D.O.I. http://dx.doi.org/10.1590/2176-4573p55831

ARTIGOS

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 / Creating a Large-Scale Audio-Aligned Parsed Corpus of Bilingual Russian Child and Child-Directed Speech (BiRCh): Challenges, Solutions, and Implications for Research

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

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 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

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,

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

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

Introdução

Este artigo apresenta o *corpus* de fala de crianças bilíngues em russo (BiRCh, *Bilingual Russian Child Speech*, http://birch.ling.brandeis.edu), que se encontra no processo de construção na Universidade Brandeis (Waltham, MA, EUA). O projeto envolve 10 crianças bilíngues (na idade entre dois e nove anos) de nove famílias nas quais se fala a língua russa e que residem nos EUA, Canadá e Alemanha, representando duas situações de contato linguístico com os dois idiomas diferentes dominantes na sociedade (inglês e alemão); e também as crianças de cinco famílias monolíngues (quatro da Rússia e uma da Ucrânia), pareadas pela idade com os participantes bilíngues, como um grupo controle. BiRCh integra as gravações de áudio de interações naturalísticas entre as crianças e seus cuidadores (geralmente os pais) em contextos familiares. As gravações de áudio são distribuídas entre os três grupos de idiomas dominantes na sociedade de maneira equilibrada. O *corpus* inclui as transcrições de um milhão de palavras das gravações de áudio com as informações sobre as disfluências de fala (principalmente falsos inícios) e fenômenos discursivos (como as elaborações e repetições intra-sentenciais), sendo que uma subdivisão com 500 mil palavras é morfologicamente anotada e sintaticamente analisada. Além disso, o *corpus* fornece as informações sociolinguísticas sobre cada família participante, tais como, a quantidade e o tipo de contato linguístico que as crianças têm com a língua falada em casa e a língua da sociedade, grau de escolaridade dos pais e sua proficiência na língua dominante da sociedade, etc. Todas as transcrições são alinhadas com o sinal de áudio, e os dados

anotados são conectados ao áudio e à transcrição, o que torna possível realizar as pesquisas tanto textuais quanto gramaticais para encontrar no áudio um ponto relevante.

Uma característica singular e crucialmente importante do BiRCh é a anotação morfológica e sintática detalhada. Em uma língua morfologicamente rica como o russo, muitos fenômenos linguísticos são impossíveis de serem estudados sem as informações morfológicas detalhadas que se estendam além da marcação da classe gramatical de palavras (mais adiante, POS, *part-of-speech*) e sem a anotação sintática. Por exemplo, o estudo da voz passiva na língua russa inclui a investigação de três tipos de construções: aquelas que são formadas com a ajuda de uma forma verbal de particípio passivo, aquelas que incluem o sufixo multifuncional *-sia*, e as construções impessoais que muitas vezes têm um significado passivo, mas têm uma forma verbal ativa com o sujeito nulo. A busca por sujeitos nulos e particípios usados no contexto da voz passiva requer a anotação morfológica e sintática.

O BiRCh é o primeiro projeto desse tipo[1], desenvolvido de forma única para investigar fatores que afetam o desenvolvimento e a mudança da competência gramatical em crianças bilíngues, pois é baseado em dados longitudinais que ocorrem naturalmente, desde a primeira infância. O *corpus* traça os caminhos de aquisição da língua de crianças bilíngues e monolíngues antes do momento em que a assimetria entre o *input* e o uso da língua começa a crescer em contextos bilíngues no início da educação formal (BENMAMOUN; MONTRUL; POLINSKY, 2010). Ele registra uma ampla gama de fenômenos linguísticos em múltiplas instâncias de uso por crianças participantes e seus pais, o que facilita as generalizações estatisticamente significativas, as comparações viáveis e as correlações confiáveis ao comparar pais bilíngues e monolíngues, crianças bilíngues e monolíngues e pais bilíngues e seus filhos. Além disso, a fala dos pais registrada no *corpus* BiRCh apresenta dados importantes não apenas para o estudo das propriedades de *input*, mas também para a investigação das mudanças de linguagem que ocorrem ao longo da vida dos pais adultos bilíngues e para as comparações entre os diferentes tipos de bilingues.

[1] Gostaríamos de agradecer o projeto imprescindível RUEG (Emerging Grammars in Language-Contact Situations: A Comparative View) que está sendo realizado na Alemanha; , no entanto, ao contrário de BiRCh, ele é baseado em amostras de fala produzidas em contextos experimentais por adolescentes (14-18 anos de idade) e jovens adultos.

Para a pesquisa em aquisição da linguagem, o BiRCh será uma fonte de dados fecunda e importante porque oferece um olhar mais atento aos desvios da trajetória da aquisição da linguagem monolíngue à medida que essas se acumulam ao longo do tempo e (potencialmente) levam a uma gramática de herança. Os falantes de herança[2] (HSs, *heritage speakers*) adultos são frequentemente comparados às crianças aprendizes de línguas, e de fato, os linguistas identificaram várias áreas da gramática em que ambos os grupos parecem apresentar alguns padrões em comum e distinguir-se dos falantes de L1 adultos (BENMAMOUN et al., 2014; ARSLAN, 2015; SEKERINA; SAUERMANN, 2015; ARSLAN; BASTIAANSE, 2020). No entanto, essas características convergentes não devem ser tomadas como evidência de que a gramática da língua de herança 'congelou' no meio do caminho para a gramática de falante de L1 adulto (para as evidências de reanálise, ver, por exemplo, POLINSKY, 2011).

Há pelo menos quatro processos que podem resultar em uma gramática de herança de falante adulto que se diferenciam da linha de base da gramática de um falante adulto da mesma língua como L1. As mudanças na gramática de herança podem ser causadas por características da língua dominante (transferência de idioma). Ou, de outra maneira, a gramática de L1 de herança pode inicialmente convergir com a gramática de L1 de um falante adulto e, posteriormente, perder alguns traços ou alterar-se após um período de desuso (atrito linguístico). Uma terceira possibilidade é que a gramática de herança de falante adulto ofereça uma solução diferente da gramática de L1 de falante adulto em termos do *input* linguístico que ambos recebem (obtenção divergente). Por fim, o *input* para o processo de aquisição – fala dos pais bilíngues e monolíngues – pode variar como o resultado de mudança no comportamento linguístico dos pais bilíngues (*input* diferente). Compreender a trajetória da aquisição da língua de herança é um componente necessário para diagnosticar e desembaraçar esses processos e, por sua vez, para obter mais informações sobre a natureza da aquisição e do conhecimento sobre a língua no geral. Um *corpus* longitudinal gramaticalmente anotado é uma ferramenta crucial para as pesquisas desse ramo.

Nas próximas seções, apresentamos em detalhes a metodologia para a construção do *corpus* BiRCh, incluindo o esboço geral da organização dos processos

[2] Os falantes de línguas de herança (HSs) são definidos como 'bilíngues simultâneos ou sequenciais (sucessivos) cuja língua mais fraca corresponde à língua minoritária de sua sociedade e cuja língua mais forte é a língua dominante dessa sociedade' (POLINSKY, 2018, p.9).

envolvidos na construção do *corpus*, descrevemos a coleta de dados, transcrição, anotação inicial (incluindo os fenômenos bilíngues, discursivos e de disfluência), anotação morfológica e análise sintática. Em cada seção, abordamos as dificuldades associadas à construção de um *corpus* profundamente anotado e descrevemos as nossas soluções para o desafio de encontrar um equilíbrio entre a construção de um recurso copioso e confiável e a necessidade de conciliar isso com os recursos finitos. Por fim, apresentamos os exemplos de uso atual do corpus BiRCh e as sugestões de aplicação.

1 O esquema geral do pipeline de desenvolvimento do *corpus*

A figura 1 mostra o esquema geral do pipeline de desenvolvimento do *corpus* BiRCh e os resultados obtidos em cada estágio (cf. PÕLDVERE et al., 2021 para uma tentativa recente de criar um *corpus* de amostras de fala em inglês britânico com as transcrições e áudio alinhados).

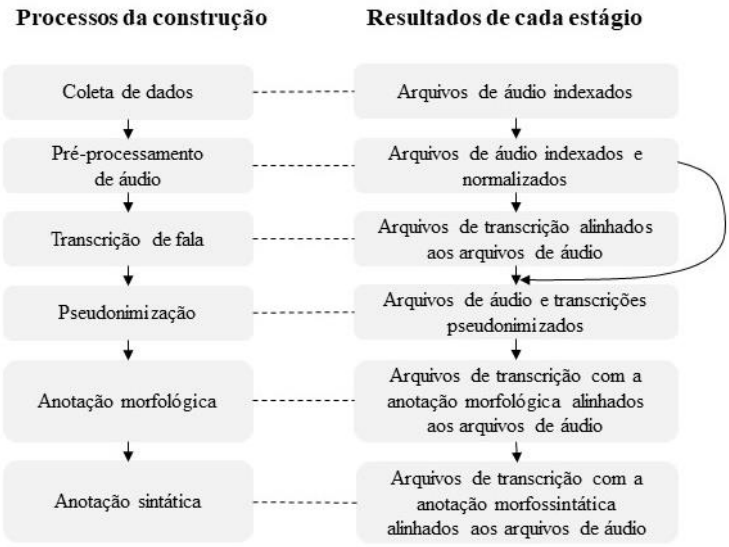


Figura 1. O esquema geral do pipeline de desenvolvimento do *corpus* BiRCh

Na etapa de coleta de dados, cada arquivo de áudio é indexado com um formato de nome padronizado. Por exemplo, o nome da quarta gravação da família da criança S no momento em que ela tinha 4 anos, 6 meses e 9 dias é S_4-6-9_3 (o número final começa com 0). No estágio de pré-processamento de áudio, cada arquivo de áudio indexado é previamente processado: a pessoa responsável junta as gravações cronologicamente adjacentes, organiza os arquivos removendo os fragmentos de áudio

sem palavras ou aumentando o volume. O nome de cada arquivo de áudio normalizado é usado como o identificador exclusivo para todos os arquivos derivados nos próximos estágios.

A transcrição de fala consiste em três etapas principais executadas manualmente por diferentes falantes nativos (NS) no ELAN, uma ferramenta padrão de acesso aberto para a anotação textual alinhada ao tempo de gravações multimídia:

- Segmentação inicial da fala, transcrição e anotação inicial com a marcação de
  - inícios falsos (a principal anotação em disfluência),
  - os fenômenos discursivos que complicam a análise sintática (como as estruturações intra-sentenciais, as orações parentéticas e as repetições intra-sentenciais com a intenção retórica), e
  - os fenômenos bilíngues (como os empréstimos e a troca de código).
- A revisão das transcrições de fala e as anotações iniciais (muitas vezes também a manipulação das transcrições de segmentos[3] com troca de código em alemão ou inglês).
- Revisão da segmentação de fala.

Este é o processo minimamente viável para garantir a qualidade da anotação dentro do padrão de excelência[4]. Tomamos a decisão de não usar as ferramentas automáticas em nenhuma etapa, principalmente, devido à peculiaridade de nossos dados: a fala de criança e a fala dirigida à criança (que não envolve o uso de inglês) por famílias russas bilíngues e monolíngues no ambiente de casa. Corrigir a transcrição automática para esse tipo de dados exigiria mais trabalho do que fazer a transcrição manual desde o começo. O resultado obtido na etapa da transcrição de fala é um conjunto de arquivos com as transcrições (no formato de anotação ELAN baseado em XML, isto é, EAF), alinhados no tempo com os arquivos de áudio no nível de segmentação, com a marcação de tempo nos momentos em que aparecem as informações pessoalmente identificáveis, que são substituídos em arquivos de áudio e transcrição no próximo passo da pseudonimização.

Os arquivos de transcrição pseudonimizados posteriormente são usados na etapa de anotação morfológica, incluindo:

- A tokenização automática, isto é, a divisão de segmentos em listas de tokens de palavras.
- A anotação morfológica automática de tokens de palavras em russo, que consiste em lematização e marcação de características morfológicas e POS.
- A revisão manual de anotação morfológica.

[3] Nós denominamos os tokens do nível de sentença como ‘segmentos’

[4] Em linguística computacional, ‘padrão de excelência’ refere-se à precisão e consistência da anotação humana.

- A revisão final manual da anotação morfológica.

Desenvolvemos o nosso próprio tokenizador baseado em regras e uma ferramenta de marcação morfológica para maximizar o ajuste com a transcrição feita pela nossa equipe e com os marcadores de anotação morfológica, respectivamente. A ferramenta de marcação morfológica usa como o seu núcleo Mystem, a melhor opção em termos de desempenho e a clareza na organização do conjunto de marcadores[5] entre as ferramentas de marcação morfológica mais usadas para a língua russa (KOTELNIKOV; RAZOVA; FISHCHEVA, 2018). Esse núcleo é continuamente enriquecido por regras complementares que surgem ao longo da prática de anotação. O resultado obtido na etapa de processamento automático é um conjunto de arquivos salvos no formato FoLiA baseado em XML. Esse formato destaca-se por sua versatilidade, facilidade de compreensão e portabilidade, pois acopla vários tipos de anotação linguística com os conjuntos arbitrários de marcadores e inclui todas as camadas de anotação em um único arquivo (VAN GOMPEL; REYNAERT, 2013). Também é acompanhado por FLAT, uma ferramenta de anotação baseada na tecnologia *web*, cuja interface de usuário pode mostrar diferentes camadas de anotação linguística ao mesmo tempo (VAN GOMPEL et al., 2017). Os arquivos resultantes do processamento automático são corrigidos manualmente e, em seguida, revisados por diferentes anotadores no formato FLAT, o que garante a qualidade dentro do padrão de excelência. Os softwares que reconhecem o formato FoLiA, como FoLiApy e FoLiA-tools, nos permitem manipular facilmente os dados em qualquer ponto do ciclo de construção do *corpus* (por exemplo, após as revisões das regras de anotação), tornando o fluxo de trabalho mais flexível e interativo. A escolha por formato FoLiA não só nos permite alinhar todos os aspectos de anotação em um único arquivo (cf. TORTORA, 2014), facilitando a implementação de quaisquer revisões de anotação futuras, mas também potencializa a acessibilidade do BiRCh à análise computacional ou anotação futura.

Por fim, o estágio de anotação sintática envolve as iterações intercaladas de duas etapas principais:

- A análise sintática automática.

[5] O conjunto de marcadores de Mystem é construído com a base no dicionário gramatical da língua russa de Zaliznyak e é usado para a análise morfológica do corpus nacional do russo. Também é usado no projeto RUEG.

- A revisão manual de análises sintáticas.

À medida que seguimos a metodologia de ponta de construção de *corpora* com anotação sintática existentes que focam na variação e mudança da língua (TORTORA; SANTORINI; BLANCHETTE, 2018), tais como os *Penn Parsed Corpora of Historical English* (PPCHE, Corpora analisados do inglês histórico de Penn) (KROCH et al., 2016) e *Audio-Aligned and Parsed Corpus of Appalachian English* (AAPCAPPE, Corpus do inglês de Apalaches analisado e alinhado com arquivos de áudio) (TORTORA et al., 2017), organizamos a análise sintática de cada segmento em pelo menos uma árvore de estrutura de frase, usando o software CS (*CorpusSearch 2*, Busca pelo banco de dados 2) (RANDALL; TAYLOR; KROCH, 2005) desenvolvido para a análise sintática automática, baseada em regras, e fazemos a revisão manual orientada pela busca. O nosso primeiro passo é converter os dados morfologicamente anotados do formato FoLiA para o formato do banco de dados *Penn Treebank*, mantendo os identificadores de todos os segmentos e *tokens* de palavras para a integração da anotação sintática nos arquivos FoLiA alinhados ao áudio e morfologicamente anotados. Assim, o resultado da etapa de anotação sintática inclui tanto os arquivos de análise sintática no formato *Penn Treebank* quanto os arquivos FoLiA integrados. Estes últimos são usados[6] em seguida para a implantação de uma interface de pesquisa e visualização baseadas na tecnologia *web* no ANNIS, uma ferramenta de código aberto bem desenvolvida e especializada nos *corpora* com múltiplas camadas de anotação linguística (KRAUSE; ZELDES, 2016). Nossa pesquisa interna mostra que a linguagem de consulta usada no ANNIS (AQL) é capaz de cobrir todas as funções de pesquisa implementadas no CS e, portanto, fornece aos usuários no mínimo a mesma facilidade de fazer a busca pelas estruturas sintáticas em foco. Além disso, como a expressividade do AQL não é sensível aos tipos de anotação, BiRCh pode ser explorado de novas maneiras em comparação com os seus antecessores sob o paradigma do *Penn Treebank.*

Esse esquema geral da organização dos processos na construção do *corpus* tinha sido continuamente otimizado até se tornar estável. Usamos uma plataforma de gerenciamento de projetos completa[7] para organizar, registrar, comunicar e analisar

[6] Colaboramos com Maarten Van Gompel, o principal autor de FoLiA, para converter FoLiA em SaltXML (ZIPSER; ROMARY, 2010), que por sua vez pode ser integrado ao ANNIS.

[7] Os recursos mais proveitosos incluem os modelos de tarefas que podem ser reutilizados por diferentes membros da equipe para criar as tarefas do mesmo tipo; e relatórios de tarefas com as informações de

todas as contribuições. Os recursos de bate-papo e as ferramentas desenvolvidas a base de conhecimento acumulado fornecem um espaço central para levantar as perguntas e respondê-las.

2 Coleta de dados

Este projeto não seria possível sem a boa vontade, o interesse e o comprometimento das famílias participantes que desempenharam um papel primordial na criação do *corpus* BiRCh. As mães geralmente eram a força que impulsionava a decisão da família a participar da pesquisa. Na maioria das famílias, as mães têm formação linguística, pedagógica ou filológica e foram motivadas por seus interesses profissionais. Para as famílias bilíngues, a motivação para participar incluía o engajamento com o bilinguismo de seus filhos e, para todas as famílias, um interesse genuíno no avanço da pesquisa linguística desempenhou um papel decisivo. Mantivemos o contato com os participantes por meio de contato pessoal regular e de e-mails informativos anuais do projeto BiRCh, nos quais relatamos o progresso do projeto até o momento e fornecemos às famílias as orientações baseadas em pesquisas para ajudar no desenvolvimento linguístico de seus filhos. Também compartilhamos alguns dos arquivos anotados com cada família para motivar a participação nas próximas etapas.

Cada família participante foi solicitada a fazer gravações de áudio semanais de interações verbais com seu filho com a duração de pelo menos 30 minutos. Esse cronograma de gravação continuava pelo maior tempo possível, com os intervalos para as férias de verão e inverno. A duração média da participação é de 3,26 anos, e a participação ininterrupta mais longa durou sete anos. Inicialmente, as famílias receberam os gravadores *Sony* de alta qualidade. No primeiro ano do projeto, mudamos para o dispositivo de gravação profissional *ZoomH2n* e definimos um requisito para que todos os arquivos de áudio sejam salvos no formato WAV para garantir a qualidade dos dados acústicos. Ambas as etapas garantiram que os dados BiRCh possam ser usadas para a pesquisa fonológica.

rastreamento de tempo, que nos permitem calcular a carga de trabalho e a eficiência de diferentes tarefas executadas por diferentes membros da equipe para otimizar o nosso fluxo de trabalho e ajustar o nosso orçamento.

No segundo ano do projeto, testamos os dados transcritos até o momento para a presença de fenômenos linguísticos de baixa frequência na fala infantil, como as construções passivas, e descobrimos que o nosso cronograma de gravações na época captava apenas cerca de 1% das formas linguísticas de baixa frequência. Com base nas pesquisas anteriores (ROWLAND; FLETCHER; FREUDENTHAL, 2008), convidamos uma família em cada grupo com criança de 4 anos de idade a participar de um regime de gravações mais denso: de 3 a 7 horas por semana (os participantes do *corpus* denso da Alemanha só podiam gravar de 1,5 a 3,5 horas por semana). Para tornar a participação densa mais fácil possível, fornecemos a essas três famílias gravadores em miniatura *ATTO Digital* que podiam ser colocados na roupa da criança e cuja carga de bateria era de até seis horas entre as gravações. Esse cronograma de gravação continuou por seis meses, depois os voluntários do *corpus* denso puderam retornar ao cronograma regular de gravação.

Ademais, uma vez a cada dois anos, coletamos os dados sociolinguísticos sobre todas as famílias participantes. No início, coletamos as informações básicas sobre o perfil linguístico de família, incluindo o local de nascimento dos pais e da criança, o local da residência atual, com quem a criança convive no domicílio e a idade de cada criança, participante ou não do projeto. Também usamos o questionário *Bilingual Language Exposure Calculator* BiLec, Calculadora de exposição à língua (UNSWORTH et al., 2012; UNSWORTH, 2016) para reunir as informações etnográficas e sociolinguísticas aprofundadas sobre as famílias bilíngues. Este questionário inclui perguntas detalhadas sobre a quantidade de exposição a cada um dos idiomas da criança, incluindo a porcentagem de uso diário para cada idioma, os níveis de proficiência dos pais e outros cuidadores, e também contato passivo com a língua, tal como o tempo de contato com a TV ou audiolivros.

3 Anotação inicial e segmentação

A transcrição, segmentação e anotação inicial ocorrem como um único processo, o que requer as diretrizes claras para obter a consistência. O princípio fundamental dos procedimentos adotados é permitir a recuperação confiável de exemplos por futuros usuários do *corpus*, minimizando o trabalho adicional dos anotadores. O objetivo final é

produzir um *corpus* analisado gramaticalmente, incluindo a análise sintática. Nós, portanto, anotamos apenas aqueles fenômenos de disfluência, discurso e bilinguismo que, se não estivessem marcados, poderiam interferir na anotação morfológica e sintática. A anotação inicial de BiRCh (MALAMUD; DUBININA, 2017a) e as Diretrizes de Segmentação de BiRCh (MALAMUD; DUBININA, 2017b) são baseadas nas diretrizes AAPCAppE (SANTORINI; DIERTANI, 2017), que por sua vez são baseadas nas diretrizes PPCHE (SANTORINI, 2016) e na discussão em Hindle (1983). Esclarecemos as categorias existentes na AAPCAppE e PPCHE e adicionamos categorias novas específicas à natureza da fala de criança e da fala dirigida à criança assim como à fala bilíngue.

Transcrevemos as amostras em russo usando alfabeto cirílico (codificação UTF-8) e usamos o alfabeto latino para alemão e inglês. Estabelecemos grafias padronizadas para os fenômenos de linguagem difíceis de transcrever, como os preenchimentos de pausa e outras interjeições, por exemplo, *aa*, *mm*, *nea* (‘*nope*’) para garantir que sejam encontrados na busca, o que é crucial para as futuras pesquisas sobre as disfluências.

Não realizamos uma anotação de disfluências completa (pausas, correções, etc.), em vez disso, focamos no que Hindle (1983) chamou de ‘não-fluências sintáticas’. A nossa categoria de disfluência principal é falso início; na anotação inicial também marcamos as repetições (repetições exatas com a intenção retórica), elaborações (repetições não exatas, paráfrases, que não equivalem a orações principais completas e que esclarecem constituintes que não são frases completas) e orações parentéticas.

Essas categorias de não fluência sintática identificam constituintes que não se encaixam perfeitamente no algoritmo de anotação sintática. A análise sintática inicial os ignora e, portanto, é simplificada, mas as estruturações e as orações parentéticas são posteriormente analisadas e se tornam rótulos sintáticos no *corpus* final (SANTORINI; DIERTANI, 2017). As nossas diretrizes fornecem esclarecimento amplo e introduzem algumas mudanças na definição de elaborações e orações parentéticas. A última categoria em nosso *corpus* abrange dois tipos de orações: (i) as orações parentéticas ou periféricas que representam um comentário, e (ii) as elaborações que equivalem a orações principais completas. Não marcamos as construções parentéticas que não representavam uma oração.

(1) Oj (PAREN ty znaeš') kogda ja byla devočkoj
Oh você.SG sabe quando eu era:F menina:INS.SG

(PAREN mne naverno četyre godika bylo)
Eu.DAT talvez quatro anos.DIM tinha.N
deduška Saša (ELAB moj papa) prines
vovô Sasha meu pai PRF:trouxe.PST.M.SG
vot takuju ogromnuju golovu
tãoFOC assim:F.ACC.SG enorme:F.ACC.SG cabeça.F:ACC.SG
ščuki.
de lúcio.F:GEN.SG
'Oh, você sabe, quando eu era menina, talvez de quatro aninhos, o vovô Sasha, meu pai, trouxe uma cabeça de lúcio tão enorme assim'

Existe uma estreita relação entre a marcação de orações parentéticas e a segmentação da transcrição. Para as orações principais completas que são relacionadas com o conteúdo do restante do segmento e ocorrem no início ou no final desse segmento, os anotadores precisam decidir se elas são parentéticas, constituem segmentos separados ou, às vezes, representam uma oração principal que incorpora o restante do segmento. Além disso, como em russo é possível a queda do argumento (especialmente em conversas informais), os anotadores geralmente precisam decidir se uma frase específica constitui uma oração principal completa ou não. Desenvolvemos algumas heurísticas para essas decisões. Para facilitar a recuperação, se as duas orações podem ser pensadas como os exemplos de uma construção específica, tendemos a optar para a não-divisão em segmentos separados.

Para capturar os fenômenos de interações espontâneas entre as crianças e os cuidadores em nosso *corpus* de maneira mais abrangente, introduzimos novas anotações para cantar (fala cantada), palavras pronunciadas incorretamente (apenas para os desvios mais graves) e palavras ocasionais/neologismos, bem como as anotações de fenômenos bilíngues, empréstimos ocasionais e troca de código. Em BiRCh, uma palavra é marcada como um empréstimo (ocasional) se for adaptada aos sistemas fonológicos e morfológicos da língua russa. Empréstimos morfologicamente adaptados podem ter a marcação de caso ou outra característica morfológica, como em (2), onde *oma* ('vovó' em alemão) aparece com sufixo de caso instrumental russo *-oi*.

(2) opa s om-oj
vovô COM vovó-INS[8]
'vovô e vovó'

Poplack et al. (2020) mostram que os critérios puramente fonológicos não são indicadores confiáveis de empréstimos ou troca de código[9]. Como não estávamos dispostos a considerar todas essas palavras como a troca de código (ou como empréstimos), criamos as regras que permitiam aos anotadores marcar consistentemente esses fenômenos e permitir que os usuários do corpus pudessem encontrar tais exemplos, conduzir as análises fonéticas e potencialmente argumentar que alguns exemplos devem ser reclassificados. Como a variação individual está notoriamente presente, a nossa marcação de empréstimos puramente fonológicos depende do falante: isto é, as palavras precisam ser integradas ao russo, mais do que representar uma troca de código do falante para o inglês ou alemão. Por exemplo, a palavra pafin ('puffin') usada por um dos pais participantes, que, ao contrário da pronúncia em inglês daquele falante, não tem o /p/ aspirado e tem o /f/ palatalizado, é transcrita em letras cirílicas e marcada como um empréstimo. Por outro lado, a frase baby süß ('sweet baby'), pronunciada de acordo com todas as regras fonológicas do alemão, é escrita em alemão e é considerada troca de código. Mesmo com essas heurísticas, muitas vezes é difícil distinguir os empréstimos da troca de código, pois as diferenças na pronúncia podem não ser claras e a palavra pode não mostrar outras marcas de adaptação aos sistemas linguísticos russos (por exemplo, a presença de marcas de declinação para substantivos). Nesses casos, a decisão foi optar pela troca de código.

Em BiRCh, os empréstimos são considerados palavras russas e marcados com todas as informações morfológicas pertinentes, enquanto as trocas de código não são anotadas. Por exemplo, a palavra pafin ('puffin') no exemplo acima é marcada como substantivo, gênero masculino, animado, singular, nominativo. Voltamos à discussão da marcação morfológica na seção a seguir.

[8] Usamos as regras de glossagem de Leipzig com as seguintes opções e modificações: incluímos apenas as características morfológicas imediatamente relevantes para cada exemplo. Além disso, geralmente não marcamos o tempo em verbos finitos nos tempos diferentes do passado – em russo, os verbos nos tempos diferentes do passado têm a marcação de pessoa (por exemplo, *odolžu* PRF:emprestar:1SG), enquanto para os verbos no passado a marcação de pessoa não é usada, mas o gênero fica marcado (em formas singulares) (por exemplo, *odolžila* PRF:emprestar:PST:F.SG). Assim, o leitor possa notar que os verbos finitos que não estejam no modo imperativo e que têm marcação de pessoa, mas não de gênero, não estão no tempo passado.

[9] Somos gratos ao revisor anônimo que nos apontou para a bibliografia referente ao assunto.

4 As diretrizes de anotação morfológica

O russo é uma língua morfologicamente rica com a ordem de palavras flexível, e muitas estruturas sintáticas são expressas por meios morfológicos (por exemplo, casos). Além disso, a morfologia, em particular, tem sido notada como uma área em que os HSs divergem dos padrões monolíngues (POLINSKY, 2018). Portanto, a anotação morfológica completa que vai além da marcação de POS é fundamental para viabilizar as pesquisas baseadas em *corpus* na área do desenvolvimento gramatical de falantes de russo.

Em nossa abordagem em geral, fomos inspirados pelo *Russian National Corpus* (*Corpus* nacional do russo) morfologicamente anotado (RNC, 2003). O ponto de partida para as nossas Diretrizes de Anotação Morfológica (DUBININA et al., 2019) é o conjunto de marcadores de Mystem, parecido com o utilizado em RNC, que facilita as comparações com os dados do BiRCh. Como nesses recursos anteriores, usamos dois tipos de rótulos para a anotação morfológica de cada palavra: um rótulo POS e um conjunto de rótulos de recursos morfológicos (posteriormente, recursos). Em BiRCh, anotamos vários fenômenos não marcados no RNC e, em muitos casos, nos afastamos da análise do RNC para os fenômenos existentes. No restante desta seção, nos concentramos nessas diferenças do RNC e mencionamos os desafios da anotação morfológica para os nossos dados e as soluções propostas.

4.1 Novos fenômenos não marcados no RNC

Diferentemente do RNC, os dados do BiRCh são ricos em fenômenos de linguagem que caracterizam os contextos bilíngues e a aquisição de linguagem por criança, tais como empréstimos, troca de código, palavras ocasionalmente inventadas e as formas morfológicas não padronizadas. Na seção anterior, mencionamos a marcação morfológica de empréstimos ocasionalmente inventados; aqui voltamos para os outros fenômenos.

BiRCh usa um único rótulo para as palavras ocasionalmente inventadas e os neologismos. As palavras ocasionalmente inventadas em sua definição tradicional são palavras inventadas que muitas vezes são o resultado de um jogo de palavras que a

criança faz: por exemplo, *kmiščeta*, criado por uma criança e explicado por ela como ‘uma combinação de cores amarelas e vermelhas’. Essa categoria de anotação também inclui os neologismos criados dentro de uma família, ou seja, aquelas palavras que estão presentes na fala dos pais, como apelidos familiares para as pessoas e objetos. Por exemplo, numa família, os pais e a criança usam consistentemente o neologismo *podguz* em vez de *podguznik* (‘fralda’). Há também palavras inventadas que resultam de pronúncia errada de palavras russas legítimas realizada por crianças, por exemplo, *xamil'jard* (em vez de *xameleon* ‘camaleão’). Neologismos e palavras ocasionalmente inventados são anotados morfologicamente quando reconhecidos como palavras por anotadores NS (*native speaker*, falante nativo). Caso contrário, eles recebem o rótulo POS de não-palavra.

Finalmente, as formas percebidas como erros pelos anotadores NS são marcadas como ‘inesperadas’ e incluem a forma esperada/gramaticalmente correta (3). Isso permite que as pesquisas feitas no *corpus* encontrem os erros morfológicos e, ao mesmo tempo, também permite a possibilidade de as formas marcadas inesperadas serem o resultado de diferenças dialetais entre os participantes e anotadores. Essa possibilidade de aprimoramento da anotação é essencial para a pesquisa sobre a aquisição de formas morfológicas.

(3) Moja (unexpected form, moe) učenie
Minha:FEM (unexpected form, meu.N) aprendizado:N
‘Meu aprendizado’

A outra inovação na anotação BiRCh é a marcação de diminutivos, que não são destacados separadamente no RNC ou em outros *corpora* russos, mas são particularmente interessantes do ponto de vista da aquisição. Anotamos as formas diminutivas usando o recurso DIM e inserindo uma palavra oculta indicando a forma básica não diminutiva. Portanto, a busca feita no *corpus* por um substantivo específico mostrará os exemplos com formas diminutivas e não diminutivas.

Usamos um processo de anotação semelhante para os ideófonos deverbais, ou seja, interjeições etimologicamente relacionadas a verbos que, às vezes, mantêm as propriedades de subcategorização desses verbos, como em (4). Inserimos o verbo relacionado como uma palavra oculta e marcamos o ideófono com um link para a interjeição.

(4) A lisa xvat′ (xvatat′) ego za xvost.

E raposa pega.INTJ (pegar:INF) o pelo rabo
'E a raposa pegou o pelo rabo.'

As duas outras inovações do BiRCh estão relacionadas ao fato de que o corpus fornece as anotações morfológica e sintática. Primeiro, marcamos as informações relevantes morfossintaticamente que não são evidentes da forma morfológica da palavra. Por exemplo, BiRCh tem o rótulo 'quantificacional' para aqueles advérbios que têm a regência de substantivos com o caso genitivo (mnogo 'muitos', čut′-čut′ 'um pouquinho' e vários outros). A segunda inovação diz respeito a uma característica que não será visível no corpus publicado, mas que é importante para a análise sintática de orações por conter o verbo de ligação no tempo presente, pois no idioma russo ele não é realizado nessa situação. Utilizamos o recurso 'predicado' para a palavra (normalmente ocorrida no início) no constituinte remanescente em uma frase verbal que começa com o verbo de ligação nulo; os anotadores fazem os testes da presença do verbo de ligação alternando o enunciado para os tempos passado ou futuro, para verificar se os verbos de ligação byl(a/o) 'foi' ou budet 'será' emergirão. Uma vez que o verbo de ligação nulo é inserido durante a análise, esse recurso não é mais necessário.

4.2 As inovações para os fenômenos descritos no RNC

Nas gramáticas de russo tradicionais (por exemplo, USHAKOV, 1935; OZHEGOV; SHVEDOVA, 1997), as expressões idiomáticas de várias palavras são frequentemente atribuídas a uma única categoria POS, como partícula ou conjunção. Garantimos um POS separado para cada palavra, o que faz com que a busca por palavras específicas produza resultados mais abrangentes e auxilia na análise morfológica e sintática. Também separamos as séries que contêm pronomes indefinidos que consistem de, em inglês, *wh-word* e uma partícula como *-to* ou *-nibud'* em *wh-words* e partículas. Por exemplo, *komu-to* 'someone.DAT' torna-se o pronome no caso dativo *komu* 'who:DAT' (lema *kto*) seguido pela partícula *-to* (lema *-to*). Isso nos permite unificar *wh-words* indefinidas usadas nesse contexto com as *wh-words* ocorridas em qualquer outro contexto, bem como com outros usos de algumas das partículas, por exemplo, o uso da partícula de foco *-to*, como em (5). Palavras separadas que geralmente são escritas juntas na ortografia convencional são marcadas com @.

(5) To -to on byl rad
É que.ND -TO ele.NOM SER:PST.M.SG feliz.M.SG
vstretit' kogo@ @-to!
encontrar:INF quem.ACC -TO
'É que ele estava feliz em encontrar alguém!'

As partículas geralmente formam uma categoria extensa no RNC e nas gramáticas tradicionais, abrangendo muitas palavras que servem a uma variedade de funções, bem como as expressões compostas por várias palavras. Realizamos um levantamento abrangente das partículas no RNC e nos dicionários de Ushakov e Ozhegov com o objetivo de restringir essa categoria de POS àquelas palavras que não podem ser consideradas advérbios, conjunções ou outras POS.

Além de estreitar a categoria POS de partícula, também eliminamos POS 'predicado' atribuído a várias palavras no RNC, Mystem e algumas gramáticas (por exemplo, ZALIZNYAK, 2007), por exemplo, *nužno* 'necessário' ou *izvestno* 'conhecido'. Essa categoria POS não faz parte dos conjuntos de marcadores de PPCHE, AAPCAppE ou os conjuntos de marcadores UD (*Universal Dependencies*, dependências universais) (UD POS, 2014). Marcamos estas ocorrências como advérbios, uma vez que a sua forma morfológica (o sufixo *-o*) está de acordo com essa categoria POS e, além disso, várias dessas palavras apresentam os usos adverbiais.

Um desvio das anotações RNC que queremos mencionar por último é o uso do recurso 'não declinável' (ND) para os pronomes *čto* 'quem', *vsë* 'tudo', *to* 'aquilo' (6) e, por fim, *èto* 'isto'(7), incluindo o seu uso para o preenchimento de pausa (6a, compare com 6b), em certas construções quando seu caso é difícil ou impossível de determinar.

(6) Usač - èto žuk.
Serra-pau isto:ND besouro
'O serra-pau é um besouro.'

(7) a. Daj mne èto ... kružku.
Dá me isso:ND caneca.F:ACC.SG
'Me dá, como que é, a caneca.'
b. Daj mne ètu kružku.
Dá me essa:F.ACC caneca.F:ACC.SG
'Me dá essa caneca.'

4.3 Os desafios da anotação morfológica

O problema mais comum na anotação morfológica é a ambiguidade, ou seja, a marcação de palavras com múltiplas funções morfossintáticas. As nossas diretrizes

incluem uma lista de muitas dessas palavras com as explicações detalhadas. Eliminamos a ambiguidade sempre que possível, e nos casos em que as palavras permanecem ambíguas no contexto, marcamos ambas as possibilidades morfológicas. Isso não apenas nos permite evitar as decisões arbitrárias sobre a anotação, mas também fornece aos usuários do *corpus* as informações sobre a interpretação ambígua dos dados. Destacaremos apenas dois exemplos de desambiguação na anotação morfológica – a marcação de uma palavra como particípio ou adjetivo e a anotação de uma palavra como adjetivo curto ou advérbio.

4.3.1 Particípios vs. Adjetivos

Na gramática tradicional da língua russa, um modificador etimologicamente deverbal é considerado um particípio quando possui complementos e/ou prefixos; caso contrário, é considerado um adjetivo. Em muitos casos, especificamente com sufixos *-n-* e *-en-*, essa escolha afeta a ortografia: os particípios escrevem-se com duplo *nn*, enquanto os adjetivos com um único *n*, apesar de não exibir nenhuma diferença na pronúncia (ver (8a, b)).

(8) a. U nejo vjazanaja jubka.
Ela.GEN tricotar(da):F.NOM.SG saia.F:NOM.SG
'Ela tem a saia tricotada.'
b. Vjazannaja krjučkom
Tricotar:F.NOM.SG agulha de crochê.M:INSTR.SG
ili spicami?
ou agulha de tricô:INSTR.PL
'Feita com a agulha de crochê ou a agulha de tricô?'

Para a consistência na anotação, marcamos tais palavras como particípios (isto é, verbos com o rótulo 'particípio'): em (8), tanto *vjazanaja* quanto *vjazannaja* têm o lema *vjazat* 'tricotar'. Adicionamos 'adjetivo possível' como um rótulo para que os usuários que pesquisam os verbos ou os adjetivos possam encontrar ambas as formas.

4.3.2 Adjetivos vs. Advérbios

As palavras que terminam em *-o* ou em *-e* podem ser adjetivos ou advérbios curtos. Em casos simples, um modificador de frase nominal que tem concordância com o substantivo principal é um adjetivo, enquanto um modificador de frase verbal é um advérbio. Nos casos em que a palavra ocorre numa construção com o verbo de ligação

ou numa construção relacionada a um sujeito neutro, os anotadores verificam a concordância substituindo o sujeito por palavra do outro gênero ou número – feminino ou plural –, como em (9ab).

(9) a. Ej èto interesno.
Ela.DAT isso:N.NOM.SG interessante:ADJ.N.NOM.SG
b. Ej oni interesny.
Ela.DAT eles:NOM interessantes:ADJ.NOM.PL
'Ela está interessada nisso/neles.'

Nas construções em que o teste de substituição não é possível de ser realizado, temos as regras para garantir a anotação consistente. Por exemplo, em um enunciado com um verbo de ligação (por exemplo, 'ser' ou 'tornar-se') e sem um sujeito evidente, a palavra alvo é sempre marcada como um advérbio. Por último, quando não está claro como classificar a construção, tendemos a optar para o advérbio, como no exemplo abaixo.

(10) Ej interesno, xorošo, veselo (v škole).
Ela.DAT interessante:ADV bem:ADV divertido (na escola)
'Ela está interessada, bem e divertida (na escola).'

Depois da revisão e da correção manual dos marcadores morfológicos automáticos, os dados passam para a etapa final da anotação: análise sintática.

5 Análise Sintática

Fazer anotação sintática é uma das decisões mais trabalhosas e caras que os criadores de um *corpus* podem tomar. A estrutura sintática, em contraste com a forma morfológica, muitas vezes é invisível (exceto para as pistas prosódicas ocasionais) e deve ser inferida. As possibilidades da sintaxe de qualquer linguagem humana são infinitas (modo de performance) e oferecem uma gama notoriamente ampla de ambiguidades. Nesta seção, começamos mostrando que esses dois obstáculos – a invisibilidade e a infinidade – aplicados aos dados bilíngues tornam a decisão sobre a anotação sintática uma tarefa fácil. Em seguida, revisamos os principais aspectos de construção e organização da nossa anotação sintática e mostramos como juntos eles facilitam a busca das respostas para as questões de cunho teórico.

5.1 A motivação para a anotação sintática

A gramática da língua de herança é semelhante a uma tecelagem complexa que une várias condições e processos. BiRCh mostra como essa tecelagem se desdobra no tempo e ajuda a desvendar sua trajetória de aquisição não trivial. Conforme descrito na introdução, para 'desfiar o arco iris', pelo menos quatro processos devem ser desembaraçados: a transferência linguística (o empréstimo de propriedades gramaticais da língua dominante), o atrito linguístico (modificação da gramática da L1), a obtenção divergente (a produção de recursos novos ou diferentes com a base no *input* incompleto) e o *input* produzido pelos pais que é diferente quando se trata das crianças monolíngues e bilíngues. No caso do *input* diferente, os papéis da transferência linguística e da inovação independente devem ser distinguidos na fala dos pais. É importante ressaltar que todos esses processos envolvem alguma forma de desalinhamento dentro dos pares de forma-significado que é significativamente mais comum entre as estruturas sintáticas invisíveis e infinitas do que entre as unidades morfofonológicas diretamente evidentes e finitas. A anotação sintática é o melhor lugar para examinar esses processos, o que torna necessário que o nosso *corpus* atenda às necessidades dos pesquisadores que estudam a aquisição de línguas, as línguas de herança, o contato linguístico e a sintaxe e semântica teóricas.

5.2 A arquitetura da anotação sintática

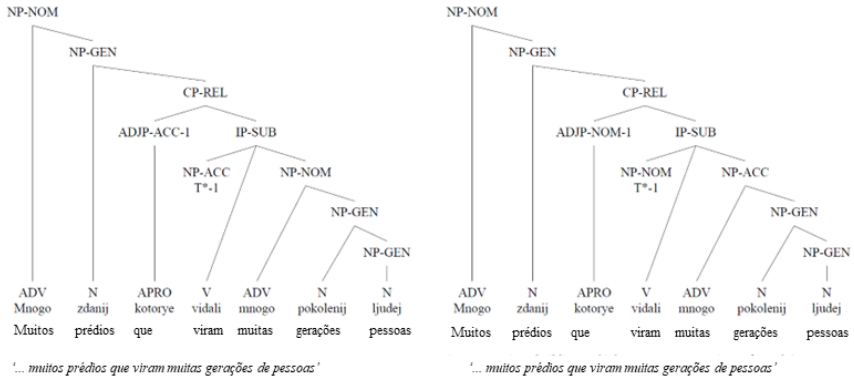
A anotação sintática busca um equilíbrio harmonioso entre os recursos e as ambições dos criadores do *corpus* e as necessidades dos possíveis usuários futuros (conforme interpretado pelos criadores do *corpus*). Um dos principais objetivos para nós era tornar o BiRCh acessível a um amplo grupo de profissionais de linguagem que podem se beneficiar de informações sobre a sintaxe. Para este fim, duas decisões prudentes que moldaram toda a anotação sintática foram tomadas desde o início.

Primeiramente, adotamos o estilo de anotação sintática *Penn Treebank* (MARCUS; SANTORINI; MARCINKIEWICZ, 1993). O objetivo principal desse estilo de anotação é facilitar a pesquisa sintática automatizada e, em busca desse objetivo, a precisão teórica às vezes pode ser desprendida em favor da simplicidade da

anotação. Especificamente, a anotação sintática *Penn Treebank* permite a ramificação *n*-ária e estruturas excêntricas (ou seja, sem um ponto central). Por outro lado, esse estilo de anotação é familiar a muitos linguistas que usaram outros *corpora* analisados que são criados seguindo os mesmos princípios (MARTINEAU, 2008; WALLENBERG et al., 2011; BECK, 2013; KROCH et al., 2016; GALVES; ANDRADE; FARIA, 2017; TORTORA et al., 2017; KROCH, 2020). Também é importante que o formato *Penn Treebank* inclua uma linguagem de consulta do CS abrangente que é usada para fazer a busca pelo *corpus* e modificá-lo. Como resultado, a nossa anotação sintática pode parecer não sensível para algumas questões difíceis da sintaxe russa (veja abaixo), porém procura fornecer os meios para que todos os pesquisadores possam facilmente encontrar os dados que procuram.

Em segundo lugar, na tentativa de abordar algumas das questões mais complicadas da sintaxe de línguas de herança, optamos por concentrar os nossos esforços em duas propriedades da linguagem: a ambiguidade e o silêncio. Em termos da primeira questão, os HSs muitas vezes precisam investir muitos esforços para lidar com as orações ambíguas (ver POLINSKY; SCONTRAS, 2020 e referências nesses trabalhos). Para os HSs de russo, tais dificuldades se manifestam em múltiplos níveis, desde sinônimos lexicais (RAKHILINA; VYRENKOVA; POLINSKY, 2016) até a solução de uso da anáfora (IVANOVA-SULLIVAN, 2014a), a ordem das palavras e o escopo de quantificadores (IONIN; LUCHKINA, 2019). No entanto, a ideia que os HSs tendem a evitar a ambiguidade em todas as situações é somente uma previsão. Para abordar essa questão, a anotação sintática em BiRCh inclui sistematicamente as informações sobre a ambiguidade sintática. Em BiRCh, cada segmento pode ser associado a várias estruturas sintáticas. Dessa forma, qualquer ambiguidade, desde que seja detectável com a quantidade de rótulos oferecidos em nossa anotação, é incluída e pode ser encontrada (como em 11).

(11)



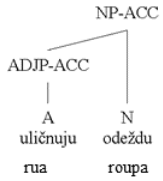
Voltando à questão do silêncio, é um outro tema recorrente na linguística de HL (*heritage language*, língua de herança) línguas de herança que diz respeito ao significado associado à ausência de elementos abertos. Suas interpretações muitas vezes levam os pesquisadores a conclusões paradoxais. Por exemplo, as gramáticas de línguas de herança são frequentemente citadas a respeito do atrito quando se trata de pronomes nulos (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). O efeito é tão prevalente e forte que os falantes de uma língua *pro-drop* produzem os pronomes abertos em sua língua de herança com uma frequência significativamente mais alta em comparação com os falantes do grupo controle, mesmo quando sua língua dominante também é *pro-drop* (ver DE PRADA PÉREZ, 2009, 2015 para os dados em espanhol e catalão). Ao mesmo tempo, há discussões sobre a substituição de alguns tipos de elipses pelo *pro-drop*. Polinsky (2016, 2018) afirma que os HSs de russo reanalisam um tipo específico de elipse chamado de VP (*verb phrase*, sintagma verbal) como queda do objeto uma vez que ambos parecem levar a formas de superfície idênticas (GOLDBERG, 2005; GRIBANOVA, 2013). Em outras palavras, os HSs parecem desfavorecer fortemente o *pro-drop*, exceto quando reinterpretam algum tipo de elipse como *pro-drop*. A combinação dessas tendências, por sua vez, sugere que o *pro-drop* de objeto possa ser usado mais do que outras quedas de argumento. A nossa anotação sintática em que o fenômeno *pro-drop* é marcado para todos os argumentos obrigatórios é bem apropriada para verificar se as reivindicações desse tipo são sustentadas.

5.3 A construção da anotação sintática

Passando das decisões sobre a organização para a construção do *corpus*, a anotação sintática em nosso *corpus* foi criada usando a linguagem de consulta CS que foi projetada para trabalhar com os *corpora* anotados sintaticamente no formato *Penn Treebank*. Para ser compatível com essa linguagem de consulta, os arquivos resultantes da anotação morfológica são primeiramente normalizados e convertidos do formato FoLiA para o formato *Penn Treebank*, de modo que, para cada segmento, fenômenos de discurso que complicam a análise sintática (por exemplo, elaborações intra-sentenciais e orações parentéticas) são colocados em colchetes separados e, portanto, podem ser analisados sintaticamente após o conteúdo principal ter sido totalmente analisado. O processo de análise do conteúdo principal de cada segmento (que posteriormente também é aplicado às elaborações e aos elementos parentéticos colocados em colchetes) foi separado em três fases que visam diferentes grupos de constituintes e fenômenos gramaticais. Cada fase consiste em duas etapas. Primeiro, a análise sintática semiautomática baseada em regras prossegue com as consultas de revisão feitas no *corpus*. Na segunda etapa, a análise é corrigida manualmente mediante consultas realizadas no *corpus*; essas identificam as classes específicas de exemplos que posteriormente são modificadas em editores de texto que trabalham com a notação de colchetes no estilo *Penn Treebank*. A divisão em três fases reflete a ideia básica de desenvolvimento da árvore sintática de baixo para cima. Nesse caso, cada fase usa as informações sintáticas coletadas e consolidadas na fase anterior.

Durante a primeira fase, as informações morfológicas (POS e os marcadores de caso) são usadas para identificar e projetar constituintes endocêntricos ‘pequenos’: NP (*noun phrase*, sintagma nominal), ADJP (*adjective phrase*, sintagma adjetival), NumP (*number phrase*, sintagma numérico), PP (*prepositional phrase*, sintagma preposicional), etc. Nessa etapa, os constituintes também são marcados com rótulos indicando o caso. As informações sobre os casos são usadas para identificar e incorporar os subconstituintes de NPs para diagnosticar a subextração e reconstruir os traços internos de NP (consulte 12 e 13 abaixo).

12



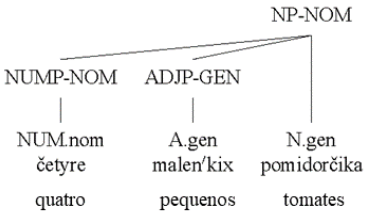
'roupa de rua'

13

'Que xícara você quer?'

Para simplificar, todos os (sub)constituintes nominais são incorporados em NP, como no exemplo abaixo:

(14)

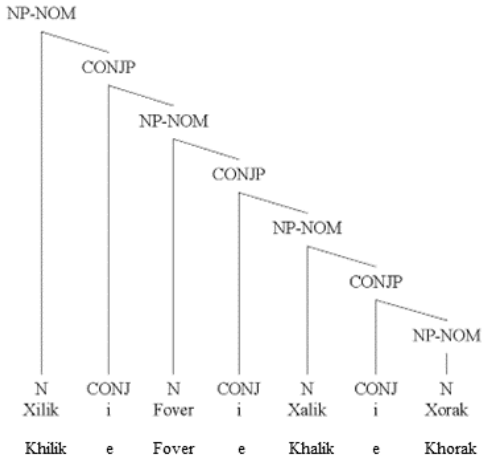


'quatro tomates pequenos'

Essa estrutura simplificada de NP também significa que os rótulos de casos em alguns contextos estruturais de uso de casos precisam ser modificados. Em (14) acima, por exemplo, o caso do NP em uma construção numérica com o genitivo de quantificação precisa corresponder ao caso de Num e não de N.

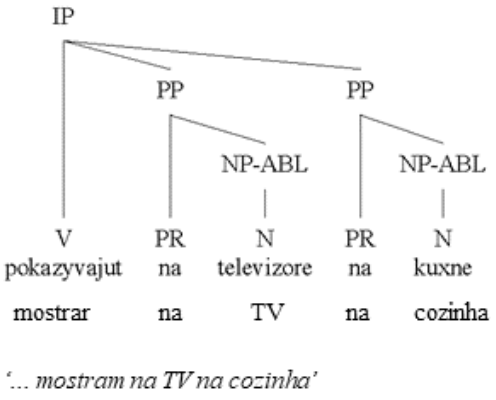
Após a identificação dos NPs, também marcamos as conjunções de NPs (15). Para facilitar a exposição, todo conjunto subsequente é incluído no conjunto anterior. A revisão manual durante a primeira fase consiste na reatribuição de PPs pós-nominais que podem ser parte de NP ou de um argumento/adjunto de oração, como em (16).

(15)



'Khilik e Fover e Khalik e Khorak'

(16)

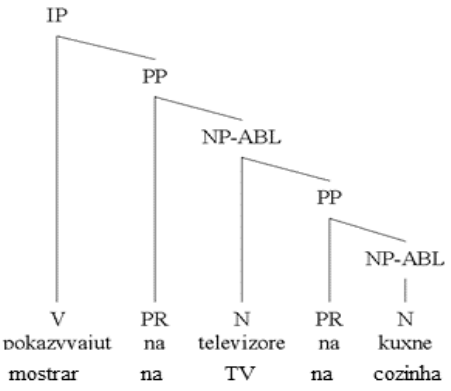


'... mostram na TV na cozinha'

Quando ambas as posições alternativas são plausíveis, duas árvores são geradas e associadas ao segmento, como em (17).

A segunda fase consolida as informações sintáticas que foram coletadas durante a primeira fase para localizar os constituintes excêntricos 'maiores' – IP (*inflectional phrase*, sintagma flexional) e CP (*complementizer phrase*, sintagma de complemento) – identificar as orações com os verbos de ligação e reconstruir os traços de movimento interno e externo definido em relação à oração, como em (18).

(17)



'... mostram na TV na cozinha'

(18)

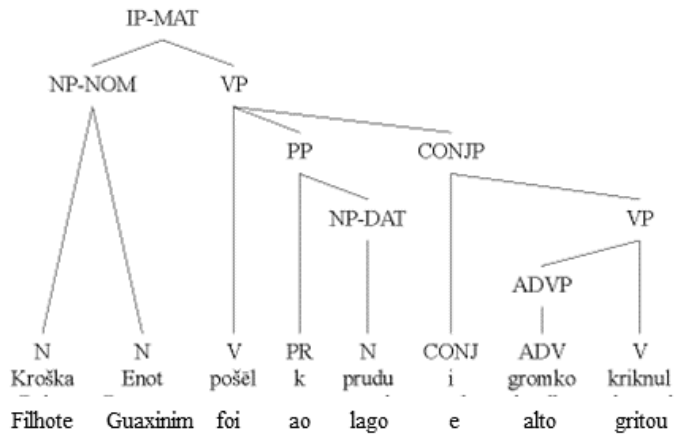
'O que você está fazendo?'

Decidimos não incluir VPs na anotação sintática regular. A localização do VP envolve uma infinidade de compromissos teóricos (por exemplo, incluímos também AspP, vP ou Voice? VP *shells*? Reconstruímos todos os NPs sujeitos dentro de VP/vP

ou apenas alguns deles? Mantemos a mesma arquitetura funcional para as orações não ergativas e não acusativas? Como analisamos os predicados de estados psicológicos? Onde anexamos os advérbios específicos?). No final, todas essas opções apenas complicam a pesquisa e inundam (e, eventualmente, afogam) o pesquisador com os meandros da sintaxe teórica russa. As duas situações nas quais marcamos explicitamente VP é quando ele é destacado por um processo sintático (o movimento para o início da oração, conjunção, etc.), como em (19), e quando uma elipse é possível.

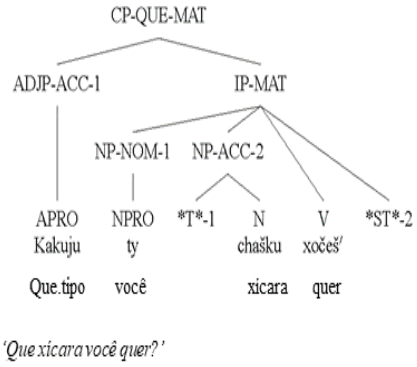
Durante a parte automatizada da segunda fase, verificamos os traços do movimento dos constituintes para a esquerda que foram identificados durante a primeira fase. Assumimos que o russo é uma língua SVO e, portanto, todas as outras permutações são criadas por *scrambling*, como em (20).

(19)



'O Filhote Guaxinim foi ao lago e gritou alto'

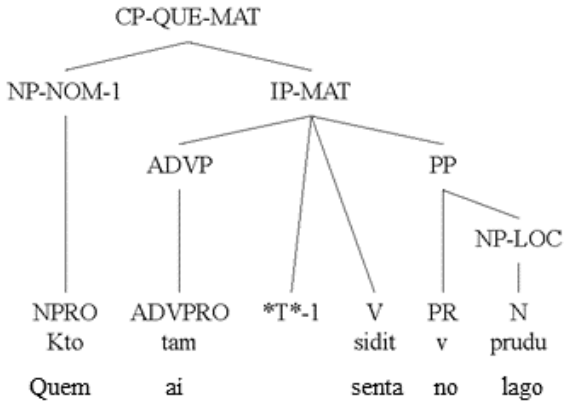
(20)



'Que xícara você quer?'

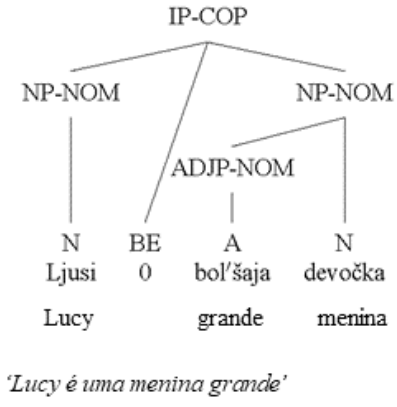
Também sustentamos que qualquer movimento para a esquerda em *left-adjoined* advérbios é um resultado do movimento para o início da sentença. Dependendo se o elemento movido é uma *wh-frase* ou um elemento em foco, distinguimos os movimentos para o início em *wh-fronting* e *focus fronting*. Os traços de ambos são reconstruídos durante a segunda etapa (21). Durante a segunda fase também reconstruímos os verbos de ligação nulos em orações que contêm essas ligações, como em (22).

(21)



'Quem é que está sentado aí no lago?'

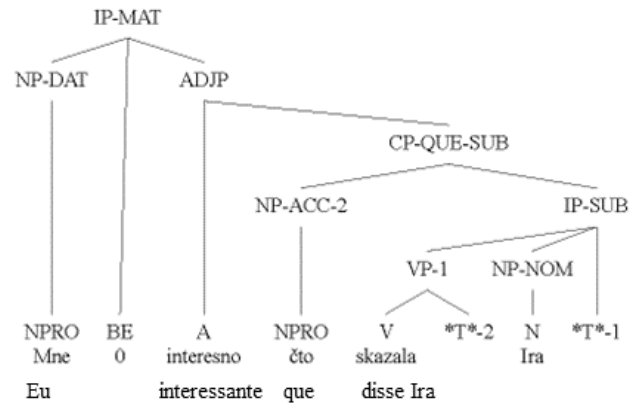
(22)



'Lucy é uma menina grande'

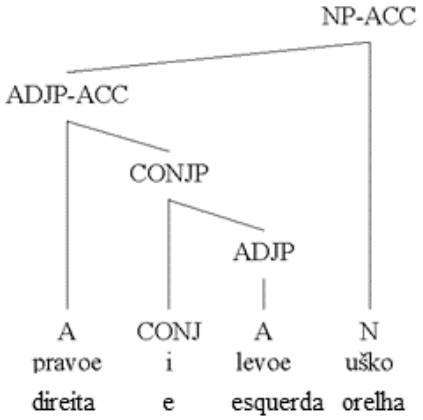
Por fim, projetamos os CPs para cima de IPs em *wh-questions* de matriz e CPs subordinados declarativos e interrogativos. Os dois últimos são ainda incluídos nos constituintes apropriados. Neste ponto, a anotação sintática contém as informações suficientes para marcar todos os diferentes subtipos de CPs e IPs (matriz, interrogativa subordinada, etc.), como em (23). A parte manual da segunda fase inclui a revisão da junção, uma verificação visual da atribuição dos traços e uma avaliação dos subtipos do IP e CP (veja o exemplo 24).

(23)



'Estou curioso em saber o que a Ira disse'

(24)

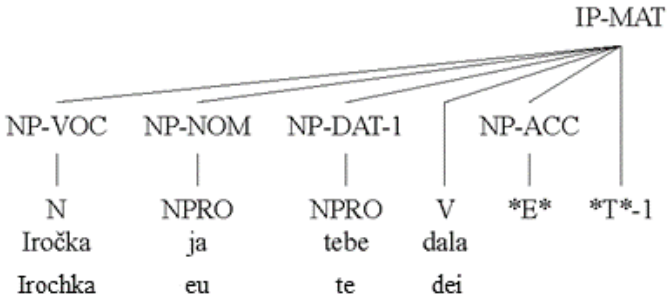


'orelhinha direita e esquerda'

Finalmente, a terceira fase visa preencher as 'lacunas' na estrutura de oração que geralmente estão associadas a elipses e respostas fragmentadas. Durante esta etapa, cada predicado e seus argumentos (identificados com a aplicação de rótulos de casos) são verificados em nosso dicionário interno de valências verbais. No caso de qualquer

argumento ausente, o argumento nulo correspondente é criado e adicionado no local apropriado, como em (25).

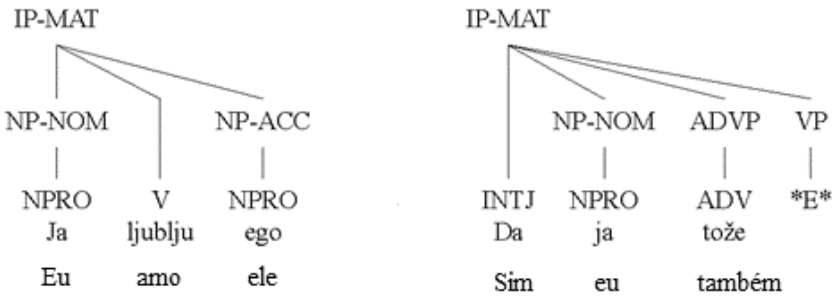
(25)



'Irochka, eu te dei'

A parte manual da terceira fase inclui a correção e a expansão das elipses não nominais para incluir os casos de elipses de VP, bem como as faltas e evasão, como em (26).

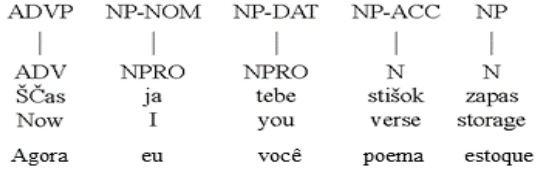
(26)



'Eu amo ele' *'Sim, eu também'*

Após a terceira fase, uma verificação manual final é realizada nas correções no *corpus* todo para excluir a possibilidade de que a resolução incorreta da ambiguidade morfológica (como, por exemplo, classificação de *zapas* como substantivo enquanto é um verbo, como em (27)) resulte em uma série de erros que se extendem até a árvore sintática.

(27)



'Agora eu recito um poema para você'
(a intenção) 'Agora tenho guardado um poema para você'

As conclusões e implicações para a pesquisa

O que torna o BiRCh singular é a sua camada de anotação final, ou seja, a análise sintática, já que é o primeiro e único banco de árvores sintáticas em russo realizado em grande escala no momento. Em termos de especificação de anotação, é diretamente comparável ao *subcorpus* de um milhão de palavras do *Wall Street Journal* do *Penn Treebank* (MARCUS et al., 1999), o conjunto de dados mais usado para a análise de constituintes sintáticos em inglês (CLARK; FOX; LAPPIN, 2013, p.241) e, portanto, pode desempenhar um papel semelhante para a análise de constituintes sintáticos em russo e contribuir para a análise sintática multilíngue, incluindo os idiomas morfologicamente ricos (SEDDAH et al. , 2013; SEDDAH; K3; TSARFATY, 2014). Além do mais, podemos criar um recurso exclusivo para a análise de constituintes sintáticos em modalidades escrita e falada, convertendo SynTagRus (BOGUSLAVSKY et al., 2002), um banco de árvores sintáticas de dependência de larga escala baseado em textos escritos em russo [10], para um banco de árvores de constituintes sintáticos, e combinando o resultado com os dados de fala de BiRCh. A conversão de SynTagRus é possível com base nas nossas Diretrizes de Anotação Morfológica e Sintática e a metodologia proposta por Luu e coautores (2016).

Além disso, o nosso *corpus* pode ser usado como um valioso conjunto de dados dentro do padrão de excelência para várias tarefas de NLP (*Natural Language Processing*) correspondentes aos seus múltiplos aspectos de anotação. Por exemplo, o reconhecimento de fala conversacional se beneficiará da transcrição de um milhão de palavras do BiRCh alinhada com aproximadamente 270 horas de áudio de alta qualidade no nível de frases (consulte JURAFSKY; MARTIN, 2020 para uma revisão de conjuntos de dados semelhantes em inglês, como SwitchBoard e CALLHOME).

Da perspectiva da linguística teórica, o *corpus* BiRCh anotado para as disfluências e morfossintaxe oferece acesso sem precedentes ao estudo da gramática russa. Por exemplo, recuperar todos os exemplos de elipses de VP, construções de negação, vários tipos de construções passivas ou orações principais com sujeitos nulos torna-se uma questão de pesquisa simples, e padrões observáveis nos dados podem ser comparados com predições teóricas. Além disso, como o *corpus* permite as

[10] Acessível ao público em https://universaldependencies.org/treebanks/ru_syntagrus/index.html.

comparações confiáveis entre os adultos monolíngues e bilíngues e seus filhos, ele fornece a base para as previsões teóricas mais fortes. Abaixo, fornecemos uma breve descrição de alguns projetos de pesquisa em andamento ou nas etapas de planejamento que se baseiam nos dados do BiRCh.

Assim que os dados se tornaram disponíveis nas etapas iniciais da construção do *corpus*, nós (membros da equipe BiRCh) usamos as transcrições anotadas da disfluência alinhadas ao áudio para estudar as propriedades de duas expressões russas – *aa* e *mm* (DUBININA et al., 2018). Essas são geralmente vistas como os preenchedores de pausa, ou seja, hesitações não silenciosas semelhantes aos *uh* e *um* do inglês, mas, como descobrimos, também podem sinalizar o comprometimento, recebimento da mensagem ou chamar atenção, como em (28):

(28) Mm! Ty golodnaja, aa? - Mm?
Mm você.NOM faminta:NOM, neh? - Mm?
- Est' xočeš'? - Aa, da.
- Comer.INF quer:2SG? - Ah, sim.
'Oh! Você esta com fome, neh? - Oi? - Quer comer? - Ah, sim.'

Descobrimos que *aa* e *mm* na fala dos pais apresentam padrões de distribuição distintos, e que há diferenças significativas entre os pais monolíngues e bilíngues no uso dessas palavras como preenchedores de pausa, mas não em suas outras funções (DUBININA et al., 2018), o que sugere o efeito do bilinguismo. Atualmente, estamos explorando correlações entre as palavras *aa* e *mm* na fala dos pais e das crianças.

Dois outros estudos atuais baseados em dados do *corpus* BiRCh fundamentam-se em anotação morfológica e sintática e visam abordar as questões teóricas e de aquisição de língua mais amplas: o marcador de polidez lexical *požalujsta* 'please' e as solicitações de forma mais geral, as construções com os verbos marcados com o sufixo *-sia*. A distribuição de vários usos de *-sia*, que podem ter significados passivos, médios, reflexivos, recíprocos e outros, na fala de pais monolíngues lançará luz sobre as questões teóricas da sintaxe e da semântica do russo, enquanto a sua distribuição no *input* e no *output* (produzido por crianças bilíngues) pode responder a perguntas sobre o desenvolvimento das interfaces sintaxe-semântica e sintaxe-pragmática em situações de contato linguístico (MALAMUD et al., 2022). Da mesma forma, o estudo que investigou o uso de *požalujsta* 'please' (DUBININA et al., em andamento) pode elucidar o desenvolvimento de estratégias de polidez em comunidades bilíngues que levam a gramáticas de línguas de herança divergentes (DUBININA; MALAMUD,

2017) e, ao mesmo tempo, avançar a nossa compreensão da gramática da modificação de atos de fala.

Para dar um exemplo concreto de pesquisa sintática possibilitada por BiRCH, podemos olhar para a aquisição infantil de *Left Branch Extraction* (LBE, Extração da posição esquerda ao núcleo) (ROSS, 1967), um tipo de subextração do NP. LBE é possível em russo, mas não está presente em inglês e alemão (as duas línguas dominantes no grupo bilíngue em BiRCh). Van Kampen (1994) discute um caso peculiar de crianças holandesas (L1) produzindo frases com LBE, embora os falantes adultos de holandês não apresentem LBE por completo. Ela liga a LBE à presença de morfologia atributiva e levanta a hipótese de que as condições restritivas de morfologia pobre são adquiridas lentamente, o que deixa às crianças holandesas uma janela de tempo para brincar com LBE. BiRCh fornece os meios perfeitos para testar ainda mais a hipótese de Van Kampen. Como o russo é uma língua morfologicamente rica, não esperamos encontrar as limitações para LBE, a menos que elas sejam precedidas de empobrecimento do inventário morfológico comum para a aquisição bilíngue. Relatamos nosso estudo em Koval et al. (2022).

Para finalizar, esperamos que, ao descrever a metodologia utilizada para o *corpus* BiRCh, mostremos uma gama completa de seu potencial para a pesquisa e para a criação de outros *corpora* orais de fala bilíngue. Os *corpora* de fala infantil sintaticamente anotados em geral, e o *corpus* BiRCh bilíngue coletado em condições naturalísticas em particular, fornecem uma ferramenta importante para a pesquisa sobre a aquisição de sintaxe, morfologia e suas interfaces com semântica e pragmática. Por fim, essa pesquisa pode lançar luz sobre a natureza da aquisição da linguagem em si, além de fornecer o conhecimento sobre a mudança linguística em crianças e adultos em situações de contato linguístico, e sobre a linha de base monolíngue.

REFERÊNCIAS

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.

LƯU, 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.

Declaração de Contribuição dos Autores

Todos os autores (Alex Luu, Pasha Koval, Sophia A. Malamud e Irina Y. Dubinina) fizeram contribuições substanciais para a elaboração do artigo "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", abrangendo integralmente os seguintes aspectos: 1) concepção, análise e interpretação dos dados; 2) redação e revisão crítica do artigo quanto ao conteúdo intelectual importante; 3) aprovação final da versão a ser publicada, 4) responsabilização por todos os aspectos do trabalho, garantindo que as questões relacionadas à precisão ou integridade de qualquer parte do trabalho sejam devidamente investigadas e resolvidas. Enquanto todos os autores trabalharam em todo o manuscrito, as seguintes seções receberam destaque especial dos seguintes autores: seção 1 – Alex Lưu, seção 2 – Irina Y. Dubinina, seção 3 – Sophia A. Malamud, seção 4 – Irina Y. Dubinina e Sophia A. Malamud, seção 5 – Pasha Koval, seção conclusiva – Irina Y. Dubinina, Sophia A. Malamud e Alex Lưu.

Traduzido por *Aleksandra S. Skorobogatova* – as.skorobogatova@gmail.com

Recebido em 01/10/2021

Aprovado em 29/08/2022

Pareceres

Tendo em vista o compromisso assumido pela *Bakhtinina*. Revista de Estudos do Discurso com a Ciência Aberta, a revista publica somente os pareceres autorizados por todas as partes envolvidas.

Disponibilidade de dados de pesquisa e outros materiais

Os conteúdos subjacentes ao texto da pesquisa estão contidos no manuscrito.

Anexo

Abreviaturas empregadas no texto

Abreviatura	Inglês	Português
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 nativo
POS	part-of-speech	classe gramatical

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 adjetival
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