D.O.I. http://dx.doi.org/10.1590/2176-4573p55747

ARTIGOS

Pesquisa com *corpora* de aprendizes de russo: estado da arte e apelo à ação / *Russian Learner* Corpora *Research: State of the Art and Call for Action*

Olesya Kisselev[*]

RESUMO
Com o aumento da disponibilidade e facilidade de uso de *corpora* de língua russa e ferramentas de análise de *corpus*, o campo do ensino da língua russa começou recentemente a empregar a linguística de *corpus* como uma abordagem para entender a dinâmica de desenvolvimento de russo como segunda língua e língua de herança. O artigo fornece uma breve visão geral do estado atual da pesquisa na área de *corpora* de aprendizes e explora os benefícios da aplicação de métodos e instrumentos de linguística de *corpus* para o estudo do russo. O artigo revisa questões pertinentes na área de *design*, compilação e anotação de *corpora*; oferece uma visão geral dos *corpora* de língua russa existentes e descreve os estudos de russo como a segunda língua/língua de herança baseados em análise de *corpus*, atualmente disponíveis. O artigo conclui com um chamado aos especialistas na área para explorar o benefício de abordagens baseadas em *corpus* para o estudo do russo.
PALAVRAS-CHAVE: Linguística de *corpus*; Pesquisa de *corpus* de aprendizes; *Corpora* de língua russa; Aquisição de segunda língua; Aquisição de língua de herança

ABSTRACT
With the increase in availability and user-friendliness of Russian language corpora *and* corpus-*analytic tools, the field of Russian language education has recently begun to employ* corpus *linguistics as an approach to understanding the dynamic of language development in users of Russian as a second and heritage language. The paper provides a brief overview of the current state of learner* corpus *research as a field and explores the benefits of application of* corpus *linguistics methods and instruments to the study of Russian. The paper reviews pertinent issues in* corpora *design, compilation, and annotation; offers an overview of the existing Russian language* corpora *and reports on the currently available* corpus-*based studies of Russian as a second/heritage language. The paper concludes with a call to the field to explore the benefit of* corpus-*based approaches to the study of Russian.*
KEYWORDS: Corpus *linguistics; Learner* corpus *research*; Corpus-*based research; Russian language* corpora*; Second language acquisition; Heritage language acquisition*

[*] University of Texas at San Antonio, Department of Bicultural Bilingual Studies, San Antonio, Texas, Estados Unidos da América; https://orcid.org/0000-0003-2514-3107; olesya.kisselev@utsa.edu

Introdução

O avanço generalizado da tecnologia computacional que ganhou velocidade na década de 1990 resultou em mudanças significativas em muitas disciplinas sociais, incluindo a linguística e os estudos da linguagem aplicada, que viram o crescimento da nova disciplina de linguística de *corpus* que foca principalmente na exploração de dados (mais do que na verificação de teorias) que são contidos em grandes e organizados bancos de dados, conhecidos como *corpora* de linguagem. Descrita tanto como uma metodologia quanto como um método (GRIES, 2009; MCENERY; HARDY, 2012), uma prática e uma "abordagem filosófica" (LEECH, 1992), a linguística de *corpus* utiliza os métodos e ferramentas de análise da linguagem assistida por computador que permitem aos pesquisadores analisar grandes quantidades de dados linguísticos autênticos para procurar padrões, regularidades e idiossincrasias na estrutura da linguagem e no uso da linguagem em várias modalidades, variedades, registros, gêneros e por grupos diferentes de falantes. O impacto da linguística de *corpus* no campo dos estudos da linguagem tem sido significativo e é descrito por muitos linguistas como nada menos que revolucionário (HUNSTON, 2002; KOPTEV; MUSTAJOKI, 2008; GRIES, 2011, entre outros), contribuindo para todas as subáreas da linguística.

A área da pedagogia da linguagem tem sido, sem dúvida, uma das que mais contribuíram no desenvolvimento das abordagens da linguística de *corpus*. Resumidamente, a convergência entre os campos de ensino de línguas e da linguística de *corpus* originou duas grandes direções (LEECH, 2014). Uma foca na aplicação do conhecimento obtido nas investigações de *corpora* padrão para melhor atender às necessidades pedagógicas de professores e alunos de línguas. Essa abordagem, por exemplo, produziu uma série de materiais modernos, desenvolvidos baseado em evidências: gramáticas de referência, dicionários de frequência, listas de frases, livros didáticos e outros materiais de ensino/aprendizagem baseados em dados de *corpus* (CONRAD; BIBER, 2009; BIBER; CONRAD, 2010; KOPOTEV; MUSTAJOKI, 2008; LU *et al*., 2018; LEBEDEVA, 2020). Além disso, os professores de línguas vêm desenvolvendo métodos e técnicas pedagógicas para aprendizagem orientada por dados, uma abordagem que permite a exploração

independente e semi-independente de dados de *corpus* por aprendizes de línguas (BOULTON, 2017).

O outro lócus da convergência está na aplicação de métodos e ferramentas da linguística de *corpus* para o estudo da língua do aprendiz, ou seja, a língua produzida por aprendizes em diferentes níveis de proficiência linguística, com vistas a uma melhor compreensão das trajetórias de desenvolvimento dos comportamentos linguísticos, lacunas e habilidades de quem aprende uma língua como segunda língua (L2), língua estrangeira (FL, *foreign language*) ou língua de herança (HL, *heritage language*) (GRANGER, 2009; LEECH, 2014).

Ambas as direções se desenvolveram de forma robusta ao longo das últimas três décadas. É certo que o maior progresso foi feito na área de inglês como segunda língua (ESL, *English as a Second Language*) / língua estrangeira (EFL, *English as a Foreign Language*), em que a disponibilidade tanto de *corpora* padrão quanto de *corpora* de aprendizes bem desenvolvidos e a adoção de métodos de linguística de *corpus* vieram cedo e foram apoiadas pela introdução de várias práticas institucionalizadas. Nos últimos anos, no entanto, houve alguns desenvolvimentos encorajadores na linguística de *corpus* de russo, tanto no que diz respeito aos *corpora* padrão quanto à linguística de *corpus* de aprendizes (KISSELEV; FURNISS, 2020; LEBEDEVA, 2020).

No presente artigo, a autora fornece uma revisão de alguns desses desenvolvimentos, especificamente na área de abordagens baseadas em *corpus* para o estudo da língua russa produzida por aprendizes, e defende a necessidade de maiores avanços para chegar em uma verdadeira convergência entre a linguística de *corpus* de russo e os estudos de aquisição de russo como a segunda língua (SLA, *Second Language Acquisition*).

1 Avanços nos estudos de russo baseado em *corpus* produzido por aprendizes

Desde o início da década de 1990, os linguistas de *corpus* têm defendido o valor dos *corpora* de aprendizes no ensino de línguas. *Corpora* de aprendizes representam uma língua produzida por falantes cujo domínio da língua ainda não atingiu a maturidade (LEECH,

2014); isso inclui *corpora* de aquisição da primeira língua/língua infantil (L1), *corpora* de aprendizes de segunda língua, coletadas em falantes de L2 ou FL que estão em diferentes níveis de proficiência e, ultimamente, *corpora* de língua de herança, que incluem os dados linguísticos obtidos de falantes de HL e/ou aprendizes de HL. O principal objetivo dos *corpora* de aprendizes é "contribuir para uma melhor compreensão dos padrões universais, específicos tanto a uma língua quanto a um grupo, da aquisição de segunda língua/ uma língua estrangeira" (KISSELEV, 2021, p.525). Como tal, os *corpora* de aprendizes são ferramentas úteis tanto para o estudo teórico da aquisição da linguagem quanto para os objetivos aplicados de criar melhores currículos, programas de estudo e materiais pedagógicos para os aprendizes de línguas.

Os estudos de SLA de russo também contribuíram para o desenvolvimento e investigação de *corpora* de aprendizes. O primeiro *corpus* abertamente disponível, o *Russian Learner Corpus of Academic Writing* (RULEC, ou *Corpus* de Escrita Acadêmica de Aprendizes de Russo), foi feito há mais de uma década. Sendo um *corpus* longitudinal de redações de nível avançado, ele contém textos escritos (tarefas de casa, ensaios e artigos científicos) criados por estudantes de língua russa que estavam todos matriculados na mesma sequência de cursos de língua russa de nível avançado em uma universidade americana. A característica única do RULEC é a distribuição de dados equilibrada em relação aos *backgrounds* dos aprendizes, sendo 19 dos 36 autores do *corpus* aprendizes de russo como língua de herança, e o restante, aprendizes de russo como língua estrangeira[1]. Essa característica exclusiva permite uma comparação sistemática de padrões de desenvolvimento da linguagem em aprendizes de russo como língua de herança e como língua estrangeira. O *corpus* também fornece outros tipos importantes de metadados, ou seja, informações sobre os textos e os aprendizes que os criam, tais como o nível de proficiência linguística (na escala de proficiência ACTFL), o nome do curso para o qual o artigo foi escrito, o tipo de texto (por exemplo, parágrafo, ensaio, artigo científico), a função para a qual a tarefa foi formulada (por exemplo, definição, narração, argumentação, etc.) e restrições de tempo (escrita cronometrada ou não cronometrada). Esses metadados ajudam os pesquisadores a criar

[1] Para uma descrição mais detalhada do *design* de RULEC, procedimentos de compilação e objetivo, bem como ideias para o uso pedagógico do *corpus*, consulte Alsufieva e colegas (2012) e Kisselev e Alsufieva (2017).

subcorpora com base nas características do aprendiz e do texto e comparar esses *subcorpora* em relação aos vários parâmetros linguísticos com o objetivo de entender os efeitos relativos do nível de proficiência, gênero, tópico e outras características dos aprendizes e dos textos que eles criam nas características linguísticas desses textos.

Os dados originais do RULEC são brutos, ou seja, os dados linguísticos não são lematizados, não possuem *tags* de classes gramaticais ou anotações sintáticas. Embora todos esses procedimentos tenham se tornado facilmente disponíveis (KISSELEV, 2021), os primeiros estudos baseados em dados do RULEC utilizaram os dados brutos. De fato, certas questões de pesquisa podem ser investigadas com sucesso usando apenas os dados não analisados, com a ajuda de procedimentos de análise de *corpus* apropriados. Tal foi a abordagem de Kisselev e Alsufieva (2017), que se propuseram a analisar a dinâmica do uso de frases complexas que contêm conjunções por aprendizes de russo com níveis de proficiência de intermediário a avançado. Com base nos dados do RULEC, os autores criaram quatro *subcorpora* separando os textos dos aprendizes pelo nível e *background* (HL Intermediário, HL Avançado, FL Intermediário e FL Avançado), extraíram uma lista de palavras para cada um dos *subcorpora* e, em seguida, pesquisaram as listas de palavras para estabelecer quais conjunções os aprendizes usaram em sua escrita. Usando a lista de conjunções extraídas como guia, os autores conduziram uma análise abrangente das linhas de concordância (ou seja, amostras de linguagem que contêm todas as conjunções em questão) extraídas dos quatro *subcorpora*. Tendo analisado e categorizado as frases complexas extraídas, os autores avaliaram as mudanças quantitativas no uso estrutural e funcional de estruturas de frases complexas, bem como as taxas de precisão e os padrões de erro nos grupos HL e L2 nos níveis Intermediário e Avançado de proficiência linguística. Por exemplo, foi observado que não houve mudanças numéricas na quantidade de estruturas subordinadas usadas pelos aprendizes de FL, mas a frequência de estruturas subordinadas usadas pelos aprendizes de HL aumentou. No entanto, de fato, os números convergiram para os níveis avançados para ambos os grupos, sugerindo que, provavelmente, os alunos de FL geralmente começam a adquirir a habilidade de conectar ideias por escrito através do uso de várias conjunções mais cedo, uma vez que a exposição deles ao russo é fortemente baseada

na leitura e escrita desde os níveis iniciantes. Os alunos de HL, que tendem a iniciar os cursos de nível universitário e começar a alfabetização acadêmica após terem desenvolvido habilidades orais de nível intermediário, acabam tendo que trabalhar *overt marking* de sintaxe complexa nos níveis intermediário e avançado. Kisselev e Alsufieva (2017) também analisaram os tipos funcionais e estruturais de frases com conjunções e descobriram que os tipos menos frequentes e estruturas mais complexas foram melhor representados nos níveis avançados para ambos os grupos, com os aprendizes de HL exibindo vantagem sobre os aprendizes de FL em relação às estruturas que requerem manipulação estrutural dos constituintes da oração subordinada (por exemplo, *to, tchto* 'que'; *tchtoby* 'a fim de'; *kotoryi* 'qual').

Um estudo subsequente (KISSELEV; KOPOTEV; KLIMOV, no prelo) abordou em grande parte a mesma questão do desenvolvimento de estruturas de frases complexas, mas empregou uma análise computacional mais avançada. Primeiro, os autores analisaram os dados brutos do RULEC usando uma ferramenta de aplicação de NLP treinável chamada UDPipe (STRAKA e STRAKOVÁ, 2017) que fornece tokenização, lematização e análise morfológica e sintática de dados de linguagem.

Em seguida, usando scripts Python feitos pela própria equipe, os pesquisadores analisaram e compararam dados produzidos por quatro grupos de aprendizes (HL Intermediário, HL Avançado, FL Intermediário e FL Avançado) em relação aos doze índices de complexidade sintática geral. Esses índices incluíram: comprimento médio das frases, proporções de orações coordenadas e subordinadas em relação ao número total de orações, proporção de tipos específicos de subordinação (orações infinitivas, modificadores de orações adverbiais, e orações relativas, gerundiais e participiais) e medidas de "profundidade" frasal (ou seja, profundidade de aninhamento máxima e média de uma frase sintática, bem como o número de frases com profundidade de aninhamento "rasa"). Os resultados do estudo confirmaram a maioria das observações do estudo anterior de Kisselev e Alsufieva (2017); por exemplo, os resultados de ambos os estudos se alinharam com as conclusões de muitos estudos anteriores sobre a complexidade sintática conduzidos com os dados de *corpus* de L2 e demonstraram uma complexidade sintática geral na escrita de alunos mais avançados. E, enquanto o estudo de Kisselev e Alsufieva (2017) foi principalmente

descritivo, a abordagem computacional do estudo de Kisselev e colegas (2021) também tem implicações para a avaliação de domínio da língua russa, mostrando como características sintáticas específicas se correlacionam com vários níveis de proficiência de aprendizes de russo.

A diferença entre os dois estudos não é simplesmente a diferença entre as possibilidades de marcação gramatical *versus* dados brutos; o fato é que diferentes questões de pesquisa podem exigir tratamento diferente dos dados e uma combinação diferente de métodos qualitativos e computacionais. Por exemplo, o foco do estudo de Peirce (2018), que também utilizou os dados do RULEC, foi rastrear o desenvolvimento da precisão na morfologia nominal envolvendo o caso genitivo de substantivos, adjetivos e determinantes. Ao se propor a analisar esse tipo específico de erro (ou seja, erros de caso genitivo), o autor teve que recorrer a um método que integrava a codificação manual de erros e a marcação desses erros para posterior análise computacional usando um *software* específico (aqui, oXygen XML Editor).

A combinação dos benefícios da análise do avaliador humano com a eficácia dos procedimentos baseados em *corpus* permitiu a Pierce considerar diferentes fatores que possivelmente afetam o desenvolvimento dessa característica morfológica em aprendizes de russo. O estudo utilizou como variáveis independentes os metadados disponíveis no *corpus* RULEC, especificamente a restrição de tempo na escrita do texto (cronometrada ou não cronometrada) e o *background* de aprendizagem de línguas (HL ou FL). A comparação da quantidade e dos tipos de erros por grupo e por característica de restrição de tempo permitiu ao autor discutir os resultados do estudo à luz do papel central que a exposição precoce/tardia à linguagem desempenha na aquisição da linguagem, tanto em possíveis representações de características funcionais nominais em dois grupos de aprendizes quanto nas restrições de processamento às quais os dois grupos de aprendizes podem estar sujeitos em condições de tarefa cronometrada. Como os estudos revisados acima demonstram, um estudo de *corpus* pode ser mais ou menos dependente de tecnologia para melhor abordar os focos da pesquisa (e talvez, para se adequar ao nível de familiaridade dos pesquisadores com procedimentos baseados em *corpus*). No entanto, o potencial de *corpora* de aprendizes com *tags* de erros

não pode ser sobrestimado. A análise sistemática de erros, tal como agrupar os erros por frequência, por características do grupo (como níveis de proficiência, idade ou envolvimento dos pais) e por propriedades estruturais e funcionais, pode trazer informações cruciais sobre os processos de desenvolvimento da linguagem e os fatores que o influenciam. A análise de erros pode ajudar a testar hipóteses sobre os efeitos relativos da interferência de L1 e a proficiência em L2/HL, entender o impacto das práticas de instrução e de diferentes histórias de aprendizagem e responder a muitas outras questões importantes que ainda são pouco pesquisadas na aquisição de línguas de herança.

Um ambicioso projeto de *corpus* de larga escala, o *Russian Learner Corpus* (RLC, *Corpus* de Aprendizes de Russo, http://web-corpora.net/RLC) promete fornecer ao campo de estudos de russo seu primeiro *corpus* com a anotação completa de erros. Embora o *corpus* ainda esteja em construção e os *subcorpora* não estejam bem equilibrados, o repositório atualmente abriga uma grande coleção de textos orais e escritos (aproximadamente três mil amostras de fala, RAKHILINA *et al*., 2016) produzidos por falantes de russo como L2 e HL que possuem diferentes níveis de proficiência linguística e uma variedade de línguas dominantes (atualmente, mais de 20 L1s diferentes estão listados no *site*). O RLC está prontamente disponível na forma de dados brutos e com *tags* de classe gramatical POS, e pelo menos uma parte significativa do *corpus* está configurada para receber as *tags* de erros.

Em um estudo recente, Eremina (2020) utilizou as partes marcadas do *corpus* RLC (indiscriminadamente, sem considerar L1) para trabalhar com a *tag* de erro “Idiom” que marca uma expressão que contém muitas palavras e é incorreta. A pesquisadora classificou as expressões incorretas extraídas em dois tipos principais, estruturais e semânticos, e então analisou esses subtipos ainda mais profundamente, levantando hipóteses sobre a natureza de cada erro. Embora o estudo não se arrisque a implementar nenhum procedimento estatístico, ele estabelece as bases para análises estatísticas subsequentes de vários tipos de expressões fraseológicas produzidas por aprendizes de russo como L2. Considerando a crescente atenção que os campos de SLA e pedagogia da linguagem estão prestando à capacidade dos aprendizes de L2 de usar com sucesso as expressões formulaicas em sua língua-alvo, os estudos que abordam o desenvolvimento da complexidade fraseológica em russo como L2 são muito necessários.

Embora o trabalho realizado pela equipe do RLC exija anotação manual, o campo da linguística computacional está enfrentando problemas de detecção e correção automática de erros. Vários projetos de pesquisa têm se dedicado à questão metodológica da detecção automática de erros em idiomas morfologicamente ricos, incluindo o russo (ROSEN *et al*., 2014; ROZOVSKAYA; ROTH, 2018). Quanto mais *corpora* de aprendizes estiverem disponíveis para esses pesquisadores, melhor eles poderão treinar modelos computacionais para reconhecer padrões de desenvolvimento específicos nos dados linguísticos.

Felizmente, o desenvolvimento de *corpora* de aprendizes de russo está em ascensão. Um desses projetos é o *Multilingual Academic Corpus of Assignments – Writing and Speech* (MACAWS, *Corpus* Acadêmico Multilíngue de Tarefas - Escrita e Fala https://sites.google.com/email.arizona.edu/macawswebinar/home), que inclui dados de aprendizes de russo coletados nas atividades regulares em sala de aula. O *corpus* está disponível *online*; atualmente possui mais de mil textos produzidos por 100 aprendizes de russo, principalmente do primeiro e segundo anos de ensino (para mais informações sobre o *corpus*, ver Novikov e Vinokurova, 2022). Dois outros projetos de *corpora* de aprendizes também estão atualmente sendo desenvolvidos (ambos estão disponíveis mediante solicitação). O *Middlebury Russian Corpus of Learner Language* (MiRuCLL, *Corpus* de Aprendizes de Russo de Middlebury, por Kisselev e colegas, a ser lançado) é um *corpus* de desenvolvimento que contém dados coletados de aprendizes de russo como L2 no início e no final de um programa intensivo de verão imersivo. A característica única desse *corpus* é a disponibilidade de informações sobre o nível de proficiência do aluno no início e no final do período de instrução. A avaliação de proficiência é baseada na escala de proficiência ACTFL, tornando os dados potencialmente comparáveis a muitas outras amostras de dados.

Outro *corpus* de aprendizes de russo apresentado na pesquisa atual é o *Russian Essay Corpus* (*Corpus* de Ensaios em Russo, KISSELEV, 2019; KISSELEV *et al*., no prelo). Esse *corpus* é compilado a partir de textos extraídos do anual *National Post-Secondary Russian Essay Contest* (NPSREC, Concurso Nacional de Ensaios em Russo no Nível Pós-Secundário) patrocinado pelo American Council of Teachers of Russian (ACTR, o Conselho Americano de Professores de Russo). Em evento anual, o NPSREC atrai ampla participação de

estudantes de russo em todos os EUA. Após a conclusão do ciclo de premiação, os ensaios dos alunos totalmente anonimizados são disponibilizados aos pesquisadores. Até agora, pelo menos um ano de dados do NPSREC foi coletado e processado como um *corpus* de dados de aprendizes de russo transversal independente. Enquanto o *corpus* RULEC tem um pequeno número de participantes que contribuíram com muitos textos durante um longo período de tempo, o *Russian Essay Corpus* representa um grande número de estudantes que vem de uma grande variedade de programas em todo o país. Os níveis de proficiência dos participantes do *corpus* são indexados como faixas de horas de instrução (o nível 1 inclui aprendizes que receberam menos de 100 horas de instrução, o nível 2 entre 100 e 200 horas de instrução e assim por diante); no entanto, uma pequena parte dos textos do *corpus* também é classificada de acordo com a escala de proficiência ACTFL. O *background* de aprendizagem de línguas difere entre os aprendizes de russo como HL e FL. O *Russian Essay Corpus* tem o potencial de produzir resultados que são facilmente generalizáveis em vários grupos de aprendizes de russo.

Esses *corpora* estão se tornando uma ferramenta importante para os pesquisadores que trabalham com a língua russa e professores de língua russa, uma vez que as pesquisas com base nesses *corpora* têm o potencial de enriquecer significativamente nossa compreensão dos caminhos de desenvolvimento de aprendizes de língua russa, auxiliar nas práticas de avaliação e ajudar a avaliar as práticas de instrução. No entanto, para que essa promessa se concretize plenamente, são necessários mais e maiores *corpora* de aprendizes e muito mais pesquisas baseadas em *corpus* no campo. Na seção a seguir, a autora descreve algumas considerações práticas e etapas específicas na criação de *corpora* de aprendizes personalizados e na condução de estudos baseados em *corpus*.

2 Começando uma pesquisa em linguística de *corpus*: alguns *know-hows*

2.1 Coleta de dados e compilação de *corpus*

Nem todo conjunto de dados linguísticos pode ser chamado de *corpus*; de fato, a compilação de *corpus* requer vários cuidados por parte do pesquisador e um planejamento considerável. Os princípios específicos de como coletar e processar os dados que são inseridos em um *corpus* têm tanta importância para um estudo baseado em *corpus* quanto os métodos computacionais usados na análise. Esses princípios incluem a autenticidade e o tamanho dos dados linguísticos, bem como a sistematicidade da seleção dos dados, a representatividade dos dados e o quanto eles são equilibrados.

Autenticidade dos dados do corpus. Um dos princípios mais importantes da linguística de *corpus* é seu foco na linguagem autêntica, ou seja, a linguagem usada por seus falantes em contextos comunicativos autênticos. Acredita-se que investigar a linguagem autêntica, em vez de amostras de linguagem criadas em experimentos linguísticos, permite a ultrapassar os possíveis vieses que invadem os dados coletados em ambientes experimentais. Muitos *corpora* contemporâneos que agora são coletados para propósitos específicos, especialmente os *corpora* de aprendizes, representam efetivamente os dados elicitados. No entanto, esses dados elicitados vêm na forma de narrativas elicitadas, entrevistas e outros tipos de discursos situacionalmente fundamentados. A autenticidade também permite a inclusão de aspectos contextuais e situacionais na análise, através do seu registro como metainformações.

Para garantir que os resultados de uma análise de *corpus* sejam generalizáveis, os *corpora* normalmente são grandes. Ao mesmo tempo, o tamanho de um *corpus* é um padrão relativo; por um lado, os *corpora* precisam ser grandes o suficiente para permitir a aplicação de análises estatísticas e operações estatísticas generalizáveis, mas podem ser menores se o objetivo do *corpus* for restrito a uma questão de pesquisa específica ou a um contexto local. Assim, um conjunto de redações em sala de aula e/ou apresentações orais coletadas em intervalos regulares durante um ano letivo ou mesmo um semestre do mesmo grupo de alunos

pode se tornar um *corpus* a ser usado para avaliar o progresso dos alunos ou a eficácia das abordagens de instrução nesse contexto de instrução específico (BIBER; CONRAD; REPPEN, 2004).

A seleção sistemática de dados garante que a amostragem de dados não seja aleatória e seja claramente relevante para questões de pesquisa. Observe que, mesmo no caso de um *corpus* em sala de aula de pequena escala, o professor-pesquisador deve atentar para a sistematicidade da coleta de dados, considerando os intervalos em que os dados são coletados, o modo de coleta de dados (por exemplo, em casa ou em sala de aula, manuscritos ou digitados, etc.), e o tipo de dados (por exemplo, o gênero e a modalidade de produção da linguagem). O princípio da sistematicidade está inerentemente ligado ao princípio da *representatividade* dos dados, que garante que os dados encontrados no *corpus* representem um modo, variedade ou gênero específico de uma língua ou de um determinado grupo de falantes da forma mais completa possível. Para ilustrar, um grande *corpus* nacional como o *Russian National Corpus* (RNC, https://ruscorpora.ru) somará muitos milhões de palavras e sua representatividade é alcançada pela inclusão de textos de vários modos de comunicação (escrito, falado, multimodal e modos intermediários, como mensagens de texto), de vários gêneros e por diferentes autores, representando diversas variedades regionais e históricas.

Assim, os resultados de uma investigação em larga escala feita sobre os dados do RNC podem ser considerados representativos do estado atual da língua russa. Os *corpora* de aprendizes de russo revisados na seção acima, por exemplo, são representativos (em grau variável) de aprendizes de russo que têm inglês como L1 e, portanto, certas observações e conclusões baseadas no estudo desses *corpora* podem não ser generalizáveis para falantes de todos os L1. E, finalmente, os dados inseridos no *corpus* devem ser equilibrados entre autores individuais, tipos de texto, registros, modos, etc. Por exemplo, no caso de *corpora* em sala de aula, o pesquisador deve garantir que o número de textos de autoria dos alunos participantes é razoavelmente igual, que esses textos são de um certo modo semelhantes em extensão e/ou que nenhum gênero ou modo textual está super-representado no *corpus*; isso é feito para evitar o efeito potencial de um (ou um pequeno subconjunto) dos parâmetros de texto. Compiladores de *corpora* de grande escala geralmente precisam se esforçar muito para

garantir que os dados do *corpus* sejam equilibrados, para garantir a eficácia e a significância dos procedimentos analíticos subsequentes.

Como se pode supor com base nos princípios descritos acima, um *corpus* não é apenas um (grande) conjunto de dados linguísticos; um *corpus* linguístico é uma coleção considerável e legível por máquina, compilada sistematicamente e equilibrada, de textos autênticos que são representativos de uma língua ou de uma variedade específica de línguas. A subseção a seguir analisa como um pesquisador pode processar e analisar os dados do *corpus*.

2.2 Anotação de *corpus*

Uma vez que o *corpus* é bem desenhado e os dados são coletados, eles devem ser sistematicamente descritos. Conforme mencionado na seção anterior, a descrição do texto é fornecida na forma de *metatags*, que normalmente acompanham cada texto ou arquivo no *corpus*. Essas *metatags* podem incluir o nome do autor do texto (ou qualquer identificador (ID) único do texto, como um pseudônimo ou um número), dados biográficos (idade, sexo, primeira língua), data de criação/ocorrência, gênero do texto e qualquer outros metadados que possam ser úteis aos propósitos do corpus. Descritores de metadados podem então ser usados como variáveis nas análises dos dados do corpus. O corpus RULEC, por exemplo, registra várias características do texto e do aluno na "Caixa de Identificação do Cabeçalho", conforme ilustrado abaixo (veja a *Figura 1*). O uso dessas informações pode ajudar o pesquisador a agrupar os dados usando alguns desses parâmetros e/ou, em geral, considerar esses parâmetros do aprendiz e do texto como variáveis que podem afetar os parâmetros linguísticos da produção linguística.

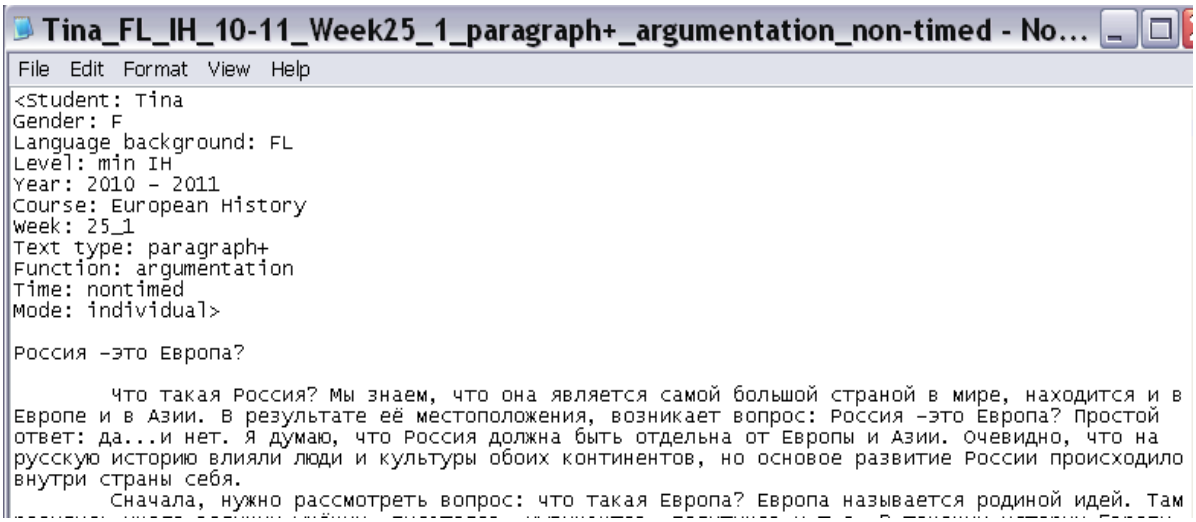

Figura 1: ID de cabeçalho de texto do *corpus* RULEC. Reproduzido com permissão de Alsufieva e colegas (2012).

Embora os metadados sejam uma condição *sine qua non* do *design* e da compilação do *corpus*, informações adicionais também podem ser adicionadas para rotular ou anotar palavras, frases e quaisquer unidades de texto com significado, mais longas ou mais curtas. A anotação (ou marcação) pode fornecer informações diferentes sobre as unidades no nível de texto e pode incluir informações morfossintáticas (por exemplo, anotação da classe gramatical, bem como pessoa, número, gênero, caso, voz, tempo verbal, aspecto), informações sintáticas (por exemplo, análise sintática de frases), informações semânticas (por exemplo, desambiguação de sentido da palavra, animacidade, contável/não contável), informações discursivas (por exemplo, atos de fala), *tags* de erro e/ou qualquer outra informação necessária para um *corpus* específico para a pesquisa.

Esta informação adicional é "anexada" a unidades linguísticas relevantes na forma de *tags*, razão pela qual a anotação é muitas vezes referida como *tagging* (etiquetagem). Consulte a Tabela 1 para obter um exemplo de uma frase de aprendiz analisada sintaticamente com o analisador UDPipe.

ID	Token	Lema	Classe gramatical	Anotação morfolófica	ID sintática	Anotação sintática
1	Мы	мы	PRON	Case=Nom|Number=Plur|Person=1	2	nsubj
2	живём	жить	VERB	Aspect=Imp|Mood=Ind|Number=Plur|Person=1|Tense=Pres|VerbForm=Fin|Voice=Act	0	root
3	в	в	ADP	_	4	case
4	мире	мир	NOUN	Animacy=Inan|Case=Loc|Gender=Masc|Number=Sing	2	obl
5	,	,	PUNCT	_	7	punct
6	где	где	ADV	Degree=Pos	7	advmod
7	ничего	ничего	ADV	Degree=Pos	4	acl:relcl
8	,	,	PUNCT	_	10	punct
9	просто	просто	PART	_	10	advmod
10	чёрная	черный	ADJ	Case=Nom|Degree=Pos|Gender=Fem|Number=Sing	7	conj
11	и	и	CCONJ	_	12	cc
12	белая	белый	ADJ	Case=Nom|Degree=Pos|Gender=Fem|Number=Sing	10	conj
13	.	.	PUNCT	_	2	punct

Tabela 1. Amostra do resultado de anotação com UDPipe. Reproduzido com permissão de Kisselev, Kopotev e Klimov (no prelo).

Conforme discutido na seção anterior, o nível de anotação precisa, antes de tudo, ser definido pelo foco de pesquisa prevista com o *corpus* e outros esquemas de anotação, disponíveis comercialmente ou publicamente ou personalizados, podem ser aplicados aos dados.

2.3 Ferramentas analíticas e procedimentos de trabalho com *corpus*

Um *corpus* bem compilado e bem descrito pode ser submetido a uma série de procedimentos estatísticos; a maioria desses procedimentos se enquadra em algum tipo de recuperação de dados, obtenção de frequências e análise estatística. Essas análises são realizadas com ajuda de *softwares* específicos para a análise de *corpus* que podem funcionar

de maneira independente (instalados no computador pessoal) ou serem acessados diretamente na *web*. Os programas mais usados entre os que podem ser instalados são o WordSmith Tools (SCOTT, 2016), para o uso do qual é necessário adquirir uma licença, e o AntConc (ANTHONY, 2019) de *download* gratuito. Uma série de ferramentas que podem ser acessadas diretamente na *web* fornecem procedimentos de análise semelhantes (veja, por exemplo, o conjunto de ferramentas NLP Tools for Social Sciences, descrito por Kyle e Crossley em 2015, ou LancsBox, descrito por Brezina e colegas em 2020). A funcionalidade desses programas pode variar, mas efetivamente todos eles são projetados para fornecer aos pesquisadores da linguagem ferramentas e formas rápidas, automáticas e significativas de organização, extração e análise de dados de *corpus*. Utilizando tais ferramentas computacionais, um pesquisador pode realizar vários procedimentos de análise. Alguns dos procedimentos comuns que ajudam a analisar os dados de *corpus* são listados abaixo.

Obtenção de estatísticas descritivas. Um conjunto de informações estatísticas descritivas gerais sobre os dados do *corpus* pode ser facilmente obtido até com ajuda de ferramentas de análise de *corpus* básicas. Um pesquisador pode obter rápida e automaticamente informações sobre o número de palavras e *tokens* de palavras, número de frases, número de parágrafos e, especialmente importante, o tamanho dessas unidades linguísticas, etc. Vários estudos mostraram que medidas baseadas em tamanho podem ser usadas com sucesso para classificar o grau de desenvolvimento linguístico do aprendiz. Por exemplo, um aumento no tamanho de um texto produzido dentro de uma certa janela de tempo, bem como o tamanho de uma oração (ou seja, um número médio de palavras por oração) e o tamanho de uma frase (ou seja, número médio de palavras por frase) podem indicar o grau de desenvolvimento linguístico geral ou da proficiência (NORRIS; ORTEGA, 2009; BULTÉ; HOUSEN, 2012; POLAT *et al*., 2019; KISSELEV *et al*., no prelo, entre outros). Até o tamanho das palavras aumenta com o nível de proficiência em russo (KISSELEV *et al*., no prelo).

As estatísticas descritivas também costumam incluir informações sobre a relação tipo/*token* (TTR, *type/token ratio*), ou seja, a porcentagem de palavras únicas (lemas) ou formas de palavras em relação a todas as palavras do *corpus*. A dinâmica de TTR que ilustra

a diversidade do vocabulário dos aprendizes também foi demonstrada ao refletir as diferenças qualitativas na proficiência linguística (LEE; JANG; SEO, 2009; KISSELEV *et al*., no prelo).

Extração de listas de palavras. Com este procedimento, um pesquisador pode compilar uma lista de palavras (lemas ou formas de palavras) utilizadas no *corpus* (ou *subcorpora*); a(s) lista(s) pode(m) ser ordenada(s) por ordem alfabética ou por frequência e podem posteriormente ser comparadas com dicionários de frequências padrão (por exemplo, "Novyi chastotnyj slovar' russkoi leksiki" ["O novo dicionário de frequência do léxico em russo"], LJASHEVSKAJA; SHAROV, 2010), dicionários de alunos (por exemplo, "A frequency dictionary of Russian: Core vocabulary for learners" ["Dicionário de frequências em russo: vocabulário básico para alunos"], SHAROFF *et al*., 2014) ou listas de um mínimo lexical (por exemplo, "Leksicheskii mínimum po russkomu kak inostrannomu" ["O mínimo lexical para russo como língua estrangeira"], 2013). Comparar a produção do aprendiz com esses recursos lexicais existentes e/ou listas de palavras escolhidas com base nos *corpora* de aprendizes com vários níveis de proficiências ou *backgrounds* pode ajudar os pesquisadores a avaliar rapidamente as habilidades lexicais dos alunos.

As listas de palavras também são um ponto de partida útil para muitas outras pesquisas qualitativas, fornecendo uma primeira impressão sobre os conhecimentos do aprendiz e alguns padrões no conhecimento lexical. Pode-se avaliar rapidamente o uso excessivo/subutilizado de itens lexicais, ver erros e padrões de erros ou simplesmente explorar os dados lexicais para uma futura análise mais aprofundada dos padrões de uso lexical. Por exemplo, um dos primeiros estudos de *corpus* de aprendizes de russo, conduzido por Pavlenko e Driagina (2007), concentrou-se na aquisição de vocabulário emocional por falantes de russo como L2 (com inglês L1). Os pesquisadores coletaram três pequenos *corpora* de narrativas orais produzidas por estudantes de russo americanos (em russo), monolíngues russos (em russo) e monolíngues americanos (em inglês). Os autores compararam as frequências de uso e o quanto o uso das palavras referentes às emoções foi adequado (por exemplo, *rasstraivaetsia* 'ficar chateado', *grustnoe* 'triste'), assim como as mesmas características para os seus radicais (por exemplo, *rasstra/o-* 'chateado', *grust-* 'triste'), entre os três grupos e descobriram que, ao contrário de monolíngues russos, que

mostraram uma forte preferência por verbos ao descrever estados emocionais, os aprendizes preferiram usar em russo as construções com adjetivos (semelhante a americanos monolíngues que falam inglês como L1); os aprendizes também usaram uma gama menor de palavras referentes às emoções e muitas vezes confundiram ou violaram as restrições conceituais sobre o uso do vocabulário referente à emoção (por exemplo, ao empregar *razozlilas* 'ficou bravo' em vez de *rasstroilas* 'ficou chateado'). Uma série de questões de pesquisa (principalmente com foco no vocabulário) pode ser conduzida usando os dados brutos de *corpora* e esses procedimentos simples.

Obtenção e classificação de linhas de concordância. Concordâncias são amostras de linguagem que podem ser extraídas de um *corpus* automaticamente. As concordâncias contêm o termo de pesquisa (geralmente uma palavra, uma frase ou mesmo um radical de palavra) que o pesquisador escolhe investigar. As concordâncias podem ser classificadas de diferentes maneiras: alfabeticamente ou por palavras à esquerda/direita do termo de pesquisa, etc. Uma classificação desse tipo permite ao pesquisador identificar diferentes padrões nos dados. Se os dados do *corpus* forem gramaticalmente analisados, pode-se também extrair concordâncias usando *tags* gramaticais como itens de busca.

Obtenção de listas de collocates *e* colligates. *Collocations* (colocações) são unidades compostas por várias palavras ou cadeias lexicais que ocorrem concomitantemente com o termo de pesquisa com mais frequência do que seria esperado por acaso. Essas expressões formulaicas (ou *nesvobodnye slovosotchetaniia* em russo) são notoriamente desafiadoras para os aprendizes de idiomas e são as principais candidatas para serem abordadas dentro de temas de pesquisa na área de SLA e intervenção pedagógica. As expressões formulaicas em russo representam uma variedade de tipos estruturais e incluem sequências de adjetivo + substantivo (por exemplo, *sloznaia problema, trudnaia zadatcca, krepkii tchai, sil'noe lekarstvo*), advérbio + verbo (*krepko zadumat'sia, sil'no tolknut'*), preposição + substantivo (*na rabote, v stole*), etc. Um pesquisador pode potencialmente extrair todos os n-gramas (ou seja, sequência de duas, três ou mais palavras) de um *corpus* e analisá-los manualmente ou empregar as análises estatísticas incorporadas a um *software* de análise de *corpus* para estabelecer uma lista de padrões recorrentes.

Da mesma forma, *colligation* refere-se ao fenômeno da formulaicidade, mas com construções gramaticais, em vez de lexicais. Por exemplo, as construções "*igrat' v* + acusativo" e "*igrat' na* + prepositivo" são *colligations*. O trabalho de Apresjan (2017) é uma ilustração apropriada de pesquisa de *colligation* em russo como L2 e HL. O estudo investiga construções possessivas em russo com e sem o uso explícito do verbo existencial *est'* usando os dados do RLC. A busca no *corpus* foi formulada como "*u* + genitivo (substantivo, pronomes) + *est'*" e "*u* + genitivo (substantivo, pronomes) + nome (substantivo)" (APRESJAN, 2017, p.86). O autor analisou as linhas de concordância extraídas com o objetivo de entender se há algumas regras específicas, semânticas e pragmáticas, que governassem o uso dessas construções por aprendizes de HL e L2. Os resultados revelaram que os aprendizes de HL podem usar as construções com êxito, considerando todos os significados semânticos e pragmáticos, enquanto os dados de L2 contêm várias instâncias errôneas. Especificamente, os aprendizes de L2 cometeram o dobro de erros nas construções onde *est'* não estaria presente, sugerindo que os aprendizes de russo como L2 podem precisar de instrução adicional em relação a essa estrutura.

O desenvolvimento da habilidade de uso de fraseologismos pelos aprendizes de L2 é uma área de grande interesse em SLA (PAQUOT; GRANGER, 2012). E as pesquisas potenciais nessa área agora são facilitadas com o desenvolvimento de dicionários frasais (por exemplo, "*Slovar' russkoi idiomatiki*", Kustova, n.d.; "*Slovar' glagol'noi sotchetaemosti nepredmetnyx imion russkogo iazyka*", Biriuk, Gusev, e Kalinina, s.d.) e plataformas para investigar *collocations* e *colligations* em grandes *corpora* padrão (por exemplo, CoCoCo, Kopotev, 2020), que fornecem informações específicas sobre padrões lexicais e gramaticais em russo padrão e podem servir como linha de base na análise de dados de aprendiz.

Os procedimentos descritos acima são somente alguns de muitos. O número – e a sofisticação – de procedimentos baseados em *corpus* disponíveis hoje está em contínua expansão; no entanto, o principal objetivo dessas ferramentas é permitir que um pesquisador se envolva com grandes quantidades de dados autênticos e extraia e examine várias amostras de unidades linguísticas produzidas na fala e na escrita pelos portadores das variedades linguísticas em foco. Ao extrair, classificar e analisar (estatística ou manualmente) as

estruturas linguísticas escolhidas para análise, o pesquisador pode procurar regularidades e padrões no uso da linguagem que, de outra forma, escapariam da atenção de pesquisadores e professores de línguas.

Conclusão e *desiderata*

De maneira geral, a tarefa dos pesquisadores que trabalham com a aquisição de L2 e línguas de herança é entender os processos mentais que fundamentam a produção e o desenvolvimento linguístico em falantes de L2 e HL. *Corpora* linguísticos compostos de dados linguísticos produzidos em ambientes autênticos para fins comunicativos tornaram-se ferramentas importantes para fornecer aos pesquisadores algumas bases para a interpretação dos processos mentais e das representações mentais do conhecimento. Aliados a sofisticadas ferramentas computacionais que permitem a extração e análise dos dados de maneira rápida e confiável, os *corpora* linguísticos têm se mostrado uma ferramenta indispensável na pesquisa linguística, e as implicações pedagógicas dessa pesquisa são significativas para as práticas de sala de aula no estudo de línguas. Ao adotar abordagens baseadas em *corpus*, os campos de SLA de russo e de ensino de russo se beneficiam tremendamente, tanto pela expansão de nossa compreensão da natureza do desenvolvimento de russo como L2 e HL, quanto pela expansão das abordagens pedagógicas e repertórios de ensino e aprendizagem da língua russa.

REFERÊNCIAS

ALSUFIEVA, A.; KISSELEV, O.; FREELS, S. Results 2012: Using Flagship Data to Develop a Russian Learner *Corpus* of Academic Writing. *Russian Language Journal*, n. 62, pp.79-105, 2012.

ANDRJUSHINA, N.; KOZLOVA, T. *Leksicheskii minimum po russkomu yazyku kak inostrannomu. Bazovyj Uroven'* [*Lexical minimum for Russian as a foreign language. Basic level*]. 5.ed. St. Petersburg: Zlatoust, 2020.

ANTHONY, L. AntConc (Version 3.5.8) [Computer Software]. Tokyo, Japan: Waseda University. Available on: http://www.laurenceanthony.net/software, 2019.

APRESJAN, V. YU. Russkie possessivnye konstrukcii s nulevym i vyraženynnym glagolom: pravila i ošibki. *Russkij jazyk v naučnom osvesčenii*, n. 33, pp.86-116, 2017.

BIBER, D.; CONRAD, S.; REPPEN, R. Corpus *Linguistics: Investigating Language Structure and Use.* Cambridge: Cambridge University Press, 2004.

BIBER, D.; CONRAD, S. Corpus *Linguistics and Grammar Teaching*. White Plains, NY: Pearson Education, 2010.

BIRIUK, O.; GUSEV, V.; KALININA, YE. *Slovar' glagol'noj sochetaemosti nepredmetnyx imion russkogo yazyka* [Dictionary of Verbal Compatibility of Non-Objective Names of the Russian Language]. Available from http://dict.ruslang.ru/abstr_noun.php.

BOULTON, A. Data-Driven Learning and Language Pedagogy. *In:* THORNE, S., MAY, S. (eds.). *Language, Education and Technology.* Encyclopedia of Language and Education. New York: Springer, Cham, 2017.

BREZINA, V.; WEILL-TESSIER, P.; MCENERY, A. #LancsBox v. 5.x. [software]. 2020. Available from http://*corpora*.lancs.ac.uk/lancsbox.

BULTÉ, B.; HOUSEN, A. Conceptualizing and Measuring Short-Term Changes in L2 Writing Complexity. *Journal of Second Language Writing*, n. 26, pp.42-65, 2014.

CONRAD, S.; BIBER, D. *Real Grammar: A* Corpus-*Based Approach to English.* New York: Pearson/Longman, 2009.

CROSSLEY, S. A.; KYLE, K. Assessing Writing with the Tool for the Automatic Analysis of Lexical Sophistication (TAALES). *Assessing Writing*, n. 38, pp.46-50, 2018.

DONRUSHINA R. N.; LEVINZON, A. I. Informatsionnye tehnologii v gumanitarnom obrazovanii: Natsional'nyj korpus russkogo yazyka [Information Technologies in Humanities Education: National *Corpus* of the Russian Language]. *Voprosy obrazovaniia*, n. 4, 2006.

EREMINA, O. S. Russkie nesvobodnye vyrazhenia v rechi inostrantsev: korpusnyi podhod [Russian Formulaic Expressions in the Speech of Foreigners: *Corpus* Approach]. *Russkii jazyk za rubezhom*, n. 6, pp.29-35, 2020.

FURNISS, E. Using a *Corpus*-Based Approach to Russian as a Foreign Language Materials Development. *Russian Language Journal*, n. 63, pp.195-212, 2013.

GRANGER, S. The Contribution of Learner *Corpora* to Second Language Acquisition and Foreign Language Teaching. *In:* AJMER, K. (ed.). Corpora *and Language Teaching*. Philadelphia/Amsterdam: John Benjamins, 2009, pp.13-32.

GRIES, S. What is *Corpus* Linguistics? *Language and Linguistics Compass*, v. 3, n. 5, pp.1225-1241, 2009.

GRIES, S. Methodological and Interdisciplinary Stance in *Corpus* Linguistics. *In*: BARNBROOK, G.; VIANA, V.; ZYNGIER, S. (eds.). *Perspectives on* Corpus *Linguistics: Connections and Controversies.* Philadelphia/Amsterdam: John Benjamins, 2011, pp.81-98.

HUNSTON, S. Corpora *in Applied Linguistics*. Cambridge: Cambridge UP, 2002.

KISSELEV, O. *Corpus*-Based Methods in the Study of Heritage Languages. *In:* POLINSKY,

M.; MONTRUL, S. (eds.). *The Cambridge Handbook on Heritage Languages.* Cambridge University Press, 2021, pp.520-544.

KISSELEV, O. Word Order Patterns in the Writing of Heritage and Second Language Learners of Russian. *Russian Language Journal,* n. 69, pp.149-174, 2019.

KISSELEV, O.; KOPOTEV, M.; KLIMOV, A. Specific Markers of Syntactic Complexity in Academic Russian: A Longitudinal *Corpus* Study. *In:* LEŃKO-SZYMAŃSKA, A.; GÖTZ, S. (eds.). *Complexity, Accuracy & Fluency in Learner* Corpus *Research*. John Benjamins, forthcoming.

KISSELEV, O.; FURNISS, E. *Corpus* Linguistics and Russian Language Pedagogy. *In:* DENGUB, E.; DUBININA, I.; MERILL, J. (eds.). *The Art of Teaching Russian*. Washington: Georgetown University Press, 2020, pp.307-332.

KISSELEV, O.; ALSUFIEVA, A. The Development of Syntactic Complexity in the Writing of Russian Language Learners: A Longitudinal *Corpus* Study. *Russian Language Journal,* n. 67, pp.27-53, 2017.

KOPOTEV, M. Ispol'zovanie èlektronnyx korpusov v prepodavanii russkogo jazyka [The Use of Electronic *Corpora* in Teaching the Russian Language]. *In*: LINDSTEDT J. *et al*. (eds.), *SLAVICA HELSINGIENSIA 35, S ljubov'ju k slovu, Festschrift in honour of Professor Arto Mustajoki on the occasion of his 60th birthday*. Helsinki, 2008, pp.110-118.

KOPOTEV, M. O samom slozhnom: Izuchenie sochetaemosti slov online [About the Most Difficult: Learning the Combination of Words Online]. *Russkij jazyk za rubezhom*, n. 6, pp.36-43, 2020.

KOPOTEV, M.; MUSTAJOKI, A. Sovremennaja korpusnaja rusistika [Modern *Corpus* Russian Studies]. *In:* MUSTAJOKI, A.; KOPOTEV, M.; BIRJULIN, L.; PROTASOVA, YU. (eds.). *Instrumentarij rusistiki: Korpusnye podxody.* Helsinki: Helsinki UP, 2008, pp.7-24.

KUSTOVA, G.I. *Slovar' russkoi idiomatiki. Sochetaniya slov so znacheniyem vysokoi stepeni* [A Dictionary of Russian Idiomology. Word Combinations with the Significance of a High Degree]. Moscow, 2008. http://dict.rislang.ru/magn.php.

KYLE, K.; CROSSLEY, S. A. Automatically Assessing Lexical Sophistication: Indices, Tools, Findings, and Application. *TESOL Quarterly*, v. 49, n. 4, pp.757-786, 2015.

LEBEDEVA M. YU. Dano mne telo – chto mne delat' s nim? Primenenie korpusnyh tehnologii v lingvodidaktike RKI [I Have Been Given a Body - What Am I to Do with It? Application of *Corpus* Technologies in Linguodidactics of Russian as a Foreign Language.]. *Russkij jazyk za rubezhom*, n. 6, pp.4-13, 2020.

LEECH, G. Corpora and Theories of Linguistic Performance. *In:* SVARTVIK, J. (ed.). *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82.* Berlin, New York: Mouton de Gruyter, 1992, pp.105-122.

LEECH, G. Teaching and Language *Corpora*: A Convergence. *In:* WICHMANN, A. *et al*. (ed.). *Teaching and Language* Corpora. London and New York: Routledge, pp.1-24, 2014.

LEE, S. H.; JANG, S. B.; SEO, S. K. Annotation of Korean Learner *Corpora* for Particle Error Detection. *CALICO Journal*, v. 26, n. 3, pp.529-544, 2009.

LJASHEVSKAJA, O. N.; SHAROV, S.A. *Chastotnyi slovar' sovremennogo russkogo yazyka:* Na materialax Natsional'nogo korpusa russkogo yazyka [Frequency Dictionary of the Modern Russian Language: On the Materials of the National Russian Corpus]. Azbukovnik, 2009.

LU, X.; YOON, J.; KISSELEV, O. Adding to Academic Formula Lists: Phrase-Frames for Research Article Introductions in Social Sciences. *Journal of English for Academic Purposes*, v. 36, pp.76-85, 2018.

MCENERY, T.; HARDIE, A. Corpus *Linguistics: Method, Theory and Practice*. Cambridge: Cambridge UP, 2012.

NORRIS, J.; ORTEGA, L. Measurement for Understanding: An Organic Approach to Investigating Complexity, Accuracy, and Fluency in SLA. *Applied Linguistics*, v. 30, n. 4, pp.555–578, 2009.

NOVIKOV, A.; VINOKUROVA, V. Learner Corpus as a Medium for Tasks. *In:* NUSS, S. V.; WHITEHEAD MARTELLE, W. (eds.). *Task-Based Instruction for Teaching Russian as a Foreign Language*. London and New York: Routledge, 2022.

PAVLENKO, A.; DRIAGINA, V. Russian Emotion Vocabulary in American Learners' Narratives. *The Modern Language Journal*, n. 91, pp.213-234, 2007.

PAQUOT, M.; GRANGER, S. Formulaic Language in Learner *Corpora*. *Annual Review of Applied Linguistics*, v. 32, n. 1, pp.130-149, 2012.

PEIRCE, G. Representational and Processing Constraints on the Acquisition of Case and Gender by Heritage and L2 Learners of Russian: A *Corpus* Study. *Heritage Language Journal*, v. 15, n. 1, pp.95-111, 2018.

POLAT, N.; MAHALINGAPPA, L.; MANCILLA, R. L. Longitudinal Growth Trajectories of Written Syntactic Complexity: The Case of Turkish Learners in an Intensive English Program. *Applied Linguistics*, v. 41, n. 5, pp.688-711, 2020.

RAKHILINA, E.; VYRENKOVA, A.; MUSTAKIMOVA, E.; LADYGINA, A.; SMIRNOV, I. Building a Learner *Corpus* for Russian. *In:* VOLODINA, E. *et al*. (ed.). *Proceedings of the Joint Workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition*. Umea, Sweden: LiU Electronic Press, 2016, pp.66-75.

ROSEN, A.; HANA, J.; ŠTINDLOVÁ, B.; FELDMAN, A. Evaluating and Automating the Annotation of a Learner *Corpus*. *Language Resources and Evaluation*, v. 48, n. 1, pp.65-92, 2014.

ROZOVSKAYA, A.; ROTH, D. Building a State-of-the-Art Grammatical Error Correction System. *Transactions of American Computational Linguistics*, v. 2, pp.419-434, 2014.

SCOTT, M. *WordSmith Tools Version 7* [Computer Program]. Stroud: Lexical Analysis Software, 2016.

SHAROFF, S.; UMANSKAYA, E.; WILSON, J. *A Frequency Dictionary of Russian: Core Vocabulary for Learners*. London and New York: Routledge, 2014.

STRAKA, M.; STRAKOVÁ, J. Tokenizing, Pos Tagging, Lemmatizing and Parsing ud 2.0 with Udpipe. In: *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 2017, pp.88-99.

Traduzido por *Anna Smirnova Henriques* – annsmile141@gmail.com

Recebido em 26/09/2021

Aprovado em 05/09/2022

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.

Parecer II

Esta é uma visão muito interessante do LCR russo.
Os autores devem fazer duas pequenas alterações:
> desde a década de 1990, houve mudanças significativas em muitas disciplinas sociais, incluindo linguística e estudos de linguagem aplicada, que viram o surgimento e o aumento da proeminência da nova disciplina de linguística de corpus
A Linguística de Corpus Computacional remonta à década de 1960. Por favor, altere isso.
> o repositório atualmente abriga uma grande coleção de textos
Por favor, especifique o tamanho (número de textos e/ou palavras).

Tony Berber Sardinha – Pontifícia Universidade Católica de São Paulo – PUC-SP, São Paulo, São Paulo. Brasil; https://orcid.org/0000-0001-8815-1521; tonycorpuslg@gmail.com

Disponibilidade de dados de pesquisa e outros materiais

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