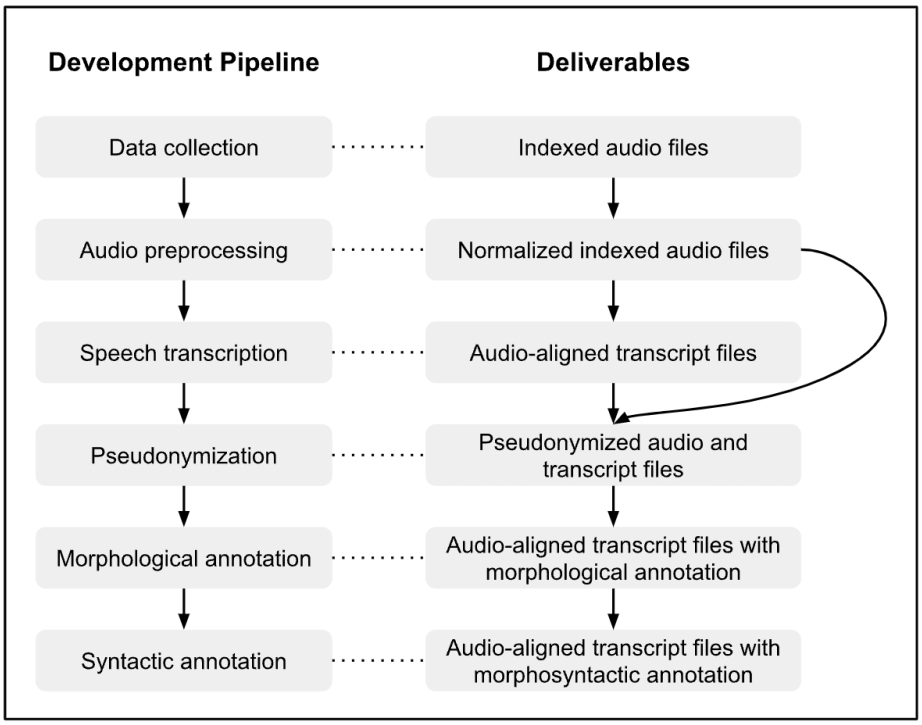


Figure 1. An overall development pipeline of BiRCh

At the data collection stage, each audio file is indexed using a standardized naming format. For example, the name of the fourth recording from the family of child S on the date when she was 4 years, 6 months and 9 days old is S_4-6-9_3 (the final counter starts at 0). At the audio preprocessing stage, each indexed audio file is preprocessed by joining chronologically adjacent recordings, separating files by removing speechless audio fragments, or increasing the volume. The name of each normalized audio file is used as the unique identifier for all derived files in the next stages.

Speech transcription consists of three main steps manually performed by different native speaker (NS) annotators in ELAN, a standard open-source annotation tool for creating time-aligned textual annotation of multimedia recordings:

- Initial speech segmentation, transcription, and initial annotation, consisting of the marking of

    - false-starts (the main disfluency annotation),
    - discourse phenomena that complicate syntactic analysis (such as intra-sentential elaborations, parenthetical clauses, and intra-sentential repetitions with rhetorical intent), and
    - bilingual phenomena (such as borrowing and code-switching).
- Checking of speech transcription and initial annotation (often also separate transcription of code-switched segments[3] in German or English).
- Speech segmentation checking.

This is the minimally viable process to assure the annotation quality at the gold standard[4] level. We decided not to use any automatic steps mainly because of the peculiarity of our data, (non-English) child and child-directed speech by bilingual and monolingual Russian families at home. Correcting the output of automatic transcription for this speech would require more work than manual transcription from scratch. The output of the speech transcription stage are transcript files (in the XML-based ELAN Annotation Format, i.e. EAF), time-aligned with audio at the segment level and accompanied with timestamps of personally identifiable information, which are replaced in both audio and transcript files at the next stage, pseudonymization.

Pseudonymized transcript files are then used as the input of the morphological annotation stage, including:

- Automatic tokenization, breaking segments into lists of word tokens.
- Automatic morphological annotation of Russian word tokens, consisting of lemmatization and POS and morphological feature tagging.
- Manual correction of morphological annotation.
- Manual final checking of morphological annotation.

We developed an in-house rule-based tokenizer and morphological tagger to maximize fit with our in-house transcription and morphological annotation guidelines, respectively. The morphological tagger uses Mystem as its core, the best option in terms of performance and tagset comprehensiveness[5] among the most popular morphological taggers for Russian (Kotelnikov; Razova; Fishcheva, 2018). This core is continuously enriched by additional rules that are learned from our annotation practice. The output of the automatic processing is saved in the XML-based FoLiA format. This format stands out for its versatility, human-readability, and portability by accommodating multiple linguistic annotation types with arbitrary tagsets and including all annotation layers in a

[3] We term sentence tokens "segments."

[4] In computational linguistics, "gold standard" refers to accuracy and consistency of human annotation.

[5] Mystem's tagset is built on Zaliznyak's Russian Grammatical Dictionary, and used for morphological analysis of the Russian National corpus. It is also used in the RUEG project.

single file (Van Gompel; Reynaert, 2013). It is also accompanied by FLAT, a web-based annotation tool whose user-interface can show different linguistic annotation layers at the same time (Van Gompel *et al.*, 2017). The output of the automatic steps is manually corrected and then checked by different annotators on FLAT, which assures the gold standard quality. The software that supports the FoLiA format, such as FoLiApy and FoLiA-tools, allows us to easily manipulate the data at any point of the corpus development cycle (e.g. due to revisions in the annotation guidelines), making the workflow more flexible and interactive. The selection of FoLiA not only enables us to align all our annotation aspects in one file (cf. Tortora, 2014), making it easy to implement any future annotation revisions, but also enhances accessibility of BiRCh to future computational analysis or annotation.

Finally, the syntactic annotation stage involves interleaved iterations of two main steps:

- Automatic syntactic parsing.
- Manual correction of syntactic parses.

As we follow the influential methodology of existing parsed corpora capturing language variation and change (Tortora; Santorini; Blanchette, 2018), such as the Penn Parsed Corpora of Historical English (PPCHE) (Kroch *et al.*, 2016) and the Audio-Aligned and Parsed Corpus of Appalachian English (AAPCAppE) (Tortora *et al.*, 2017), we parse each segment into at least one phrase structure tree, using the software CorpusSearch 2 (CS) (Randall; Taylor; Kroch, 2005) for rule-based automatic parsing and search-based manual correction. Our first step is to convert the morphologically annotated data from FoLiA into the Penn Treebank format, keeping track of the identifiers of all segments and word tokens for the integration of syntactic annotation into the audio-aligned and morphologically annotated FoLiA files. Thus, the deliverables of the syntactic annotation stage include both the parsed files in the Penn Treebank format and the integrated FoLiA files. The latter are ultimately used[6] in the deployment of a web-based search and visualization interface based on ANNIS, a well-developed open-source architecture specialized for corpora with multiple linguistic annotation layers (Krause; Zeldes, 2016). Our internal investigation shows that

[6] We collaborate with Maarten van Gompel, the key author of FoLiA, to convert FoLiA into SaltXML (Zipser; Romary, 2010), which in turn can be integrated into the ANNIS infrastructure.

ANNIS's query language, AQL, is able to cover all of the search functions implemented in CS, and therefore, provides users with at least the same power to search for target syntactic structures. Moreover, as AQL's expressiveness is agnostic with respect to annotation types, BiRCh can be explored in novel ways in comparison with its predecessors under the Penn Treebank paradigm.

This development pipeline had been continuously optimized until becoming stable. We use a one-stop project management platform[7] to help us organize, record, communicate about, and analyze everyone's contributions. Chat and knowledge base features provide a central space to raise questions and answer them.

2 Data Collection

This project would not be possible without the good will, interest, and commitment of participating families who played a paramount role in the creation of the BiRCh corpus. Mothers were usually the driving force behind the family's decision to participate. In most families, mothers have linguistic, pedagogical, or philological training and were motivated by their professional interests. For bilingual families, motivation to participate included their commitment to their child's bilingualism, and for all families, a genuine interest in the advancement of linguistic research played a decisive role. We supported participants through regular personal contact, and through yearly BiRCh project newsletters where we reported on the progress to date and provided families with research-based guidance for supporting the linguistic development of their children. We also shared some of the annotated files with each family to motivate further participation.

Each participating family was asked to make weekly audio recordings of at least 30 minutes of verbal interactions with their child. This recording schedule continued for as long as possible, with breaks for summer and winter vacations. The average participation is 3.26 years, and the longest non-stop participation ran for seven years. Initially, families were provided with high quality Sony recorders. In the first year of

[7] The most useful features include task templates, which can be reused by different team members to create tasks of the same kind, and task reports with the time tracking information, which allow us to calculate the workload and efficiency of different tasks performed by different team members to optimize our workflow and adjust our budget.

the project, we switched to the professional recording device ZoomH2n and set a requirement for all audio files to be saved in the WAV format to preserve the accuracy of acoustic data. Both of these steps ensured that BiRCh data would be useful for phonological research.

In the second year of the project, we tested the data transcribed to date for the presence of low-frequency linguistic phenomena, such as passive constructions, in children's speech, and discovered that our recording schedule at the time was capturing only about 1% of low-frequency forms. Guided by previous research (Rowland; Fletcher; Freudenthal, 2008), we invited one family with a 4-year-old child in each group to go up to a dense recording regime, to record from 3 to 7 hours per week (the dense corpus participants from Germany could only do from 1.5 to 3.5 hours per week). To make the dense sampling as easy as possible, we provided these 3 families with ATTO Digital miniature recorders that could be worn on a child's clothing and have a battery life of up to six hours between recordings. This recording schedule continued for six months, after which dense corpus volunteers could return to the regular recording schedule.

We also gathered sociolinguistic data on all participating families on a bi-yearly basis. At the initial intake, we collected basic information on the family's linguistic profile, including place of birth for the parents and the child, current place of residence, household composition, and the age of every child, whether participating in the project or not. We also employed the Bilingual Language Exposure Calculator (BiLec) (Unsworth *et al.*, 2012; Unsworth, 2016) to gather in-depth ethnographic and sociolinguistic information about bilingual families. This questionnaire included detailed questions about the amount of exposure to each of the child's languages, including the percentage of daily use for each language, proficiency levels of the parents and other caretakers, and passive language contact, such as TV time or audio books.

3 Initial Annotation and Segmentation

Transcription, segmentation, and initial annotation occur as a single process, which requires clear guidelines to achieve consistency. The main principle of our guidelines is to enable reliable retrieval of examples by future corpus users while

minimizing cognitive load on annotators. The ultimate goal is to produce a grammatically analyzed corpus, including syntactic parsing. We, therefore, only annotated those disfluency, discourse, and bilingual phenomena that, if left unmarked, would interfere with morphological and syntactic annotation. BiRCh Initial Annotation (Malamud; Dubinina, 2017a) and BiRCh Segmentation Guidelines (Malamud; Dubinina, 2017b) are based on the AAPCAppE guidelines (Santorini; Diertani, 2017), which in turn are based on the PPCHE guidelines (Santorini, 2016) and the discussion in Hindle (1983). We clarified existing categories from the AAPCAppE and PPCHE and added new categories specific to the nature of child and child-directed speech as well as bilingual speech.

We transcribe Russian using Cyrillic (UTF-8 encoding), and use the Latin alphabet for German and English. We established standardized spellings for hard-to-transcribe language phenomena, such as pause fillers and other interjections, e.g., *aa*, *mm*, *nea* 'nope' to ensure their searchability, which is crucial for future research on disfluencies.

We do not conduct a full disfluency annotation (pauses, repairs, etc.), instead focussing on what Hindle (1983) called "syntactic non-fluencies." Our main disfluency category is false-start; in the initial annotation we also mark repetitions (exact repetitions with rhetorical intent), elaborations (non-exact repetitions, paraphrases, which do not amount to full main clauses, and which clarify constituents that are not full sentences themselves), and parenthetical clauses.

These categories of syntactic non-fluency identify constituents that do not fit neatly into the syntactic annotation algorithm. Initial syntactic parsing ignores them and is thereby streamlined, but elaborations and parenthetical clauses are later parsed and become syntactic labels in the final corpus (Santorini; Diertani, 2017). Our guidelines provide extensive clarification and introduce some changes to the definition of elaborations and parenthetical clauses. The latter category in our corpus encompasses two types of clauses: (i) parenthetical or peripheral commentary clauses and (ii) elaborations that amount to full main clauses. We do not mark sub-clausal parentheticals.

(1) Oj (PAREN ty znaeš') kogda ja byla devočkoj
Oh you.SG know when I was:F girl:INS.SG
(PAREN mne naverno četyre godika bylo)
I.DAT possibly four years.DIM was.N

deduška Saša (ELAB moj papa) prines
grandpa Sasha my dad PRF:brought.PST.M.SG
vot takuju ogromnuju golovu ščuki.
FOC such:F.ACC.SG huge:F.ACC.SG head.F:ACC.SG pike.F:GEN.SG
'Oh, you know, when I was a girl, perhaps four years old, grandpa Sasha, my dad, brought a pike's head this big.'

There is a close relationship between the marking of parenthetical clauses and segmenting the transcript. For full main clauses that are related in content to the rest of the segment and occur in the beginning or end of that segment, annotators have to decide whether these constitute parentheticals, separate segments, or, sometimes, a main clause embedding the rest of the segment. In addition, since argument drop is possible in Russian (especially in informal conversations), annotators often need to decide whether a particular phrase constitutes a full main clause or not. We have developed heuristics for these decisions. For ease of retrieval, if two clauses can be thought of as examples of a specific construction, we tend to err on the side of not splitting them into separate segments.

To fully capture the phenomena of spontaneous interactions between children and caregivers in our corpus, we introduce new annotations for singing (sung speech), mispronounced words (only for gross mispronunciations) and nonce words/neologisms, as well as annotation of bilingual phenomena such as nonce borrowing and code-switching. In BiRCh, a word is marked as a (nonce) borrowing if it is adapted to the phonological and morphological systems of the Russian language. Morphologically adapted borrowings may have case-marking or other morphology, as in (2), where German *oma* 'grandma' appears with Russian instrumental case suffix *-oj*.

(2) opa s om-oj
grandpa COM grandma-INS[8]
'grandpa and grandma'

Poplack *et al.* (2020) show that purely phonological criteria are not reliable indicators of borrowing or code-switching.[9] Since we were unwilling to count all such words as code-switches (or as borrowings), we created guidelines that allow annotators to consistently tag these phenomena and allow corpus users to find such examples and

[8] We use Leipzig glossing rules with the following options and modifications: we only include morphological features immediately relevant for each example. In addition, we will generally not mark tense on finite non-past verbs – in Russian, non-past verbs are marked for person (e. g., *odolžu* PRF: borrow:1SG), while past verbs are not marked for person and instead are marked for gender (in singular forms) (e. g., *odolžila* PRF:borrow:PST:F.SG). Thus, the reader will note that those finite non-imperative verbs that are glossed for person but not gender are non-past.

[9] We are thankful to an anonymous reviewer for pointing us towards this literature.

conduct phonetic analysis and potentially argue that some examples should be reclassified. Since individual variation is famously present, our marking of purely phonological borrowings is speaker-dependent: that is, they are more closely integrated into Russian than that speaker's extended code-switches into English or German. For example, the word *pafin* 'puffin' used by one of the parent participants, which, unlike that speaker's English pronunciation, does not have the aspirated /p/ and has the palatalized /f/, is spelled in Cyrillic letters and tagged as a borrowing. In contrast, the phrase *Baby süß* 'sweet baby', pronounced in accordance with all the phonological rules of German, is written in German and is considered to be an instance of code-switching. Even with these heuristics it is often difficult to differentiate between borrowings and code-switching as differences in pronunciation can be unclear, and the word may show no other marks of adaptation to the Russian linguistic systems (e.g., the presence of case markings for nouns). In these cases, the decision was to err on the side of code-switching.

In BiRCh, borrowings are treated as Russian words and are tagged with all the pertinent morphological information while code-switched items are left unannotated. For example, the word *pafin* 'puffin' in the example above is tagged as noun, masculine gender, animate, singular, nominative. We turn to the discussion of morphological tagging next.

4 Morphological Guidelines

Russian is a morphologically rich language with flexible word order, and many syntactic structures are distinguished by morphological means (e.g., case). Moreover, morphology, in particular, has been noted as an area where HSs diverge from monolingual baselines (POLINSKY, 2018). Therefore, full morphological annotation that goes beyond the POS tagging is imperative to enable any corpus-based research into the grammatical development of Russian speakers.

In our overall approach, we were inspired by the morphologically tagged Russian National Corpus (RNC, 2003). The starting point for our Morphological Annotation Guidelines (Dubinina *et al.*, 2019) is the Mystem tagset, which is close to that used by the RNC, facilitating comparisons with BiRCh data. Like these prior

resources, we use two types of labels for morphological annotation of each word: a POS label and a set of morphological feature labels (further, features). In BiRCh, we annotate several phenomena not marked in the RNC, and in many instances we depart from the RNC analysis for existing phenomena. In the rest of this section, we focus on these differences with the RNC, and mention challenges of morphological annotation for our data and their solutions.

4.1 New Phenomena not Marked in the RNC

Differently from the RNC, BiRCh data is rich in language phenomena which characterizes bilingual contexts and child language acquisition, such as borrowings, code-switching, nonce words, and non-standard morphological forms. In the previous section, we mention morphological tagging of nonce borrowings; here we turn to other phenomena.

BiRCh uses a single tag for nonce words and neologisms. Nonce words in their traditional definition are made-up words which are often the result of children's word play: e.g., *kmiščeta,* created by a child and explained by her as "a combination of yellow and red colors." This annotation category also includes neologisms in a single-family community, i.e., those words that are present in parental speech as familial nicknames for people and objects. For example, in one family, the parent and child consistently use the neologism *podguz* instead of *podguznik* 'diaper.' There are also nonce words that result from children's mispronunciation of legitimate Russian words, e.g., *xamil'jard* (instead of *xameleon* 'chameleon'). Neologisms and nonce words are morphologically annotated if they can be accepted as words by NS annotators. Otherwise, they receive a non-word POS tag.

Finally, forms that feel like errors to NS annotators are marked as "unexpected" and include the expected/grammatically correct form (3). This allows corpus searches to find morphological errors while also allowing that there may also be unexpected forms based on dialectal differences between participants and annotators. This annotation enrichment is essential for research into the acquisition of morphological forms.

(3) Moja (unexpected form, moe) učenie
My:FEM (unexpected form, my.N) learning:N
'My learning'

Another innovation of the BiRCh annotation is the marking of diminutives, which are not singled out in the RNC or other Russian *corpora*, but are particularly interesting from the point of view of acquisition. We annotate diminutives by using the feature DIM and by inserting a hidden word indicating the non-diminutive base form. Therefore, a corpus search for a specific noun will turn up examples with both diminutive and non-diminutive forms.

We use a similar annotation process for deverbal ideophones, i.e., interjections etymologically related to verbs and sometimes retaining the subcategorization properties of these verbs, as in (4). We insert the related verb as a hidden word and mark it as having an ideophone link to the interjection.

(4) A lisa xvat′ (xvatat′) ego za xvost.
And fox grab.INTJ (grab:INF) him by tail
'And the fox grabbed him by the tail.'

Two other BiRCh innovations are related to the fact that the corpus provides both morphological and syntactic annotation. First, we mark morpho-syntactically relevant information that is not apparent from the morphological form of the word. For example, BiRCh has the feature "quantificational" for those adverbs that can govern the genitive case of nouns (*mnogo* 'many', *čut'-čut'* 'a tiny bit', and several others). The second innovation concerns a feature that will not be visible in the published corpus, but that is important for the syntactic analysis of sentences containing the present-tense copula, which is null in Russian. We use the feature "predicate" on a word (typically the head) in the remnant constituent in a verb phrase headed by the null copula, and annotators test for the presence of the copula by changing the utterance into the past or future tenses, to see if *byl(a/o)* 'was' or *budet* 'will be' emerge. Once the null copula is inserted during parsing, this feature is no longer needed.

4.2 Innovations for Phenomena Described in the RNC

In traditional Russian grammars (e.g., Ušakov, 1935; Ožegov; Švedova, 1997), idiomatic multi-word expressions are often assigned a single POS category, such as particle or conjunction. We ensure a separate POS for each word, which makes searching for specific words yield more comprehensive results and aids both morphological and syntactic analysis. We also separate *wh*-indefinite series (indefinite

pronominals consisting of a *wh*-word and a particle such as *-to* or *-nibud'*) into *wh*-words and particles, e.g., *komu-to* 'someone.DAT' becomes the dative pronoun *komu* 'who:DAT' (lemma *kto*) followed by the particle *-to* (lemma *-to*). This allows us to unify *wh*-indefinites with the *wh*-words on any of their uses, as well as with other uses of some of the particles, e.g., other uses of the focus particle *-to*, as in (5). Separate words that are usually spelled together in conventional orthography are marked with @.

(5) To -to on byl rad vstretit' kogo@ @-to!
That.ND -TO he.NOM BE:PST.M.SG glad.M.SG meet:INF who.ACC -TO
'That's when he was so glad to meet someone!'

Particles more generally form a sprawling category in the RNC and traditional grammars, encompassing many words that serve a variety of functions, as well as multi-word expressions. We conducted a comprehensive survey of particles in the RNC and in the Ušakov or Ožegov dictionaries, aiming to narrow down this POS category to those words that cannot be considered adverbs, conjunctions, or other POS.

In addition to narrowing down the particle POS, we also eliminated the POS 'predicate' assigned to several words in the RNC, Mystem, and some grammars (e.g., Zaliznyak, 2007), e.g., *nužno* 'needed' or *izvestno* 'known'. This POS is not part of the PPCHE, AAPCAppE, or Universal Dependencies tagsets (UD POS, 2014). We mark these as adverbs, since their morphological form (the *-o* suffix) conforms to that POS, and several of these words also have adverbial uses.

A final departure from the RNC annotations that we want to mention is the use of the feature 'non-declinable' (ND) for pronouns *čto* 'who', *vsë* 'all', *to* 'that' (6), and finally, *èto* 'this'(7), including its use as a pause filler (6a, compare with 6b), in certain constructions when their case is difficult or impossible to ascertain.

(6) Usač - èto žuk.
Longicorn this:ND beetle
'A longicorn is a beetle.'

(7) a. Daj mne èto ... kružku.
Give me this:ND mug.F:ACC.SG
'Give me, um, a mug.'

b. Daj mne ètu kružku.
Give me this:F.ACC mug.F:ACC.SG
'Give me this mug.'

4.3 Challenges of Morphological Annotation

The most pervasive problem in morphological annotation is ambiguity, i.e., marking words with multiple morphosyntactic functions. Our guidelines include a list of

many such words with detailed explanations. We eliminate ambiguity whenever possible, but in cases when words remain ambiguous in context, we mark both morphological possibilities. This not only allows us to avoid making arbitrary annotation decisions, but also provides corpus users with information about ambiguous interpretation of data. We will highlight just two examples of disambiguation in morphological annotation - marking a word as a participle or an adjective and annotating a word as a short-form adjective or an adverb.

4.3.1 Participles vs Adjectives

In traditional Russian grammar, an etymologically deverbal modifier is considered to be a participle when it has complements and/or prefixes; otherwise, it is considered an adjective. In many cases, specifically with suffixes *-n-* and *-en-*, this choice affects the spelling: participles are spelled with double <nn>, while adjectives with a single <n>, despite not exhibiting any difference in pronunciation (see (8a,b)).

(8) a. U nejo vjazanaja jubka.
At her.GEN knit(ted):F.NOM.SG skirt.F:NOM.SG
'She has a knitted skirt.'
b. Vjazannaja kr̨učkom ili spicami?
Knitted:F.NOM.SG hook.M:INSTR.SG or needles:INSTR.PL
'Made with a crochet hook or with knitting needles?'

For consistency of annotation, we mark such words as participles (that is, verbs with the feature "participle"): in (8), both *vjazanaja* and *vjazannaja* have the lemma *vjazat′* 'to knit.' We add "possible adjective" as a feature so that users searching for either verbs or adjectives would find these forms.

4.3.2 Adjectives vs Adverbs

Words ending in *-o* or *-e* could be short-form adjectives or adverbs. In simple cases, a noun phrase modifier which agrees with the head noun is an adjective, while a verb phrase modifier is an adverb. In cases where the word occurs in a copular or related construction with a neuter subject, annotators check for agreement by substituting a feminine or plural subject, as in (9ab).

(9) a. Ej ètо interesno.
Her.DAT this:N.NOM.SG interesting:ADJ.N.NOM.SG

b. Ej oni interesny.
Her.DAT they:NOM interesting:ADJ.NOM.PL
'She's interested in this/them.'

In constructions where the substitution test is not available, we have rules to ensure consistent annotation. For instance, in an utterance with a copula or related verb (e.g., 'become') and without an overt subject, the target word is always marked as an adverb. Finally, when it is not clear how to classify the construction, we err on the side of adverbs, as in the example below.

(10) Ej interesno, xorošo, veselo (v škole).
Her.DAT interesting:ADV good:ADV merrily (in school)
'She's interested, well, and merry (in school).'

After automatic morphological tags are corrected and checked manually, the data moves to the ultimate stage of annotation: syntactic parsing.

5 Syntactic Parsing

Developing syntactic annotation is one of the most laborious and expensive decisions corpus creators can make. Syntactic structure, in contrast to morphological form, is often invisible (except for occasional prosodic cues) and has to be inferred. Syntax of any human language is countably infinite (modulo performance) and offers a notoriously wide range of ambiguity. In this section we start by showing that these two obstacles—invisibility and infinity—applied to the bilingual data make the decision about the syntactic annotation a no-brainer. We then review the main architectural and construction aspects of our syntactic annotation and show how together they facilitate addressing theory-charged questions.

5.1 Motivation for the Syntactic Annotation

Heritage grammar is similar to a complex tapestry weaving together various influences and processes. BiRCh shows how this tapestry unfolds in time and helps unravel its non-trivial acquisition trajectory. As outlined in the introduction, to "unweave the rainbow" at least four processes have to be disentangled: language transfer (borrowing of grammatical properties from the dominant language), language attrition (modification of L1 grammar), divergent attainment (producing new or

different features based on incomplete input), and different parental input to monolingual and bilingual children. In the case of different input, the roles of language transfer and independent innovation must be distinguished in parents' speech. Importantly, all these processes involve some form of misalignment within form-meaning pairs that is significantly more common among invisible and infinite syntactic structures than the directly observable and finite morphophonological units. The syntactic annotation is the best place to scrutinize these processes, which makes it a necessity for our corpus to meet the needs of researchers in language acquisition, heritage languages, language contact, and theoretical syntax and semantics.

5.2 Architecture of the Syntactic Annotation

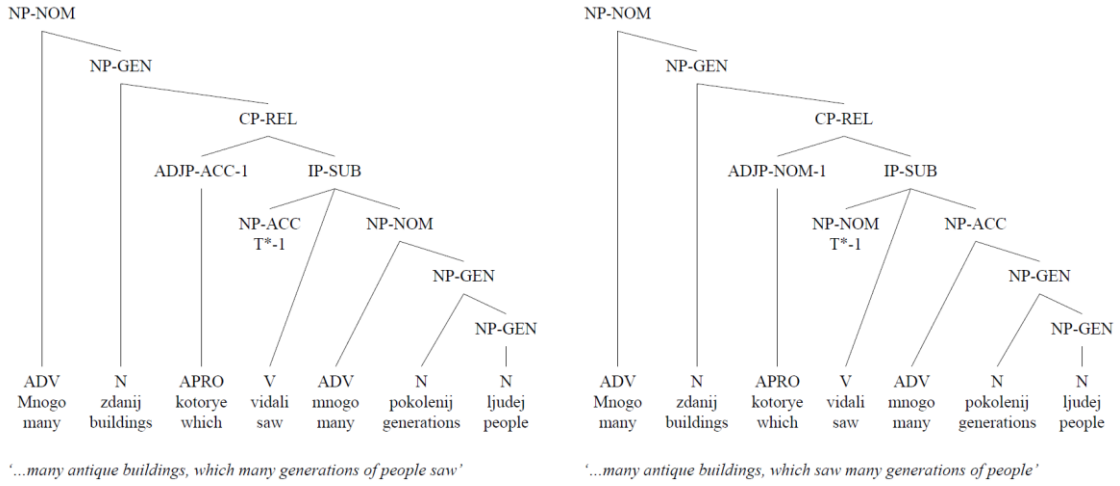
Syntactic annotation seeks a harmonious balance between the resources and ambitions of corpus creators and the needs of potential users (as interpreted by the corpus creators). One of the main goals for us was to make BiRCh accessible to a wide range of language professionals that may benefit from syntactic insight. To this end two principled decisions were made early on that shaped the syntactic annotation.

First, we adopted the Penn treebank style of syntactic annotation (Marcus; Santorini; Marcinkiewicz, 1993). The goal at the heart of this style is to facilitate automated syntactic search, and in pursuit of this goal theoretical accuracy can sometimes be abnegated in favor of annotation simplicity. Specifically, the Penn Treebank syntactic annotation allows *n*-ary branching and exocentric (i.e., head-less) structures. On the other hand, this style of annotation is familiar to many linguists who have used other parsed corpora that are created following the same principles (Martineau, 2008; Wallenberg *Et Al.*, 2011; Beck, 2013; Kroch *Et Al.*, 2016; Galves; Andrade; Faria, 2017; Tortora *Et Al.*, 2017; Kroch, 2020). It is also important that the Penn Treebank format includes the powerful query language of CS that is used for both searching and modifying the corpus. As a result, our syntactic annotation may appear agnostic about some difficult questions of Russian syntax (see below) while seeking instead to provide means for all researchers to easily find the data they need.

Second, in an attempt to help address some of the more complicated questions in heritage language syntax, we chose to concentrate our efforts on two language

properties: ambiguity and silence. In terms of the former, HSs often struggle to navigate ambiguous utterances (see Polinsky; Scontras, 2020 and references therein). For HSs of Russian, such difficulties manifest on multiple levels, from lexical synonyms (Rakhilina; Vyrenkova; Polinsky, 2016) to anaphora resolution (Ivanova-Sullivan, 2014a) to word order and quantifier scope (Ionin; Luchkina, 2019). Yet it is still only a conjecture that HSs tend to avoid ambiguity altogether. To address this question, the syntactic annotation in BiRCh systematically includes information about syntactic ambiguity. In BiRCh, every segment can be associated with multiple syntactic structures. In this way, any ambiguity, provided that it is detectable with the amount of detail offered in our annotation, gets included and can be found (as in 11).

(11)



Turning to silence, another recurrent theme in HL linguistics concerns the meaning associated with the absence of overt material. Its interpretations can often lead researchers to paradoxical conclusions. For example, heritage grammars are often cited for the attrition of null pronouns (Montrul, 2004; Serratrice; Sorace; Paoli, 2004; Tsimpli *et al.*, 2004; Polinsky; Kagan, 2007; Haznedar, 2010; Keating; Vanpatten; Jegerski, 2011; Nagy *et al.*, 2011; Ivanova-Sullivan, 2014b). The effect is so prevalent and strong that speakers of a *pro*-drop language produce significantly more overt pronouns in their heritage language (in comparison to baseline speakers) even when their dominant language is also *pro*-drop (see De Prada Pérez, 2009, 2015 for the Spanish-Catalan data). At the same time, some types of ellipsis are argued to be substituted by the *pro*-drop. Polinsky (2016, 2018) claims that Russian HSs reanalyze a

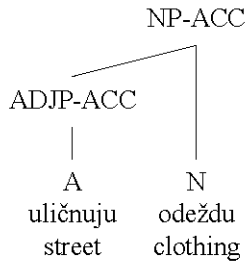
specific kind of ellipsis called the verb-stranding VP ellipsis as an object drop since both appear to lead to identical surface forms (Goldberg, 2005; Gribanova, 2013). In other words, HSs seem to strongly disfavor *pro*-drop, except when they re-interpret some kind of ellipsis as *pro*-drop. The combination of these tendencies, in turn, suggests that object *pro*-drop may be more frequent than other argument drops. Our syntactic annotation in which *pro*-drop is marked for all obligatory arguments is well-suited to check whether claims of this kind are supported.

5.3 Construction of the Syntactic Annotation

Turning from architectural decisions to corpus construction, the syntactic annotation in our corpus was created using the CS query language that was designed to work with Penn Treebank format parsed corpora. To be compatible with this query language, the output files of morphological annotation are first normalized and converted from the FoLiA format into the Penn Treebank format such that for each segment, discourse phenomena that complicate syntactic analysis (e.g. intra-sentential elaborations and parenthetical clauses) are encapsulated in separate bracket pairs and therefore can be parsed after the main content is fully parsed. The parsing process for the main content of each segment (which is later also applied to the initially encapsulated elaborations and parentheticals) was separated into three phases that target different groups of constituents and grammatical phenomena. Each phase consists of two steps. First, semi-automated rule-based syntactic parsing proceeds using corpus revision queries. As a second step, the parsing is corrected manually through corpus searching queries that identify specific classes of examples which are then changed in text editors working on the Penn Treebank style bracket notation. The separation into three phases reflects the basic idea of growing syntactic trees from the bottom up. In this case each phase uses the syntactic information gathered and consolidated in the previous phase.

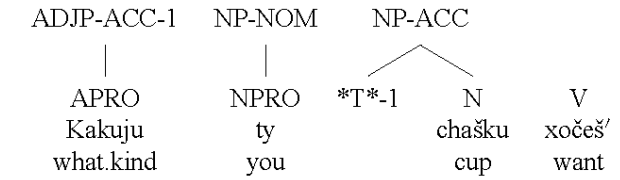
During the first phase, the morphological information (POS and case features) is used to identify and project "small" endocentric constituents (NP, AP, NumP, PP, etc.). At this step constituents are also supplied with the dash-tags for case. The case information is used to identify and embed sub-constituents of NPs, to diagnose sub-extraction, and reconstruct NP-internal traces (see 12 and 13 below).

(12)



'street clothes'

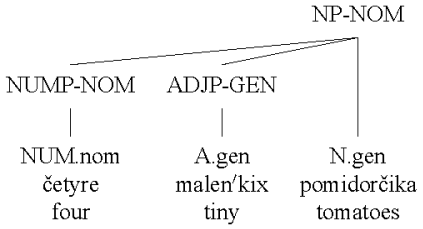
(13)



'What cup do you want?'

For simplicity, all nominal (sub-)constituents are embedded within NP, as in the example below:

(14)

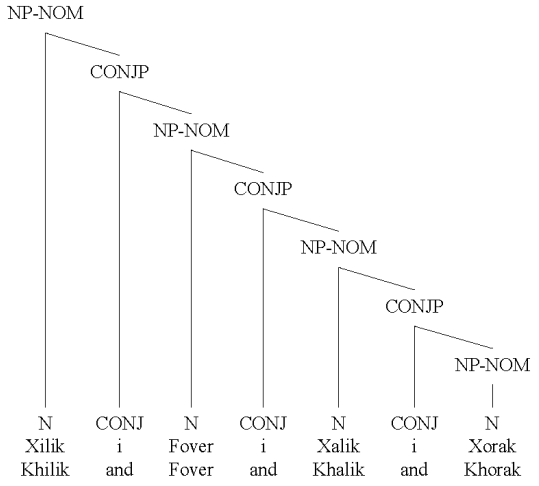


'four tiny tomatoes'

This simplified NP structure also means that case dash-tags in some structural case contexts need to be fixed. In (14) above, for example, the case of the NP in a numerical construction with the genitive of quantification needs to match the case of Num and not of N.

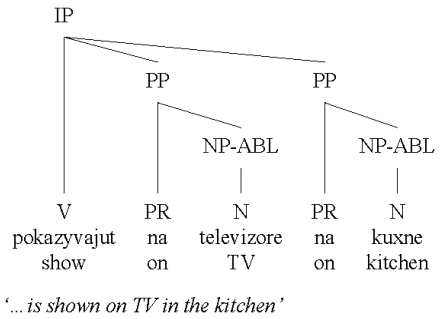
After the NPs are identified, we also mark conjunctions of NPs (15). For ease of exposition, every subsequent conjunct is included in the preceding conjunct. Manual correction during the first phase consists of re-assignment of post-nominal PPs that can be a part of NP or a clausal argument/adjunct, as in (16).

(15)



'Khilik and Fover and Khalik and Khorak'

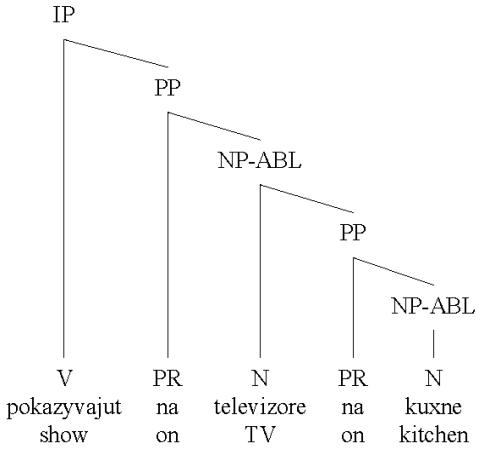
(16)



'... is shown on TV in the kitchen'

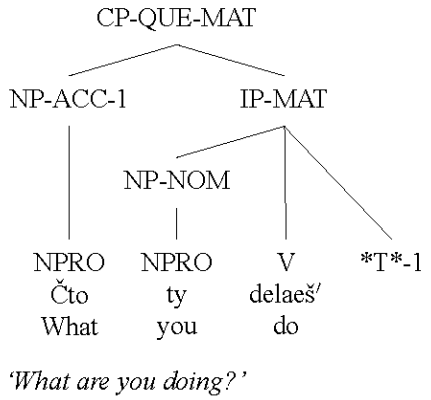
When both alternative positions are plausible, two trees are generated and associated with the segment, as in (17). The second phase consolidates the syntactic information that was gathered during the first phase to localize "large" exocentric constituents (IP and CP), identify copular clauses, and reconstruct traces of clause-internal and clause-external movement, as in (18).

(17)



'... is shown on a TV in the kitchen'

(18)



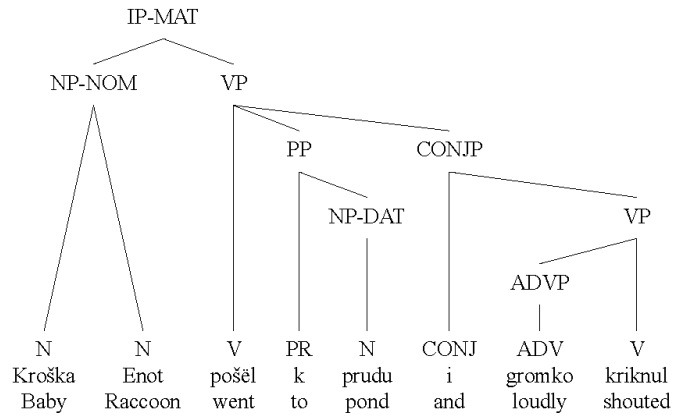
'What are you doing?'

We decided against including VPs in regular syntactic annotation. Localizing VP involves a plethora of theoretical commitments (e.g., do we also include AspP, *v*P, or VoiceP? VP shells? Do we reconstruct all subject NPs inside VP/*v*P or only some of them? Do we keep the same functional architecture for unergatives and unaccusatives? How do we analyze *psych*-predicates? Where do we attach specific adverbs?). In the end, all these options only complicate search and inundate (and, eventually, drown) the language researcher with the intricacies of Russian theoretical syntax. The only two

places where we explicitly mark VP is when it is highlighted by a syntactic process (fronting, conjunction, etc.), as in (19), and when it is one of potential ellipsis sizes.

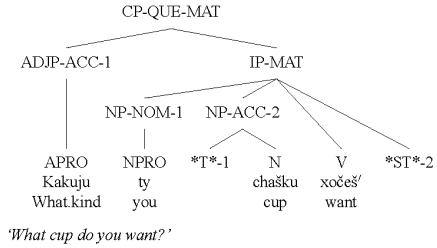
During the automated part of the second phase, we fill in traces of the leftward movement of constituents that were identified during the first phase. We assume that Russian is an SVO language and so all other permutations are created by scrambling, as in (20).

(19)



'Baby Raccoon went to the pond and screamed loudly'

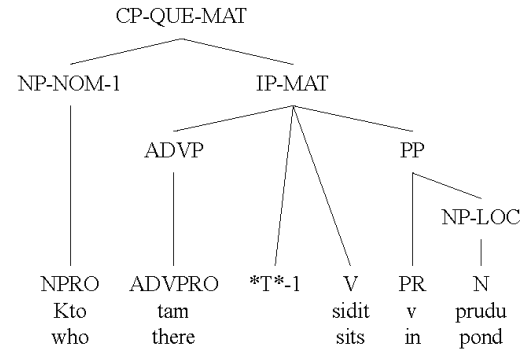
(20)



'What cup do you want?'

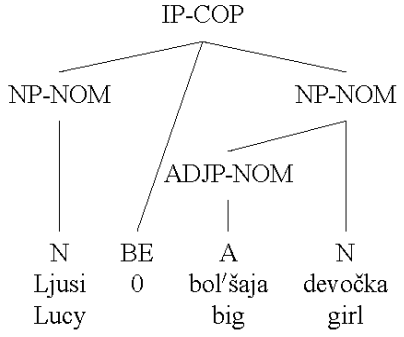
We also maintain that any leftward movement across left-adjoined adverbs is the result of fronting. Depending on whether the moved element is a *wh*-phrase or a focused element we separate fronting movement into *wh*-fronting and focus fronting. The traces of both are reconstructed during the second step (21). During the second phase we also reconstruct null copulas in copula clauses, as in (22).

(21)



'Who is that sitting there in the pond?'

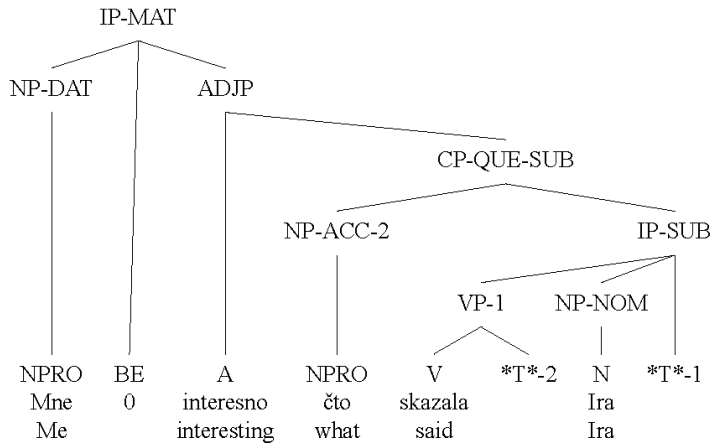
(22)



'Lucy is a big girl'

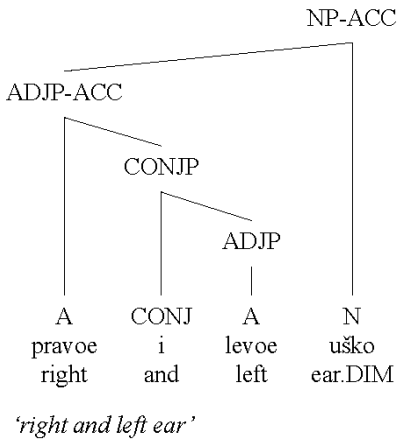
Lastly, we project CPs on top of IPs for matrix *wh*-questions and declarative and interrogative subordinate CPs. The latter two are further included in appropriate constituents. At this point the syntactic annotation contains enough information to mark all different subtypes of CPs and IPs (matrix, subordinate interrogative, etc.), as in (23). The manual part of the second phase includes correction of conjunction, a visual check of the assignment of traces, and an evaluation of the subtypes of IP and CP (see Example 24).

(23)



'I am curious what Ira said'

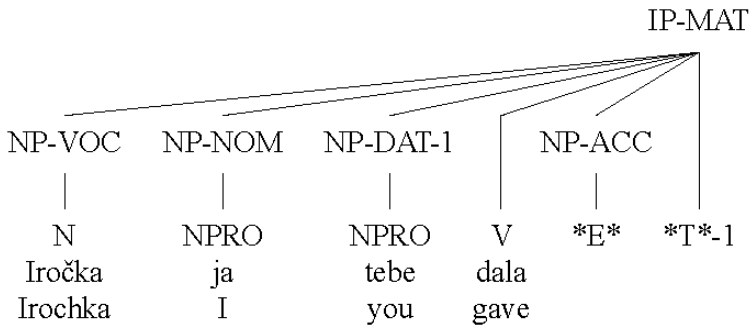
(24)



'right and left ear'

Finally, the third phase aims to fill in the "gaps" in the clausal structure that are usually associated with ellipsis and fragment answers. During this step every predicate and its arguments (identified using case dash-tags) are checked against our in-house dictionary of verb valencies. In case of any missing arguments, the corresponding null argument is created and added into the appropriate place, as in (25).

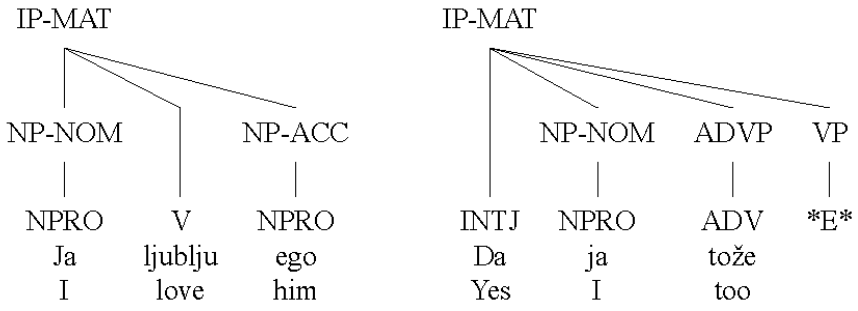
(25)



'Irochka, I gave it to you'

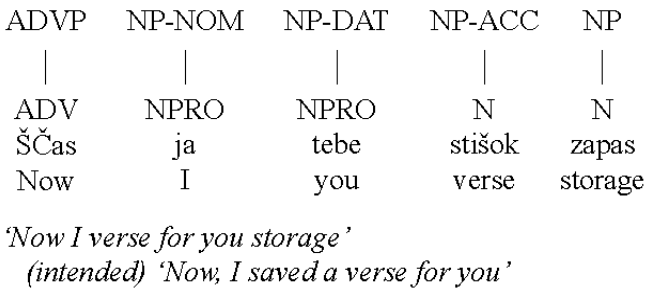
The manual part of the third phase includes correction and expansion of the non-nominal ellipsis phenomena to include cases of VP ellipsis as well as gapping and sluicing, as in (26).

(26)



After the third phase one final manual check-up for corrections is conducted for the entire corpus to exclude the possibility that the incorrect resolution of morphological ambiguity (such as classifying *zapas* as noun when it is a verb, as in (27)) resulted in a series of errors crawling up the syntactic tree.

(27)



Conclusions and Implications for Research

What makes BiRCh unique is its ultimate annotation layer, i.e., syntactic parsing, as it is the first and only large-scale constituency treebank in Russian at the moment. In terms of annotation specification, it is directly comparable to the 1-million word Wall Street Journal subcorpus of the Penn Treebank (Marcus *et al.*, 1999), the most frequently used dataset for constituency parsing in English (Clark; Fox; Lappin, 2013, p.241), and therefore, can play a similar role for Russian constituency parsing and contribute to the multilingual parsing, including morphologically rich languages

(Seddah *Et Al.*, 2013; Seddah; Kübler; Tsarfaty, 2014). Moreover, we can create a unique resource for constituency parsing across written and spoken modalities by converting SynTagRus (Boguslavsky *et al.*, 2002), a large-scale dependency treebank of written Russian texts,[10] into a constituency treebank and combining the result with BiRCh's spoken data. The conversion of SynTagRus is plausible based on our Morphological and Syntactic Annotation Guidelines and the methodology proposed in Luu; Malamud; Xue (2016).

In addition, our corpus can be used as a valuable gold standard dataset for various NLP tasks corresponding to its multiple annotation aspects. For example, conversational speech recognition will benefit from BiRCh's 1-million word transcript aligned with ~270 hours of high-quality[11] audio at the sentence level (see Jurafsky; Martin, 2020 for a review of similar datasets in English, such as SwitchBoard and CALLHOME).

On the theoretical linguistics side, the BiRCh corpus annotated for disfluency and morphosyntax offers unprecedented access to the study of Russian grammar. For instance, retrieving all examples of VP ellipsis, sentential negation, various types of passive constructions, or main clauses with null subjects becomes a matter of a simple search, and observable patterns in the data can be compared against theoretical predictions. Additionally, as the corpus allows for reliable comparisons between monolingual and bilingual adults and their children, it supports stronger theoretical predictions. Below we provide a brief description of several research projects in progress or in the planning stages based on BiRCh data.

As soon as data became available at the early stages of corpus-building, we (members of the BiRCh team) used the audio-aligned disfluency-annotated transcripts to study the properties of two Russian expressions - *aa* and *mm* (Dubinina *et al.*, 2018). These are generally viewed as pause fillers, i.e., non-silent hesitations similar to the English *uh* and *um*, but, as we discovered, can also signal commitment, receipt of message, or call for attention, as in (28):

(12) Mm! Ty golodnaja, aa? - Mm? - Est' xočeš'? - Aa, da.
Mm you.NOM hungry:NOM, eh? Mm? Eat.INF want:2SG? - Oh, yes
'Oh! Are you hungry, eh? - Huh? - Want to eat? - Oh, yes.'

[10] Publicly accessible from https://universaldependencies.org/treebanks/ru_syntagrus/index.html.
[11] 32-bit/48kHz WAV

We discovered that *aa* and *mm* in the parents' speech have distinct distribution patterns, and that there are significant differences between monolingual and bilingual parents in the use of these words as pause fillers, but not in their other functions (Dubinina *et al.*, 2018), which suggests the effect of bilingualism. We are currently exploring correlations between the *aa* and *mm* words in parents' and children's speech.

Two other current BiRCh-based studies rely on both morphological and syntactic annotation and aim to address larger theoretical and language acquisition questions: the lexical politeness marker *požalujsta* 'please' and requests more generally, and the constructions involving verbs marked with the suffix *-sja*. The distribution of various uses of *-sja*, which can have passive, middle, reflexive, reciprocal, and other meanings, in the speech of monolingual parents will shed light on theoretical questions about the syntax and semantics of Russian, while their distribution in the input and output (produced by bilingual children) can answer questions about the development of the syntax-semantics and syntax-pragmatics interfaces in language contact situations (Malamud *et al.*, 2022). Similarly, the study investigating the use of *požalujsta* 'please' (DUbinina *et al.*, in progress) can elucidate the development of politeness strategies in bilingual communities that lead to divergent heritage grammars (Dubinina; Malamud, 2017) and at the same time also advance our understanding of the grammar of speech act modification.

To give a concrete example of syntactic research made possible by BiRCH, we can look at children's acquisition of the Left Branch Extraction (LBE) (Ross, 1967), a type of sub-extraction from NP. LBE is possible in Russian, but is not present in English and German (the two dominant languages in the bilingual group in BiRCh). Van Kampen (1994) discusses a peculiar case of L1 Dutch children producing sentences with LBE, even though adult Dutch lack LBE entirely. She links LBE to the presence of attributive morphology and hypothesizes that the restrictive conditions of poor morphology are acquired slowly, which leaves Dutch children a time window to play with LBE. BiRCh provides perfect means to further test Van Kampen's hypothesis. Since Russian is a morphologically rich language, we expect to find no limitations for LBE, unless they are predated by the impoverishment of the morphological inventory common for bilingual acquisition. We report our study in (Koval *et al.*, 2022).

In closing, we hope that by describing the methodology used for the BiRCh corpus, we show a full range of its potential for research and for informing the creation of other bilingual spoken language corpora. Syntactically annotated child speech corpora in general, and the longitudinal naturalistic bilingual BiRCh corpus in particular, provide an important tool for research in the acquisition of syntax, morphology, and their interfaces with semantics and pragmatics. Ultimately, such research can shed light on the nature of language acquisition itself, in addition to providing a window into language change in children and adults in language contact situations, and into the structure of the monolingual baseline.

REFERENCES

ARSLAN, S. *Neurolinguistic and Psycholinguistic Investigations on Evidentiality in Turkish*. 2015. University of Groningen, 2015.

ARSLAN, S.; BASTIAANSE, R. Chapter 6. First Language Exposure Predicts Attrition Patterns in Turkish Heritage Speakers' Use of Grammatical Evidentiality. *In*: *Studies in Bilingualism*. Edited by Fatih Bayram. Amsterdam: John Benjamins Publishing Company, 2020. pp.105–126.

BECK, J. E. Penn Parsed Corpora of Historical Greek (PPCHiG). Disponível em: https://www.ling.upenn.edu/~janabeck/greek-corpora.html . Acesso em: 22 de julho 2021.

BENMAMOUN, E. et al. Arabic Plurals and Root and Pattern Morphology in Palestinian and Egyptian Heritage Speakers. *Linguistic Approaches to Bilingualism*. v. 4, no. 1, pp.89--123. 2014.

BENMAMOUN, E.; MONTRUL, S.; POLINSKY, M. Prolegomena to Heritage Linguistics. 2010. Disponível em: https://dash.harvard.edu/handle/1/23519841 . Acesso em: 3 de março 2022.

BOGUSLAVSKY, I. et al. Development of a Dependency Treebank for Russian and its Possible Applications in NLP. In: Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02), Las Palmas, Canary Islands - Spain. *Anais*... In: LREC 2002. Las Palmas, Canary Islands - Spain: European Language Resources Association (ELRA), May 2002 Disponível em: http://www.lrec-conf.org/proceedings/lrec2002/pdf/161.pdf . Acesso em: 7 de agosto 2021.

CLARK, A.; FOX, C.; LAPPIN, S. *The Handbook of Computational Linguistics and Natural Language Processing*. John Wiley & Sons, 2013.

DE PRADA PÉREZ, A. *Subject Expression in MinorcaN Spanish: Consequences of Contact with Catalan*. 2009. (Doctoral dissertation) – The Pennsylvania State University, 2009.

DE PRADA PÉREZ, A. First Person Singular Subject Pronoun Expression in Spanish in Contact with Catalan. *In*: *Subject Pronoun Expression in Spanish: A Cross-Dialectal Perspective*, 2015.

DUBININA, I. Y. *et al*. Variability in Input: A Corpus Study of Discourse Markers in Immigrant Parents' Speech. In: Los Angeles, CA. *Anais*... *In*: Panel on Variability and Change in Bilingual Language Acquisition: Longitudinal Perspectives, The Third International Conference on Heritage/Community Languages. Los Angeles, CA: Feb. 2018.

DUBININA, I. Y. *et al*. Razmetka morfologicheskoj informacii BiRCh [BiRCh Morphological annotation guidelines]. Disponível em: https://brandeis.app.box.com/file/451776894902?s=pzyzu57p9bl0s7zkqvsv6ecepwjtp5aj . Acesso em: 30 de setembro 2021.

DUBININA, I. Y. et al. Requests with and without Požalujsta 'Please' in Monolingual and Bilingual Acquisition. Ms., Brandeis University (in progress).

DUBININA, I. Y. *et al*. Razmetka morfologicheskoj informacii BiRCh [BiRCh Morphological Annotation Guidelines]. Disponível em: https://brandeis.app.box.com/file/451776894902?s=pzyzu57p9bl0s7zkqvsv6ecepwjtp5aj. Acesso em: 30 de setembro 2021.

DUBININA, I. Y.; MALAMUD, S. A. Emergent Communicative Norms in a Contact Language: Indirect Requests in Heritage Russian. *Linguistics*. v. 55, no. 1, pp.67–116. 1 Jan. 2017. Disponível em: https://www.degruyter.com/document/doi/10.1515/ling-2016-0039/html Acesso em: 28 de fevereiro 2021.

GALVES, C.; ANDRADE, A. L. de; FARIA, P. *Tycho Brahe Parsed Corpus of Historical Portuguese*, 2017.

GOLDBERG, L. *Verb-Stranding VP Ellipsis: A Cross-Linguistic Study*. 2005. (Doctoral dissertation) – McGill University, Montréal, Québec, Canada, 2005

GRIBANOVA, V. Verb-Stranding Verb Phrase Ellipsis and the Structure of the Russian Verbal Complex. *Natural Language & Linguistic Theory*. v. 31, no. 1, pp.91–136. Feb. 2013. Disponível em: http://link.springer.com/10.1007/s11049-012-9183-3 Acesso em: 22 de setembro 2021.

HAZNEDAR, B. Transfer at the Syntax-Pragmatics Interface: Pronominal Subjects in Bilingual Turkish. *Second Language Research*. v. 26, no. 3, pp.355–378. Jul. 2010. Disponível em: http://journals.sagepub.com/doi/10.1177/0267658310365780 Acesso em: 22 de setembro 2021.

HINDLE, D. Deterministic Parsing of Syntactic Non-fluencies. In: 21st Annual Meeting of the Association for Computational Linguistics, Cambridge, Massachusetts, USA. *Anais*... *In*: ACL 1983. Cambridge, Massachusetts, USA: Association for Computational Linguistics, Jun. 1983. Disponível em: https://www.aclweb.org/anthology/P83-1019 Acesso em: 23 de junho 2021.

IONIN, T.; LUCHKINA, T. Scope, Syntax, and Prosody in Russian as a Second or Heritage Language. *In*: *Exploring Interfaces*. Edited by Mónica Cabrera and José Camacho. Cambridge University Press, 2019, pp.141–170.

IVANOVA-SULLIVAN, T. Anaphora Resolution in Globally Ambiguous Contexts. *In*: *Theoretical and Experimental Aspects of Syntax-Discourse Interface in Heritage Grammars*. Empirical Approaches to Linguistic Theory. Brill, 2014a. pp.125-141.

IVANOVA-SULLIVAN, T. *Theoretical and Experimental Aspects of Syntax-Discourse Interface in Heritage Grammars*. BRILL, 2014b.

JURAFSKY, D.; MARTIN, J. H. Chapter 26: Automatic Speech Recognition and Text-to-Speech. *In*: *Speech and Language Processing (Draft of December 30, 2020)*, 2020.

KEATING, G. D.; VANPATTEN, B.; JEGERSKI, J. WHO WAS WALKING ON THE BEACH?: Anaphora Resolution in Spanish Heritage Speakers and Adult Second Language Learners. *Studies in Second Language Acquisition*. v. 33, no. 2, pp.193–221. Jun. 2011. Disponível em: https://www.cambridge.org/core/product/identifier/S0272263110000732/type/journal_article Acesso em: 22 de setembro 2021.

KOTELNIKOV, E.; RAZOVA, E.; FISHCHEVA, I. A Close Look at Russian Morphological Parsers: Which One Is the Best? Edited by Andrey Filchenkov; Lidia Pivovarova; and Jan Žižka In: Artificial Intelligence and Natural Language, Cham. *Anais*... Cham: Springer International Publishing, 2018.

KOVAL, P. *et al*. The Acquisition of the Left Branch Extraction by Bilingual Russian Children. In: Los Angeles, CA (virtual). *Anais*... *In*: NHLRC Fourth International Conference on Heritage/Community Languages. Los Angeles, CA (virtual): Jun. 2022.

KRAUSE, T.; ZELDES, A. ANNIS3: A New Architecture for Generic Corpus Query and Visualization. *Digital Scholarship in the Humanities*. v. 31, n. 1, pp.118–139. 1 Apr. 2016. Disponível em: https://doi.org/10.1093/llc/fqu057 . Acesso em: 11 de junho 2021.

KROCH, A. *et al*. Penn Parsed Corpora of Historical English. Disponível em: https://www.ling.upenn.edu/hist-corpora/ . Acesso em: 11 de junho 2021.

KROCH, A. *Penn Parsed Corpora of Historical English LDC2020T16*. Philadelphia, 2020.

LUU, A.; MALAMUD, S. A.; XUE, N. Converting SynTagRus Dependency Treebank into Penn Treebank Style. In: Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016), Berlin, Germany. *Anais*... Berlin, Germany: Association for Computational Linguistics, Aug. 2016 Disponível em: https://aclanthology.org/W16-1703 Acesso em: 5 de agosto 2021.

MALAMUD, S. A.; DUBININA, I. Y. Konvencii transkripcii i annotaciya neplavnostej BiRCh [BiRCh guidelines for transcription and disfluency annotation]. Disponível em: https://brandeis.app.box.com/s/h15um924ygz3t5zdvfwmsdx5kesjrzoq Acesso em: 23 de junho 2021a.

MALAMUD, S. A.; DUBININA, I. Y. Konvencii segmentacii transkripcii na predlozheniya v BiRCh [BiRCh convensions for segmenting transcripts into sentences]. Disponível em: https://brandeis.app.box.com/file/297247548157?s=woyvgzm21u28tm43anvlda9hqla0491c Acesso em: 30 de setembro 2021b.

MALAMUD, S. A. *et al*. Russian "sja" Verbs in Bilingual and Monolingual Acquisition. In: Los Angeles, CA (virtual). *Anais*... *In*: NHLRC Fourth International Conference on Heritage/Community Languages. Los Angeles, CA (virtual): Jun. 2022.

MARCUS, M. P. *et al*. Treebank-3Linguistic Data Consortium, 1999. Disponível em: https://catalog.ldc.upenn.edu/LDC99T42 Acesso em: 5 de agosto 2021

MARCUS, M. P.; SANTORINI, B.; MARCINKIEWICZ, M. A. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*. v. 19, no. 2, pp.313–330. 1993. Disponível em: https://aclanthology.org/J93-2004 Acesso em: 5 de agosto 2021.

MARTINEAU, F. Un corpus pour l'analyse de la variation et du changement linguistique. *Corpus*. no. 7. 10 Nov. 2008. Disponível em: http://journals.openedition.org/corpus/1508 Acesso em: 22 de setembro 2021.

MONTRUL, S. Subject and Object Expression in Spanish Heritage Speakers: A Case of Morphosyntactic Convergence. *Bilingualism: Language and Cognition*. v. 7, no. 2, pp.125–142. Aug. 2004. Disponível em: https://www.cambridge.org/core/product/identifier/S1366728904001464/type/journal_article Acesso em: 22 de setembro 2021.

NAGY, N. G. *et al*. Null Subjects in Heritage Languages: Contact Effects in a Cross-linguistic Context. *In*: *Anais*...2011.

OŽEGOV, S. I.; ŠVEDOVA, N. Ju. Explanatory Dictionary of the Russian Language. Disponível em: https://dic.academic.ru/dic.nsf/ogegova/ Acesso em: 2 de setembro 2021.

PÕLDVERE, N. *et al*. Challenges of Releasing Audio Material for Spoken Data: The Case of the London–Lund Corpus 2. *Research in Corpus Linguistics*. v. 9, no. 1, pp.35–62. 7 Jun. 2021. Disponível em: https://ricl.aelinco.es/index.php/ricl/article/view/157 Acesso em: 16 de julho 2021.

POLINSKY, M. Reanalysis in Adult Heritage Language: New Evidence in Support of Attrition. *Studies in Second Language Acquisition*. v. 33, no. 2, pp.305–328. Jun. 2011. Disponível em: https://www.cambridge.org/core/journals/studies-in-second-language-acquisition/article/reanalysis-in-adult-heritage-language/FC20F543D25513287F4FC8CB3E0B6ACF Acesso em: 27 de setembro 2021.

POLINSKY, M. Structure vs. Use in Heritage Language. *Linguistics Vanguard*. v. 2, no. 1. 1 Dec. 2016. Disponível em: https://www.degruyter.com/document/doi/10.1515/lingvan-2015-0036/html. Acesso em: 22 de setembro 2021

POLINSKY, M. *Heritage Languages and Their Speakers*. Cambridge University Press, 2018.

POLINSKY, M.; KAGAN, O. Heritage Languages: In the 'Wild' and in the Classroom: Heritage Languages: In the 'Wild' and in the Classroom. *Language and Linguistics Compass*. v. 1, no. 5, pp.368-395. Sep. 2007. Disponível em: https://onlinelibrary.wiley.com/doi/10.1111/j.1749-818X.2007.00022.x Acesso em: 22 de setembro 2021.

POLINSKY, M.; SCONTRAS, G. Understanding Heritage Languages. *Bilingualism: Language and Cognition*. v. 23, no. 1, pp.4–20. Jan. 2020. Disponível em: https://www.cambridge.org/core/product/identifier/S1366728919000245/type/journal_article Acesso em: 22 de setembro 2021.

POPLACK, S. *et al*. Revisiting Phonetic Integration in Bilingual Borrowing. *Language*. v. 96, no. 1, pp.126–159. 2020. Disponível em: https://muse.jhu.edu/article/751035. Acesso em: 7 de março 2022.

RAKHILINA, E.; VYRENKOVA, A.; POLINSKY, M. Linguistic Creativity in Heritage Speakers. *Glossa: a journal of general linguistics*. v. 1, no. 1, p.43. 26 Oct. 2016. Disponível em: http://www.glossa-journal.org/article/10.5334/gjgl.90/ Acesso em: 22 de setembro 2021.

RANDALL, B.; TAYLOR, A.; KROCH, A. *CorpusSearch 2*. 2005.

RNC. Russian National Corpus. Disponível em: https://ruscorpora.ru/new/en/index.html Acesso em: 2 de setembro 2021.

ROSS, J. R. *Constraints on Variables in Syntax*. 1967. MIT, Cambridge, Massachusetts, USA, 1967.

ROWLAND, C. F.; FLETCHER, S. L.; FREUDENTHAL, D. How Big Is Big Enough? Assessing the Reliability of Data from Naturalistic Samples. *Corpora in Language Acquisition Research*. 9 Apr. 2008. Disponível em: https://www.jbe-platform.com/content/books/9789027290267-tilar.6.04row Acesso em: 27 de setembro 2021.

SANTORINI, B. Syntactic Annotation Manual for the Penn Historical Corpora and the York-Helsinki Corpus of Early English Correspondence. Disponível em: https://www.ling.upenn.edu/hist-corpora/annotation/index.html Acesso em: 11 de junho 2021.

SANTORINI, B.; DIERTANI, A. Syntactic Annotation Manual for Audio-Aligned Parsed Corpora. Disponível em: https://www.ling.upenn.edu/~beatrice/annotation-audio-aligned-corpora/index.html Acesso em: 11 de junho 2021.

SEDDAH, D. *et al*. Overview of the SPMRL 2013 Shared Task: A Cross-Framework Evaluation of Parsing Morphologically Rich Languages. In: Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages, Seattle, Washington, USA. *Anais*... Seattle, Washington, USA: Association for Computational Linguistics, Oct. 2013. Disponível em: https://aclanthology.org/W13-4917 Acesso em: 7 de agosto 2021.

SEDDAH, D.; KÜBLER, S.; TSARFATY, R. Introducing the SPMRL 2014 Shared Task on Parsing Morphologically-rich Languages. *In*: Proceedings of the First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages, Dublin, Ireland. *Anais*... Dublin, Ireland: Dublin City University, Aug. 2014. Disponível em: https://aclanthology.org/W14-6111 Acesso em: 7 de agosto 2021.

SEKERINA, I. A.; SAUERMANN, A. Visual Attention and Quantifier-Spreading in Heritage Russian Bilinguals. *Second Language Research*. v. 31, no. 1, pp.75–104. Jan.

2015. Disponível em: http://journals.sagepub.com/doi/10.1177/0267658314537292 Acesso em: 22 de setembro 2021.

SERRATRICE, L.; SORACE, A.; PAOLI, S. Crosslinguistic Influence at the Syntax–Pragmatics Interface: Subjects and Objects in English–Italian Bilingual and Monolingual Acquisition. *Bilingualism: Language and Cognition*. v. 7, no. 3, pp.183–205. Dec. 2004. Disponível em: https://www.cambridge.org/core/product/identifier/S1366728904001610/type/journal_article Acesso em: 22 de setembro 2021.

TORTORA, C. The Audio-Aligned and Parsed Corpus of Appalachian English: Design and Use. *In*: WORKSHOP ON DATABASES AND CORPORA IN LINGUISTICS. Stony Brook University, NY, 17 Oct. 2014. Disponível em: https://aapcappe.commons.gc.cuny.edu/wp-content/blogs.dir/3140/files/2019/03/tortora_sb_corpus_handout_101614.pdf Acesso em: 10 de junho 2021.

TORTORA, C. *et al*. The Audio-Aligned and Parsed Corpus of Appalachian English (AAPCAppE), version 0.1. Disponível em: https://www.aapcappe.org/ Acesso em: 11 de junho 2021.

TORTORA, C. *et al*. Corpus of New York City English (CUNY-CoNYCE). Disponível em: https://conyce.commons.gc.cuny.edu/

TORTORA, C.; SANTORINI, B.; BLANCHETTE, F. Romance Parsed Corpora: Editors' Introduction. *Linguistic Variation*. v. 18, no. 1, pp.1–22. 1 Jan. 2018. Disponível em: https://www.jbe-platform.com/content/journals/10.1075/lv.00002.tor#html_fulltext Acesso em: 11 de junho 2021.

TSIMPLI, I. et al. First Language Attrition and Syntactic Subjects: A Study of Greek and Italian near-Native Speakers of English. *International Journal of Bilingualism*. v. 8, no. 3, pp.257–277. Sep. 2004. Disponível em: http://journals.sagepub.com/doi/10.1177/13670069040080030601 Acesso em: 22 de setembro 2021.

UD POS. Universal Dependencies POS tags. Disponível em: https://universaldependencies.org/u/pos/index.html Acesso em: 2 de setembro 2021.

UNSWORTH, S. *et al*. The Role of Age of Onset and Input in Early Child Bilingualism in Greek and Dutch. *Applied Psycholinguistics*. v. 35, no. 4, pp.765–805. Dec. 2012. Disponível em: https://www.cambridge.org/core/journals/applied-psycholinguistics/article/role-of-age-of-onset-and-input-in-early-child-bilingualism-in-greek-and-dutch/1B686FAC86608EB5F4EBB5F44B9B0FF0 Acesso em: 30 de setembro 2021.

UNSWORTH, S. Bilingual Language Exposure Questionnaire. Disponível em: https://www.iris-database.org/iris/app/home/detail?id=york%3A928327&ref=search

UŠAKOV, D. N. Explanatory Dictionary of the Russian Language. Disponível em: https://dic.academic.ru/contents.nsf/ushakov Acesso em: 2 de setembro 2021.

VAN GOMPEL, M. *et al*. FoLiA in Practice: The Infrastructure of a Linguistic Annotation Format. *In*: *CLARIN in the Low Countries*. Edited by Jan Odijk and Arjan van Hessen. Ubiquity Press, 2017. pp.71–82.

VAN GOMPEL, M.; REYNAERT, M. FoLiA: A Practical XML Format for Linguistic Annotation – a Descriptive and Comparative Study. *Computational Linguistics in the Netherlands Journal*. v. 3, pp.63–81. 1 Dec. 2013. Disponível em: https://www.clips.uantwerpen.be/clinjournal/clinj/article/view/26 Acesso em: 10 de junho 2021.

VAN KAMPEN, J. The Learnability of the Left Branch Condition. *Linguistics in the Netherlands*. v. 11, pp.83–94. 6 Oct. 1994. Disponível em: http://www.jbe-platform.com/content/journals/10.1075/avt.11.10kam Acesso em: 22 de setembro 2021.

WALLENBERG, J. C. *et al*. *Icelandic Parsed Historical Corpus (IcePaHC)*, 2011.

ZALIZNYAK, A. A. A Grammatical Dictionary of the Russian Language. Disponível em: https://www.morfologija.ru Acesso em: 2 de setembro 2021.

ZIPSER, F.; ROMARY, L. A model oriented approach to the mapping of annotation formats using standards. *In*: *Anais... In*: Workshop on Language Resource and Language Technology Standards, LREC, 2010, 18 May. Disponível em: https://hal.inria.fr/inria-00527799 Acesso em: 11 de junho 2021.

Statement of Authors' Contribution

All authors (Alex Luu, Pasha Koval, Sophia A. Malamud and Irina Y. Dubinina) made substantial contributions to the elaboration of the article "Creating a Large-Scale Audio-Aligned Parsed Corpus of Bilingual Russian Child and Child-Directed Speech (BiRCh): Challenges, Solutions, and Implications for Research", fully covering the following aspects: 1) conception, analysis and interpretation of data; 2) drafting and revising the article critically for important intellectual content; 3) final approval of the version to be published, 4) accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. While all authors worked on the entire manuscript, the following sections received particular focus from the following authors: section 1 – Alex Luu, section 2 – Irina Y. Dubinina, section 3 – Sophia A. Malamud, section 4 – Irina Y. Dubinina and Sophia A. Malamud, section 5 – Pasha Koval, the concluding section – Irina Y. Dubinina, Sophia A. Malamud and Alex Luu.

Received October 01, 2021
Accepted August 29, 2022

Reviews

Due to the commitment assumed by *Bakhtiniana*. Revista de Estudos do Discurso [Bakhtiniana. Journal of Discourse Studies] to Open Science, this journal only publishes reviews that have been authorized by all involved.

Research Data and Other Materials Availability

The contents underlying the research text are included in the manuscript.

APPENDIX

Abbreviations used in the text

Abbreviation	English	Portuguese
AAPCAppE	the Audio-Aligned and Parsed Corpus of Appalachian English	*Corpus* do inglês de Apalaches analisado e alinhado com arquivos de áudio
BiLec	the Bilingual Language Exposure Calculator	Calculadora de exposição à língua
BiRCh	the corpus of Bilingual Russian Child Speech	*Corpus* de fala em russo de crianças bilíngues
CS	CorpusSearch 2	Busca pelo banco de dados 2
HL	heritage language	língua de herança
HS	heritage speaker	falante de herança
LBE	left branch extraction	extração da posição esquerda ao núcleo
NS	native speaker	falante native
POS	part-of-speech	classe grammatical
PPCHE	Penn Parsed Corpora of Historical English	*Corpora* analisados do inglês histórico de Penn
RNC	the Russian National Corpus	*Corpus* nacional do russo
UD	Universal Dependencies	dependências universais
Labels of syntactic constituents:		
ADJP	adjective phrase	sintagma adjectival
CP	complementizer phrase	sintagma de complemento
IP	inflectional phrase	sintagma flexional
NP	noun phrase	sintagma nominal
NumP	number phrase	sintagma numérico
PP	prepositional phrase	sintagma preposicional
VP	verb phrase	sintagma verbal