Caligrama: Revista de Estudos Românicos
v. 29, n. 1, p. 24–42, 2024
ISSN: 2238-3824 · DOI: 10.17851/2238-3824.29.1.24-42
Submetido em: 30/05/2023 · Aprovado em: 26/11/2023

caligrama

Parâmetros para a compilação de um *corpus* em português brasileiro feito a partir de textos voltados para crianças

Guidelines for a Brazilian Portuguese Corpus Composed by Texts Written for Children

Larissa Moreira Brangel
Universidade Federal do Rio Grande do Sul (UFRGS) | Porto Alegre | RS | BR
larissabrangel@gmail.com
http://orcid.org/0000-0002-6850-3590

Beatriz Nogueira Sartori
Universidade Federal do Rio Grande do Sul (UFRGS) | Porto Alegre | RS | BR
biansart@gmail.com
http://orcid.org/0009-0003-2389-2849

Margot Luiza Pedron da Camara
Universidade Federal do Rio Grande do Sul (UFRGS) | Porto Alegre | RS | BR
margotcamaraestudante@gmail.com
http://orcid.org/0000-0002-6158-5554

Resumo: O presente estudo discute as possíveis aproximações entre a Linguística de Corpus e a Lexicografia Pedagógica no desenvolvimento de dicionários escolares de língua portuguesa voltados para crianças do 4º e 5º ano do Ensino Fundamental. Tradições lexicográficas desenvolvidas, como a de língua inglesa, costumam explorar a pesquisa com *corpus* como uma importante aliada no desenvolvimento de obras lexicográficas de excelência, e isso pode ser verificado nos mais diversos tipos de dicionários publicados por essas tradições. No caso de dicionários escolares, é possível encontrar estudos que demonstram como a compilação de *corpora* escritos para e por crianças pode ser útil no desenvolvimento de obras lexicográficas escolares. Com base em estudos desenvolvidos pela Lexicografia inglesa, o presente trabalho buscou transpor para a realidade brasileira diretrizes e métodos para o desenvolvimento de um *corpus* de textos para crianças e encontrou nas políticas públicas do livro didático uma grande aliada para essa tarefa. O Brasil, além de dispor de um ambiente favorável para o estudo ora proposto, devido à massiva publicação de dicionários escolares por editoras brasileiras e à intensa circulação desse produto nas salas de aulas do país, conta também com o Programa Nacional do Livro e do Material Didático (PNLD) como uma ferramenta adequada para o desenvolvimento de *corpus* de textos voltados para crianças, demonstrando, assim, um grande potencial para alavancar a pesquisa sobre dicionários escolares brasileiros.

Palavras-chave: Linguística de Corpus; Lexicografia Pedagógica; Programa Nacional do Livro Didático.

Abstract: The present study discusses the possible approaches between Corpus Linguistics and Pedagogical Lexicography in the development of a Brazilian Portuguese school dictionary for 4th and 5th grade children. Advanced lexicographical traditions, such as the English-speaking ones, usually explore corpus research as an important ally in the development of excellent lexicographical works, and this can be confirmed in the most diverse types of dictionaries published by them. In the case of school dictionaries, it is possible to find studies that demonstrate how the compilation of corpora written for and

by children can be useful in the development of school dictionaries. Based on studies developed by English lexicography, the present work tried to transpose to the Brazilian reality guidelines and methods for the development of a corpus of texts for children, and found in public policies of textbooks a great ally in this task. Brazil, besides having a favorable environment for the study proposed here, due to the massive publication of school dictionaries by Brazilian publishers and the intense circulation of this product in the country's classrooms, also counts on the National Textbook Program as an adequate tool for the development of a corpus of texts for children, thus demonstrating a great potential to leverage research on Brazilian school dictionaries.

Keywords: Corpus Linguistics; Pedagogical Lexicography; National Textbook Program.

1 Introdução

A prática lexicográfica pode ser considerada milenar, se remetermos suas origens aos "protodicionários" desenvolvidos pelos sumérios no berço da civilização mesopotâmica (Welker, 2004). Ainda que a história da Lexicografia seja marcada pelo caráter eminentemente prático da produção dicionarística, isto é, pela necessidade de produção de obras voltadas ao atendimento de necessidades pontuais de comunidades linguísticas, o século XX imprimiu aos dicionários importantes modificações, tanto ao instaurar a reflexão teórica a respeito dos dicionários, com a Metalexicografia (Bugueño Miranda, 2013; Bugueño Miranda; Borba, 2018), quanto ao alavancar novas abordagens para o estudo da linguagem, que impactaram fortemente na elaboração dos dicionários. Sobre esse segundo aspecto, evidenciamos a importância da aproximação entre a Lexicografia e a Linguística de Corpus (Landau, 2001; Atkins; Rundell, 2008; Kilgarriff, 2009).

A Linguística de Corpus teve origem antes do advento da informática, em projetos voltados essencialmente para o ensino e aprendizado de língua (Sardinha, 2000). Impulsionada pelo Corpus Brown, o primeiro *corpus* cujo conteúdo ostentou a marca de um milhão de palavras, a Linguística de Corpus ganhou novos contornos na década de 60, instaurando novas propostas de estudos da língua e colocando o empirismo novamente no cerne dos estudos da linguagem.[1] Na década de 80, o ambicioso projeto Cobuild, uma parceria entre a Editora Collins e a Universidade de Birmingham, publicou o primeiro dicionário baseado em exemplos reais da língua inglesa, mostrando ao mercado editorial de dicionários pedagógicos o forte potencial da Linguística de Corpus na produção de obras para aprendizes (Moon, 2007).

Desde então, as grandes editoras têm fomentado pesquisas com *corpus* e desenvolvido *corpora* cada vez mais arrojados. Atualmente, dicionários renomados no campo do ensino de

[1] Conforme discutido em Sardinha (2000), a obra *Syntactic Structures*, publicada por Chomsky em 1957, trouxe o racionalismo para o centro dos estudos linguísticos. Anos mais tarde, com o advento do Corpus Brown e as sucessivas inovações nas pesquisas com *corpus* proporcionadas pelo avanço da informática, o empirismo volta a ter reconhecimento na pesquisa linguística.

línguas atingem altos índices de excelência devido, em grande parte, à análise de *corpus*.[2] No entanto, no mercado brasileiro, onde prevalece a produção de obras lexicográficas escolares, chama a atenção a falta de materiais lexicográficos que mencionem a pesquisa com *corpus* no processo de elaboração das obras. Esse fato coloca a pesquisa lexicográfica brasileira na contramão da pesquisa de ponta sobre elaboração e uso de dicionários (cf. Brangel, 2016a).

Nesse contexto, o presente artigo apresenta os primeiros resultados da pesquisa intitulada "Bases teórico-metodológicas para a compilação de um dicionário para crianças fundamentado na semântica cognitiva". O projeto, como o próprio título sugere, almeja desenvolver princípios que auxiliem a compilação de dicionários de excelência, fazendo uso do aparato teórico da Semântica Cognitiva, principalmente no que diz respeito à Semântica de Frames e à Semântica Prototípica. A pesquisa prevê o desenvolvimento de duas etapas distintas, porém complementares, a saber: 1) compilação e análise de um *corpus* de textos voltados para crianças, e 2) desenvolvimento de metodologias para elaboração de dicionários. As páginas que seguem discorrerão sobre a etapa 1.

2 Dicionários infantis e a necessidade de um *corpus* de textos voltados para o público infantil

Na história da Lexicografia, a pesquisa com *corpus* tem ocupado uma função de destaque por revelar fatos sobre a língua que, em outras épocas, passavam despercebidos aos linguistas (Moon, 2010) e por auxiliar os lexicógrafos na seleção das informações que devem compor cada tipo de dicionário (Atkins; Rundell, 2008). Atualmente, grandes editoras do mercado lexicográfico fazem questão de divulgar seus trabalhos com *corpus*, de modo a sugerir a confiabilidade e a qualidade de seus produtos finais, como os dicionários de inglês das linhas Cambridge (Cambridge English Corpus, 1,8 bilhão de palavras), Oxford (Oxford English Corpus, 2,1 bilhões de palavras) e Collins (Collins Corpus, 4,5 bilhões de palavras).[3]

As editoras supracitadas são historicamente reconhecidas pela excelência na produção de dicionários para aprendizes de inglês como língua estrangeira (os chamados *learner's dictionaries*).[4] Essa excelência advém de diversos fatores, dentre os quais evidenciamos a longa tradição inglesa na elaboração de dicionários pedagógicos para falantes não nativos (como o primeiro exemplar, o *New Method English Dictionary*, publicado em 1935) e o fator mercadológico, que impulsiona o constante aprimoramento de materiais voltados para o ensino de inglês para falantes não nativos, uma vez que a língua inglesa desempenha o papel de língua franca na sociedade contemporânea. Um fato curioso, e de total importância para o presente

[2] A página na internet da linha de dicionários Collins é um bom exemplo. Hoje em dia, o consulente possui acesso a informações que vão muito além da ortografia e significado do item lexical pesquisado, encontrando um verdadeiro dossiê sobre as palavras consultadas, como vídeos mostrando diferentes pronúncias da palavra, a indicação de frequência dentro do *corpus* em língua inglesa, listas de expressões idiomáticas que apresentam a palavra em sua estrutura, indicação de sinonímia e antonímia, citações de obras literárias e/ou de pessoas famosas que contenham a palavra, lista de colocações, gráfico indicando tendências de uso da palavra ao longo dos anos, equivalentes tradutórios em outras línguas, dentre outras informações. Para maiores detalhes, ver: *Collins online dictionary*. Glasgow: HarperCollins, 2023. Disponível em: https://www.collinsdictionary.com/. Acesso em: 2 dez. 2023.

[3] Dados obtidos nos sites das editoras em 5 de fevereiro de 2024.

[4] Para uma discussão mais detalhada sobre essa classe de dicionários, conferir Oliveira (2010).

trabalho, é que o investimento massivo no aprimoramento de obras lexicográficas em língua inglesa voltadas para falantes não nativos impactou de maneira positiva outra classe de obras lexicográficas pedagógicas, os dicionários escolares.

Na taxonomia lexicográfica proposta por Bugueño Miranda (2014), os dicionários escolares são classificados como obras monolíngues, voltadas para falantes nativos, de discurso livre, com ênfase no significado, semasiológicas e diassistemicamente restritivas. Sob um ponto de vista funcional, os dicionários escolares resguardam a singularidade de serem obras elaboradas com o intuito de auxiliar o ensino de língua materna para estudantes em idade escolar. Essas características evidenciam que o dicionário escolar possui, além de uma série de traços caracterizadores altamente específicos, um público-alvo bastante delimitado e, consequentemente, uma função específica a cumprir junto a esse público-alvo (Farias, 2009).

Deste modo, a elaboração de um dicionário escolar deve estar respaldada por orientações teóricas e metodológicas que assegurem as especificidades da obra e, com isso, permitam satisfazer os anseios de seus consulentes. Cignoni *et al*. (1996, p. 662), por exemplo, concluem, a partir dos resultados obtidos em seus experimentos com um público infantil, que as crianças italianas preferem utilizar taxonomias, funções e atributos para definir substantivos concretos, ao passo que, para definir substantivos abstratos, a preferência recai sobre experiências pessoais e exemplos retirados do cotidiano. Uma descoberta dessa natureza pode ser útil à Lexicografia Pedagógica, pois serve de base para reflexões sobre a geração de definições de substantivos em dicionários escolares italianos, por exemplo, possibilitando a aplicação de descobertas sobre a compreensão do significado na reflexão e na compilação de dicionários.

Transpor descobertas como a citada acima para a compilação de dicionários escolares equivale a refletir sobre métodos adequados de definição para o consulente de dicionários escolares, assegurando que as informações oferecidas pela obra estejam em conformidade com a capacidade de interpretação de seu usuário. Um dos principais desafios da Lexicografia Pedagógica consiste, dessa forma, na elaboração de métodos que permitam estabelecer uma ponte de acesso eficaz entre o dicionário escolar e o seu pretenso consulente.

Retomando a pesquisa lexicográfica na língua inglesa, destacamos as contribuições que as pesquisas com *corpus* têm oferecido para a Lexicografia Pedagógica, nas últimas décadas, no sentido de orientar o lexicógrafo quanto às necessidades do público aprendiz (Moon, 2007). É possível, assim, observar, dentre outras investigações, a compilação e o estudo de *corpora* de textos em língua inglesa escritos para e por crianças em idade escolar (Banerji *et al.*, 2013; Wild; Kilgarriff; Tugwell, 2012). Inspirada pelo sucesso da Lexicografia inglesa, Brangel (2016a) já havia assinalado a importância e a necessidade de pesquisas com *corpus* voltadas para o desenvolvimento de obras lexicográficas escolares brasileiras. Análises de obras aprovadas pelo Ministério da Educação para serem distribuídas em escolas da rede pública mostraram que, embora essas obras tivessem passado por uma avaliação prévia de especialistas (cf. Brasil, 2012), elas ainda resguardavam uma série de problemas que colocavam em dúvida a sua qualidade e adequação ao público-alvo (cf. Brangel, 2013a, 2013b). Nos próximos parágrafos, listamos alguns exemplos.

Comecemos pela macroestrutura, componente canônico do dicionário semasiológico que corresponde à lista ordenada das entradas, ou seja, a progressão vertical do dicionário, geralmente disposta em ordem alfabética (Hausmann; Wiegand, 1989; Haensch *et al.*, 1982). Em uma análise macroestrutural de dicionários escolares, é pertinente verificar a quantidade e o tipo de palavras lematizadas na obra, uma vez que os itens lexicais que integram a macroestrutura devem estar em consonância com o público-alvo do dicionário (Bugueño

Miranda, 2007). A obra *Dicionário Júnior da língua portuguesa* (DiJr) (2005), aprovada pelo PNLD Dicionários 2012[5] e indicada para crianças do 2º ao 5º ano do Ensino Fundamental, apresenta um bom exemplo da falta de adequação macroestrutural ao lematizar *topless*, um item lexical cuja lematização é bastante questionável quando se trata de um dicionário voltado para crianças. *Saraiva Júnior* (SaJr) (2010), também aprovado pelo PNLD 2012 e indicado para o mesmo período escolar de DiJr (2005), curiosamente lematiza o item lexical *mesa*, mas não lematiza *cadeira*, uma falha notória da obra, haja vista a relação semântica dos dois itens lexicais e a alta frequência no vocabulário infantil. Esses dois exemplos ressaltam alguns problemas presentes na elaboração de macroestruturas de obras brasileiras.

Em outro nível de análise, é possível mencionar também problemas relacionados ao âmbito microestrutural de obras escolares brasileiras. A microestrutura de um dicionário semasiológico corresponde às informações internas ao verbete (Hartmann; James, 2002, *s.v. microstructure*; Hausmann; Wiegand, 1989), que podem ser divididas em informações relativas ao signo linguístico como significante (comentário de forma) ou informações sobre o signo linguístico como significado (comentário semântico). Fazem parte do comentário de forma, por exemplo, as informações relativas à ortografia, à gramática e à pronúncia, enquanto a definição, os exemplos e as marcas de uso integram o comentário semântico (Hartmann; James, 2002, *s.v. comment*). Para a presente discussão, trataremos apenas do comentário semântico, mais especificamente das definições lexicográficas, que, embora sejam um segmento informativo de suma importância na microestrutura, nem sempre cumprem, satisfatoriamente, a função de informar ao consulente o significado das palavras (Farias, 2013; Brangel, 2016b).

Em dicionários escolares, é possível encontrar definições lexicográficas de difícil compreensão, inclusive para adultos letrados. Exemplos disso são a definição de *arroz* oferecida por SaJr (2010), "Erva de até 1 metro, originária na Ásia, que produz pequenos grãos finos e longos" (SaJr, 2010, *s.v. arroz*) e a definição de *maracujá*, "planta trepadeira de frutos comestíveis" (SaJr, 2010, *s.v. maracujá*). No primeiro caso, a informação contida na definição não está em consonância com o conhecimento semântico das crianças brasileiras que estão nas etapas iniciais do Ensino Fundamental, pois elas não detêm conhecimentos técnicos suficientes a ponto de saber que o arroz é um tipo de erva; já no segundo caso, o problema repousa no alto grau de abrangência da definição, uma vez que o maracujá não é a única planta trepadeira de frutos comestíveis cultivada pelo homem (a exemplo da framboesa, da uva e do kiwi), ocasionando um problema no viés extensional da paráfrase.[6]

Os problemas mencionados acima, nos âmbitos macro e microestrutural[7] das obras analisadas, exemplificam falhas graves que temos identificado nos últimos anos em dicionários escolares brasileiros. Conforme vem sendo apontado por nós e por outros pesquisadores (Farias, 2009; Bugueño Miranda; Farias, 2009; Pires, 2012; Brangel 2016b), esses problemas advêm, principalmente, da falta de diretrizes teóricas e metodológicas voltadas para a elabo-

[5] Na seção 4, discutiremos mais a fundo o Programa Nacional do Livro Didático e as suas contribuições para a Lexicografia Pedagógica brasileira.

[6] Para a discussão sobre intensão e extensão na Lexicografia, ver: BRANGEL, L. M.; CHISHMAN, R. Intensão e extensão na descrição de cenários do futebol. *Fórum Linguístico*, v. 17, n. 4, p. 5416–5428, 2020. DOI: https://doi.org/10.5007/1984-8412.2020.e70832;

[7] Além da macroestrutura e da microestrutura, integram também a megaestrutura de um dicionário semasiológico a medioestrutura (sistema de remissivas) e os textos externos (informações que não compõem a nominata do dicionário, como guia do usuário, tabelas, listas e referências bibliográficas) (Hausmann; Wiegand, 1989).

ração de dicionários escolares. Nessa lacuna verificada, chamamos a atenção para a carência de *corpora* elaborados para fins de estudo e de produção de dicionários.

Atkins e Rundell (2008) abordam a pesquisa, a compilação e o estudo do *corpus* na parte intitulada "pré-lexicografia" de seu manual. Ao posicionarem o estudo com *corpus* antes da compilação da obra lexicográfica em si e dispensarem uma significativa parte do manual para discutir a importância do *corpus* na Lexicografia moderna (chamando a atenção, inclusive, para as especificidades dos *corpora* elaborados para projetos lexicográficos), os autores legitimam a importância de *corpora* na compilação de dicionários. A elaboração de dicionários pedagógicos de excelência, portanto, passa necessariamente pela etapa de elaboração de um *corpus* com propósitos lexicográficos.

O *corpus* elaborado para fins lexicográficos é capaz de orientar a tomada de decisões do lexicógrafo durante a elaboração de diversos componentes da obra, tais como aspectos macroestruturais da obra (a supressão de vocábulos do tipo *topless* e a inserção de vocábulos do tipo *cadeira* na macroestrutura de um dicionário voltado para crianças, por exemplo) e aspectos microestruturais (a seleção das melhores informações semânticas para a redação da definição de palavras como *arroz* e *maracujá*, assegurando definições funcionais ao público infantil, por exemplo). Valendo-se dessas reflexões, o presente estudo propõe uma reflexão sobre como elaborar *corpora* que sirvam para esses propósitos lexicográficos. Para isso, na próxima seção, será apresentado o Oxford Children's Corpus (OCC), um *corpus* compilado pela Universidade de Oxford que serve de base para a elaboração dos dicionários escolares produzidos por essa editora e que contribui de maneira substancial para a excelência das obras da linha Oxford.

3 O Oxford Children Corpus: boas práticas que podem servir de inspiração para a Lexicografia brasileira

A Lexicografia Pedagógica de língua inglesa demonstra, em trabalhos de excelência, a importância de se compilar um *corpus* de forma metodológica, de modo a garantir que as informações oferecidas pelo dicionário sejam adequadas ao seu público-alvo. Um *corpus* que espelhe os textos lidos e produzidos por falantes nativos em idade escolar parece, portanto, ser um elemento pré-lexicográfico necessário para a compilação de um bom dicionário escolar. Por esse motivo, apresentaremos e discutiremos o Oxford Children's Corpus, um *corpus* composto por mais de 30 milhões de *tokens* retirados de textos em língua inglesa escritos para crianças de 5 a 14 anos (Wild; Kilgarriff; Tugwell, 2012). A tabela a seguir apresenta a composição do OCC em três critérios estabelecidos pelos autores: gênero das obras, período histórico e faixa etária.

Tabela 1 – A composição do Oxford Children's Corpus

Gênero	*Tokens*
Ficção	23.139.119
Não ficção	6.755.691
Escritos por crianças	1.421.720
Sem classificação	397.352
Total	31.713.882

Período histórico	*Tokens*
Pré-1900	4.429.132
1900–1964	10.139.421
1965–1999	2.058.857
2000–presente	14.816.110
Sem classificação	270.362
Total	31.713.882
Faixa etária	***Tokens***
1 (5–7 anos)	1.802.762
2 (7–11 anos)	14.003.042
3 (11–14 anos)	7.772.753
Sem classificação	8.135.325
Total	31.713.882

Fonte: Wild, Kilgarriff e Tugwell (2012).

Em relação ao gênero, o OCC é composto majoritariamente por livros de ficção. Aproximadamente 23 milhões de *tokens* (número total de palavras, considerando a repetição no texto) correspondem a obras ficcionais, ou seja, ¾ do *corpus*. Além disso, há textos de não ficção retirados de sites voltados para crianças, como CBBC Newsround e Nickelodeon. O *corpus* também contém livros não ficcionais, como enciclopédias e livros didáticos, além de uma pequena parcela de textos escritos por crianças, retirados de sites onde as crianças postam *reviews*, poemas e histórias. Em relação ao período histórico, o *corpus* é composto majoritariamente por textos escritos no século XXI, ainda que tenham sido incluídas obras clássicas de ficção devido à presença desse gênero no repertório de leitura das crianças. Por fim, os textos estão classificados por faixas etárias (*Key Stages*, ou KS) referentes ao *National Curriculum in England, Wales and Northern Ireland* (Wild; Kilgarriff; Tugwell, 2012, p. 194). As faixas etárias estão marcadas como KS1 (de 5 a 7 anos), KS2 (de 7 a 11 anos) e KS3 (de 11 a 14 anos).

De maneira geral, os compiladores do *corpus* demonstraram uma preocupação em formar uma base de dados com textos utilizados por crianças, o que nos parece extremamente positivo haja vista o propósito lexicográfico por trás da elaboração do *corpus*. Por outro lado, os estudos publicados não especificam os critérios para a elaboração do *corpus*, o que nos leva a concluir que a seleção dos materiais tenha sido feita intuitivamente, ou seja, com base no que os estudiosos acreditavam ser os textos mais utilizados pelas crianças do Reino Unido no momento da elaboração do *corpus*.

Esse último fato pode ser interpretado como um reflexo direto das políticas educacionais do Departamento da Educação do Reino Unido, que não possuem o costume de indicar os títulos dos livros a serem adotados pelas escolas, cabendo a cada instituição organizar sua própria lista de leituras obrigatórias.[8] Diferentemente do Brasil, onde há um programa nacio-

[8] Na discussão sobre a adoção de materiais didáticos em escolas do Reino Unido, a carência de diretrizes é criticada por teóricos da área. Oates (2014), por exemplo, compara o desempenho escolar dos alunos ingleses, que

nal voltado para a avaliação e distribuição de livros didáticos nas escolas, o *National Curriculum* limita-se a estabelecer diretrizes sobre as habilidades que as crianças devem adquirir em cada *Key Stage*. Provavelmente por isso, os lexicógrafos da Oxford recorreram massivamente às obras literárias, uma vez que os clássicos infantis integram a maior parte das leituras presentes nos *syllabus* das escolas (Wild; Kilgarriff; Tugwell, 2012, p. 212–213).

Posteriormente, em uma nova etapa da pesquisa, os pesquisadores ampliaram a parcela do OCC referente aos textos escritos por crianças, visando analisar esses textos de maneira aprofundada (Banerji *et al.*, 2013). Essa nova parte do *corpus* parece útil por possibilitar a compreensão de aspectos lexicais, gramaticais e semânticos da escrita dos consulentes. Por fim, cabe mencionar que o OCC tem um recorte sincrônico, ou seja, está preocupado, sobretudo, com textos contemporâneos, um aspecto essencial que trata do tipo de léxico que as crianças necessitam e com o qual têm contato.

Em nosso entendimento, o OCC consiste em um *corpus* de excelência para os propósitos da Lexicografia Pedagógica. Além de conseguir reunir uma abundância de material textual, o *corpus* também apresenta a proeza de reunir diferentes tipos de texto, oferecendo, assim, uma base de dados exaustiva e representativa do acervo textual disponibilizado para crianças falantes de língua inglesa.

Um exemplo real das contribuições que um *corpus* desse tipo oferece para uma obra lexicográfica pode ser observado no *Oxford Primary Dictionary* (OPD) (2011), um dicionário escolar compilado para crianças de nove anos ou mais que cursam o *Primary School*[9] do sistema educacional do Reino Unido. Conforme avaliado por Brangel (2015, 2016a), o OPD (2011) desponta como uma obra de excelência no âmbito das obras lexicográficas escolares. Análises pontuais do *front matter* da obra e da macro, micro e medioestrutura revelaram que o OPD (2011) consegue fornecer informações muito bem calculadas para seus consulentes. Essas análises serão sintetizadas nos parágrafos que seguem.

No nível da microestrutura, por exemplo, foi possível observar que a obra teve o cuidado de lematizar as palavras utilizadas nas definições lexicográficas, assegurando, assim, que seu consulente encontre, no próprio dicionário, o significado de todas as palavras que integram a microestrutura da obra. Apesar dessa medida assumir importância singular em qualquer obra lexicográfica escolar, nem toda equipe lexicográfica possui esse cuidado no momento da compilação de uma obra, conforme demonstrado por Brangel (2016a) no cotejo de OPD (2011) com dicionários escolares brasileiros. No nível da macroestrutura, além de utilizar o OCC como fonte de dados para a seleção macroestrutural, assegurando confiabilidade aos itens lexicais arrolados pela obra por integrarem textos escritos para e por crianças, cumpre também mencionar a forma de ordenação dos lemas em relação à sua progressão. Nesse caso, a obra optou por oferecer uma estrutura lisa, ou seja, uma progressão vertical, ordenada alfabeticamente e recuada à esquerda (Bugueño Miranda, 2018, p. 20), em vez de uma organização em nicho léxico (progressão vertical e horizontal sem interrupção alfabética) ou ninho léxico (progressão vertical e horizontal com interrupção alfabética), apontando, mais uma vez, para a adequação da obra aos seus consulentes.

não utilizam materiais didáticos desenvolvidos em consonância com o currículo nacional, com o desempenho dos alunos de outros países, como da Finlândia e da cidade-Estado Singapura, que elaboram materiais didáticos coesos e pedagogicamente orientados – além de aprovados pelo Governo – como a forma principal de apoio a professores e alunos em sala de aula, e aponta, nessa comparação, para o baixo desempenho dos alunos ingleses.

[9] No Reino Unido, esse período escolar é destinado a crianças entre 5 e 11 anos de idade.

Com base nas considerações apresentadas, o presente estudo busca discutir possibilidades de adaptação da excelência do OCC para a realidade brasileira. Diante da conjuntura brasileira, apostamos no Programa Nacional do Livro e do Material Didático como um elemento central para uma proposta desse tipo, uma vez que o Brasil possui políticas de livros e materiais didáticos muito mais centralizadas e estabelecidas se comparadas ao Reino Unido, sendo esse um elemento favorável no desenvolvimento de um *corpus* feito de textos escritos para crianças. Abordaremos esse tema na próxima seção.

4 O Programa Nacional do Livro e do Material Didático: bases para um *corpus* em português brasileiro formado a partir de textos voltados para crianças

O mercado de livros didáticos no Brasil sempre movimentou discussões políticas e financeiras na história da educação no país. A preocupação com a avaliação dos livros didáticos e com seus processos de compra e políticas públicas teve início no Estado Novo, com a instituição da primeira Comissão Nacional de Livros Didáticos, que visava a desenvolver e sustentar regras de compra e consumo desses livros no Brasil.

Mais tarde, na década de 60, novas mudanças desencadearam o crescimento e a modificação no modo de elaboração dos livros didáticos no Brasil, dentre as quais se destacam o decréscimo do tempo de permanência do livro na escola e a diversificação na autoria (que passaram do punho de cientistas, intelectuais e professores de instituições universitárias para professores do Ensino Fundamental). Com o crescimento da rede de ensino, e maior ingresso de alunos na rede escolar, houve um aumento da velocidade no processo de industrialização do livro didático, que resultou em mudanças no conteúdo e na didática dos livros comercializados. Para Soares (1996), o crescimento cada vez mais acelerado dos conhecimentos impulsionou consideráveis alterações nos livros didáticos. É também pertinente considerar que, com uma vendagem maior, o livro didático acabou recebendo destaque comercial, sendo, então, considerado um produto vendável a partir do ponto de vista de editores interessados, suplantando os livros de literatura e científicos.

Na década de 80, com o advento do período democrático, problemas existentes nos livros didáticos do Brasil foram discutidos pela Fundação de Assistência ao Estudante (FAE), criada em 1983 por intermédio da Lei nº 7.091. Em 1984, o Ministério da Educação (MEC) assumiu a compra e venda dos livros produzidos por empresas participantes do Programa do Livro Didático. Em 1985, através do Decreto nº 91.542, o Programa do Livro Didático para o Ensino Fundamental do Instituto Nacional do Livro (PLIDEF/INL, 1971–1976) passou a se chamar Programa Nacional do Livro Didático (PNLD) e através desse programa começou a disseminação do livro didático pelo Brasil.

Com a implementação do PNLD, diversas disciplinas escolares foram sendo incluídas e os professores foram responsabilizados pelas escolhas dos livros. No ano de 1996, ocorreram novas mudanças no programa: o governo, antes responsável apenas pela compra e distribuição dos livros, criou um comitê para estudar o valor dos conteúdos e do enfoque pedagógico-metodológico dos livros. Além disso, no mesmo ano, ficou acertado que todos os alunos da 5ª à 8ª série (terminologia utilizada na época) receberiam seus livros didáticos. Na mesma época,

em 1997, a FAE deixou de existir e o PNLD passou a ser orientado pelo Fundo Nacional do Desenvolvimento para a Educação (FNDE), com a garantia de recursos para manutenção.

Recentemente, o nome do programa passou por uma pequena alteração para incluir novas mudanças. Assim, a palavra *material* passou a integrar o título, culminando no novo nome Programa Nacional do Livro e do Material Didático, sem alterações na sigla PNLD. Essa mudança reflete uma importante alteração no programa, que passou a avaliar e distribuir outros tipos de materiais nos últimos anos, como livros de ficção e dicionários.

Sobre a entrada dos dicionários no programa, cabe, aqui, uma pequena discussão devido à sua importância para a Lexicografia Pedagógica brasileira. O movimento pela avaliação de dicionários por instâncias do governo começou a ganhar corpo nos primeiros anos do século XXI, fortemente impulsionado pelo descontentamento de professores e educadores com os dicionários existentes no mercado. De fato, naquela época, pouco se discutia sobre as necessidades dos consulentes em fase escolar, o que resultava na adoção de obras do tipo mini para esses alunos. Na época, era comum que a mesma obra lexicográfica acompanhasse o aluno durante toda a sua trajetória escolar, ou seja, diferentemente dos livros didáticos, alternados a cada novo ano escolar, os dicionários de língua portuguesa permaneciam os mesmos desde o período de alfabetização das crianças até a conclusão do Ensino Médio pelo jovem quase adulto.

No início do século XXI, o Ministério da Educação lançou, pela primeira vez na história do país, um programa voltado exclusivamente para a avaliação e a distribuição de dicionários escolares, ao qual deu o título de PNLD Dicionários. Na época, foram formadas bancas avaliadoras compostas por professores que, com base em diretrizes promulgadas pelo MEC, avaliaram obras escolares submetidas à apreciação por editoras. As obras, se aprovadas, eram distribuídas em escolas da rede pública de todo o país, seguindo o mesmo percurso dos livros didáticos. Além do advento da avaliação, que obrigou as editoras a se preocuparem mais com a qualidade de seus dicionários, o programa também estabeleceu classificações de obras dicionarísticas, com o intuito de estabelecer diferentes tipos de dicionários para diferentes etapas da educação básica. Assim, tivemos, pela primeira vez, um grupo de dicionários voltado para os anos de alfabetização (dicionários tipo 1), outro voltado para a fase de consolidação da escrita (dicionários tipo 2), outro voltado para os últimos anos do Ensino Fundamental (dicionários tipo 3) e outro para o Ensino Médio (dicionários tipo 4). Além dessa classificação, o programa publicou um material para professores (Brasil, 2012) com o intuito de orientar esses profissionais quanto à natureza dos dicionários escolares e ao seu uso em sala de aula.

O PNLD Dicionários lançou dois editais, o primeiro em 2006 e o segundo em 2012, e embora apresentasse muitas falhas no âmbito de seu planejamento, especialmente no que diz respeito à caracterização dos dicionários, representou um grande avanço para o Brasil no âmbito da Lexicografia Pedagógica. Por volta de 2016 o programa começou a ser descontinuado, de modo que, atualmente, não há novas movimentações na página oficial do programa[10] e tanto as avaliações como a distribuição dos dicionários estão estagnadas.

Com a interrupção do PNLD Dicionários, o PNLD Didático (que avalia e distribui livros didáticos) e o PNLD Literário (que avalia e distribui livros de ficção) assumem, atualmente, o importante papel de avaliar e distribuir os principais materiais escolares que transitam pelas escolas brasileiras. Deste modo, por intermédio dos acervos aprovados pelo PNLD, um pesquisador pode ter acesso aos textos utilizados pelos alunos das escolas brasileiras. O PNLD,

[10] Ver: BRASIL. Ministério da Educação. *PNLD Dicionários.* Brasília: MEC, 2012. Disponível em: http://portal.mec.gov.br/pnld/dicionarios. Acesso em: 2 fev. 2024.

portanto, permite acesso ao "*habitat* natural" em que os alunos são inseridos durante suas atividades textuais e, por isso, representa uma importante fonte de dados para a compilação de *corpora*. Na próxima seção, discutiremos essa possibilidade.

5 Parâmetros para a compilação de um *corpus* de textos em português brasileiro voltados para crianças

Em Brangel (2017), foram traçadas as características essenciais de um dicionário escolar voltado para crianças do 4º e 5º ano do Ensino Fundamental, ao qual foi dado o nome de dicionário intermediário. Com base nos parâmetros lexicográficos propostos por Bugueño Miranda e Farias (2008, 2009), que definem os dicionários pedagógicos em seus traços fundamentais a partir de três axiomas (o enquadramento taxonômico, o perfil do usuário e a função da obra), Brangel (2017) chegou a uma série de características que asseguram o desenho de um dicionário intermediário. Para fins do presente estudo, aproximaremos os postulados de Brangel (2017), sobre o enquadramento taxonômico do dicionário intermediário, aos postulados de Sardinha (2000), sobre classificação de *corpora*, com o intuito de delinear as características fundamentais que um *corpus* deve apresentar para servir de base para a compilação de um dicionário escolar do tipo intermediário.

Antes de adentrarmos na relação entre o dicionário pedagógico e o seu *corpus* correspondente, é importante aprofundar algumas considerações sobre o enquadramento taxonômico de obras lexicográficas. O ato de enquadrar taxonomicamente um dicionário consiste em classificá-lo de acordo com determinados critérios. Na literatura, importantes autores forneceram propostas para a classificação de obras lexicográficas, a exemplo de Biderman (1998), Landau (2001), Swanepoel (2003), Welker (2004) e Atkins e Rundell (2008). Ainda que distintas entre si, todas essas propostas apresentam como característica comum a tentativa de propor um modelo classificatório de dicionários baseado em critérios fenomenológicos, funcionais ou linguísticos (Bugueño Miranda; Farias, 2009).

Os critérios fenomenológicos dizem respeito às percepções físicas da obra e correspondem às classificações que consideram características como o tamanho, o formato e número de entrada de uma obra. O segundo critério diz respeito às classificações funcionais, que consideram o uso efetivo do dicionário. O terceiro critério, linguístico, considera fatores como o tipo de informação fornecida (linguística ou enciclopédica), o número de línguas (dicionário monolíngue, bilíngue ou multilíngue), a perspectiva do ato da comunicação (perspectiva onomasiológica ou semasiológica) ou uma concepção sinsistêmica ou diassistêmica da linguagem (vocabulário de uso geral ou marcado diassistemicamente) (Farias, 2009, p. 37). Importante mencionar que, para os propósitos do presente estudo, os critérios linguísticos diferenciam-se dos critérios fenomenológicos e funcionais por atenderem às necessidades dos lexicógrafos. Assim, de acordo com Farias (2009), enquanto os critérios fenomenológicos e funcionais estão relacionados aos potenciais consulentes do dicionário, os critérios linguísticos auxiliam os lexicógrafos nas tarefas de concepção, redação e desenho da obra.

Com base nesses princípios, Brangel (2017, p. 89–90) propõe, a partir de uma classificação do tipo linguística, que o dicionário intermediário seja uma obra de cunho linguístico, monolíngue, semasiológico, geral, seletivo e sincrônico, sendo:

a) linguístico porque oferece informações relativas à língua;

b) monolíngue porque se destina ao aprendizado formal da língua materna;

c) semasiológico porque o ato da consulta parte da lista das palavras arroladas;

d) geral porque a seleção do léxico parte do vocabulário geral do português;

e) seletivo porque não contempla vocabulário desusado ou de baixa frequência;

f) sincrônico porque contempla palavras e significados do português contemporâneo.

Esse conjunto de traços é relevante ao lexicógrafo por auxiliá-lo em, pelo menos, duas importantes tarefas: a de compilar e a de avaliar a obra lexicográfica. Conhecer os traços que definem (e classificam) o dicionário é relevante para a compilação da obra, pois permite que se estabeleça com precisão as informações que devem estar disponíveis ao consulente, bem como a sua forma de apresentação. É relevante também para a avaliação, pois permite que se estabeleçam parâmetros para julgar a pertinência das informações contidas no dicionário. Assim, ao apresentar o traço *seletivo* em oposição ao traço *exaustivo*, depreende-se por esse traço que a macroestrutura do dicionário intermediário deve ser composta por itens lexicais que contemplem o vocabulário em uso de seus consulentes, como os itens lexicais *mesa* e *cadeira*, e que a seleção da nominata deve preterir itens lexicais inapropriados para o público-alvo, como a palavra *topless*. Caso apresentasse o traço exaustivo, essas restrições não seriam necessárias.

Uma vez estabelecida a importância da classificação taxonômica de uma obra lexicográfica, propomos que essa discussão seja estendida ao âmbito da pesquisa com *corpus* e chamamos a atenção para a correspondência direta que um *corpus* elaborado para fins de produção dicionarística deve estabelecer com o produto final almejado. Assim, se a pesquisa com *corpus* tem oferecido grandes contribuições para a Lexicografia no sentido de auxiliar a produção de dicionários, a exemplo de Banerji *et al*. (2013) e Wild, Kilgarriff e Tugwell (2012), é pertinente haver uma reflexão sobre as características que um *corpus* deve apresentar para que seja uma fonte de dados confiável para a compilação de um dicionário.

Sardinha (2000) organiza os principais critérios utilizados na literatura para se definir o conteúdo e o propósito de um *corpus* e adiciona alguns critérios complementares para chegar a uma proposta de tipologia de *corpus*. O quadro abaixo ilustra os critérios arrolados por Sardinha (2000):

Tabela 2 – Tipologia de *corpus*

Modo	Falado: Composto de porções de fala transcritas. Escrito: Composto de textos escritos, impressos ou não.
Tempo	Sincrônico: Compreende um único período de tempo. Diacrônico: Compreende vários períodos de tempo. Contemporâneo: Representa o período de tempo corrente. Histórico: Representa um período de tempo passado.

Seleção	De amostragem (*sample corpus*): Composto por porções de textos ou de variedades textuais, planejado para ser uma amostra finita da linguagem como um todo. Monitor: A composição é reciclada para refletir o estado atual de uma língua. Opõe-se a *corpora* de amostragem. Dinâmico ou orgânico: O crescimento e diminuição são permitidos, qualifica o *corpus* monitor. Estático: Oposto de dinâmico, caracteriza o *corpus* de amostragem. Equilibrado (*balanced*): Os componentes (gêneros, textos etc.) são distribuídos em quantidades semelhantes (por exemplo, mesmo número de textos por gênero).
Conteúdo	Especializado: Os textos são de tipos específicos (normalmente gêneros ou registros definidos). Regional ou dialetal: Os textos são provenientes de uma ou mais variedades sociolinguísticas específicas. Multilíngue: Inclui idiomas diferentes.
Autoria	De aprendiz: Os autores dos textos não são falantes nativos. De língua nativa: Os autores são falantes nativos.
Disposição interna	Paralelo: Os textos são comparáveis (por exemplo, original e tradução). Alinhado: As traduções aparecem abaixo de cada linha do original.
Finalidade	De estudo: O *corpus* que se pretende descrever. De referência: Usado para fins de contraste com o *corpus* de estudo. De treinamento ou teste: Construído para permitir o desenvolvimento de aplicações e ferramentas de análise.
Outros critérios	Pluralidade de autoria: Os textos foram produzidos por um autor apenas ou mais? Origem da autoria: Os textos foram produzidos por falantes nativos ou não nativos? Meio: Os textos foram escritos ou falados? Integralidade: Os elementos do *corpus* são textos integrais ou fragmentos? Especificidade: O *corpus* é composto de tipos variados de texto ou textos específicos? Dialeto: As variedades presentes no *corpus* são do tipo "padrão" ou regionais / dialetais? Equilíbrio: As variedades do *corpus* são distribuídas equitativamente ou não? Fechamento: É permitida a inclusão de conteúdos novos ou não? Renovação: O conteúdo do *corpus* reflete um período definitivo de tempo ou se renova? Temporalidade: O *corpus* é planejado para retratar períodos históricos de tempo ou não? Plurilinguismo: O *corpus* possui só textos originais ou também as traduções destes textos para uma ou mais línguas?

Fonte: Sardinha (2000).

Com o intuito de cruzar os postulados de Sardinha (2000) com os de Brangel (2017), selecionamos alguns critérios de classificação de *corpora* para os fins do presente estudo. Neste momento, é importante retomar que, uma vez que o *corpus* precisa estar em consonância com o produto lexicográfico para o qual servirá de fonte de dados, vislumbramos em nossa proposta de *corpus* uma base de dados formada a partir dos acervos indicados pelo PNLD Literário e pelo PNLD Didático destinado às etapas escolares correspondentes ao 4º e ao 5º ano do Ensino Fundamental. Esse recorte ampara-se no fato de que o dicionário intermediário discutido por Brangel (2017) tem como público-alvo crianças do 4º e 5º ano do Ensino Fundamental. Assim, com base no cruzamento dos postulados de Brangel (2017), sobre o enquadramento taxonômico do dicionário intermediário, e nos postulados de Sardinha (2000), sobre os critérios para a classificação de *corpora*, propomos que o *corpus* vislumbrado no presente estudo (*corpus* de textos em português brasileiro voltados para crianças cuja função seja servir de fonte de dados para a elaboração de um dicionário intermediário) deve apresentar as seguintes características:[11]

[11] Os critérios "conteúdo" e "disposição interna" foram retirados de nossa classificação porque não se aplicam à discussão sobre *corpora* formados por textos voltados para crianças. Os critérios "origem da autoria" e "fechamento" também foram suprimidos porque retomam os critérios "autoria" e "seleção", respectivamente.

a) Modo: escrito, porque deve ser composto por textos escritos para crianças. Tanto os livros didáticos como os livros de ficção indicados pelo PNLD são escritos para crianças, observando faixas etárias e fases de escolarização específicas. Os livros são cuidadosamente catalogados para atenderem um público-alvo bem delimitado. Diferentemente do OCC, que inseriu em sua base de dados também transcrições de textos falados, nossa proposta inicial é que um *corpus* do português brasileiro de textos escritos para crianças seja composto apenas por textos escritos. Primeiramente, porque a transcrição de textos falados envolve uma etapa mais elaborada de pesquisa, que não temos condições de desenvolver no momento, e, em segundo lugar, porque os acervos do PNLD parecem fornecer bases confiáveis para um *corpus* desse tipo, mesmo que não entrem no âmbito de textos transcritos;

b) Tempo: contemporâneo, porque representa o período atual. O PNLD possui avaliação periódica, sendo que, a cada avaliação, novas coleções de livros podem ser integradas ou descartadas do programa. Dessa forma, ao assegurar o fluxo de material a cada nova avaliação, o programa garante a atualização de seus materiais;

c) Seleção: monitor, porque é constantemente atualizado; dinâmico, porque permite o acréscimo de novos materiais. Nesse caso, as atualizações constantes do PNLD, citadas no item acima, também servem de suporte para garantir essas qualidades ao *corpus*;

d) Autoria: de língua nativa, porque os textos são escritos por falantes nativos do português brasileiro. Ainda que o PNLD Literário trabalhe com traduções e adaptações de obras literárias, as versões dos acervos são sempre cuidadosamente elaboradas por autores brasileiros preocupados com o leitor em fase escolar;

e) Finalidade: de estudo, porque o *corpus* formado a partir dos acervos servirá de base para a pesquisa que se planeja desenvolver. Em nossa proposta, um *corpus* elaborado a partir dos acervos do PNLD servirá de base para a pesquisa, não para comparação com outros *corpora*, por exemplo;

f) Pluralidade de autoria: de autoria múltipla, porque o *corpus* é formado por textos de múltiplos autores. O PNLD se preocupa em fornecer uma vasta opção de livros de ficção, bem como uma vasta possibilidade de coleções de livros didáticos, para os professores terem a liberdade de escolher os livros que melhor atendem suas demandas. Essa preocupação assegura que múltiplos autores integrem a lista de acervos;

g) Meio: textos escritos, porque os textos voltados para o 4º e 5º ano do Ensino Fundamental foram escritos para serem lidos, não para serem falados, como roteiros ou palestras, por exemplo;

h) Integralidade: integral, porque os textos são dispostos integralmente no *corpus* (livros completos), não de maneira fragmentada;

i) Especificidade: de tipo variado porque o *corpus* é composto por tipos variados de textos. O PNLD fornece uma variada gama de gêneros textuais em seus acervos, tanto nos conteúdos dos livros didáticos como também nos livros de

ficção. São explorados diversos tipos de textos (letras de música, reportagens, crônicas, contos, histórias em quadrinho etc.) e há a recorrente preocupação de que os textos sejam adequados ao período escolar para o qual são indicados;

j) Dialeto: padrão, porque as variedades presentes no *corpus* não marcam um dialeto ou regionalismo específico;

k) Equilíbrio: desigual, porque as variedades do *corpus* não são distribuídas equitativamente (a quantidade de livros do PNLD Literário é inferior ao do PNLD Didático);

l) Renovação: de renovação, porque o conteúdo do *corpus* se renova a cada edição do PNLD. As constantes edições do programa asseguram a inclusão e exclusão de materiais aos acervos;

m) Temporalidade: temporal, porque o *corpus* é planejado para retratar o período histórico atual. Os textos do PNLD retratam a língua atual em uso, não há espaço para formas arcaicas da língua;

n) Plurilinguismo: monolíngue, porque o *corpus* possui apenas textos em português brasileiro. O PNLD não trabalha com livros em língua estrangeira (com exceção dos livros didáticos para aulas de língua estrangeira, mas esses, obviamente, não integrariam um *corpus* de textos em português voltados para crianças).

Os critérios elencados acima, para a composição de um *corpus* de textos em português brasileiro escritos para crianças, evidenciam a possibilidade de aproximação dos acervos do PNLD ao estudo e compilação de *corpus*. Além de fornecerem uma representação confiável do material textual que circula pelas salas de aula brasileiras, os acervos do PNLD se encaixam de maneira muito harmoniosa nos critérios acima promulgados para a elaboração de um *corpus* de textos brasileiros nos moldes do OCC.

6 Considerações finais

Neste trabalho, procuramos transpor para a realidade brasileira importantes discussões da Lexicografia inglesa sobre o estudo de *corpus* com fins lexicográficos. Ao longo de nossa pesquisa, descobrimos que as políticas de livros didáticos no Brasil possuem longa tradição e consolidação, diferentemente do cenário do Reino Unido, e que essa particularidade das políticas educacionais brasileiras pode ser de grande valor para a elaboração de um *corpus* do tipo do OCC.

Enquanto o PNLD Didático indica os livros didáticos utilizados nas disciplinas que compõem a grade curricular escolar (como Língua Portuguesa, História, Matemática), o PNLD Literário fornece a lista de livros de ficção utilizados nas atividades de leituras dos alunos, também respeitando as diferentes etapas de ensino. Atualmente, os acervos do PNLD Didático e do PNLD Literário passam por uma fase de avaliação por especialistas, o que assegura a adequação das obras ao público-alvo, além de serem distribuídos em escolas da rede pública de todo o país, o que confere aos textos dos livros do PNLD o *status* de amostragem real e confiável dos textos utilizados pelos alunos brasileiros nas atividades de sala de aula.

Assim, chegamos à conclusão de que um *corpus* composto pelo material fornecido pelo PNLD seria útil para a compilação de dicionários, podendo corroborar para o desenho dos

mais diversos componentes das obras. No nível macroestrutural, por exemplo, as palavras e expressões registradas no *corpus* teriam o potencial de espelhar tanto a macroestrutura quantitativa da obra (densidade macroestrutural, ou seja, quantidade de lemas a serem arrolados) como também a macroestrutura qualitativa (tipos de lemas a serem arrolados). Já no nível microestrutural, o *corpus* teria o potencial de elucidar significados e formas de uso das palavras arroladas pelo dicionário, auxiliando o lexicógrafo a discriminar o(s) significado(s) a ser(em) descrito(s) em cada acepção de um verbete, bem como o tipo e a quantidade de informações a serem oferecidos nas definições. No caso da apresentação de exemplos nos verbetes, o *corpus* poderia servir também de fonte direta para a coleta de dados para a elaboração desse segmento informativo.

Como o propósito do presente estudo é fixar parâmetros para a compilação de um *corpus* de textos escritos para crianças, essas sugestões ainda não foram testadas de maneira concreta, mas assentam suas bases nos estudos com *corpus* em língua inglesa de Banerji *et al*. (2013) e Wild, Kilgarriff e Tugwell (2012) e em pesquisas brasileiras da Lexicografia Pedagógica (Farias, 2009; Bugueño Miranda; Farias, 2009; Pires, 2012; Brangel, 2015, 2016b, 2017).

Com tais parâmetros em vista, foi iniciada, no segundo semestre de 2022, a compilação de um *corpus* baseado nas discussões aqui apresentadas. Ainda que a análise do material tenha se revelado muito satisfatória, nosso principal obstáculo foi e tem sido o acesso aos livros didáticos, distribuídos somente em formato impresso e diretamente para as escolas da rede pública. Por estarem protegidas por leis de direitos autorais, as obras não são disponibilizadas gratuitamente na internet, cabendo-nos contar com as doações de algumas editoras que se disponibilizaram a nos enviar suas obras via correios, sem custos, ou com o empréstimo pontual de obras por parte de algumas bibliotecas de escolas públicas.

Atualmente,[12] o acervo do PNLD Literário conta com 180 obras de ficção indicadas para os alunos dos 4º e 5º anos do Ensino Fundamental; já o PNLD Didático organiza seus acervos em categoria 1 (por área): Língua Portuguesa, Arte e Educação Física, Matemática, Ciências da Natureza e Ciências Humanas, e categoria 2 (por componente): Arte, Educação Física, Geografia, História e Projetos Integradores. Somadas as coleções de livros aprovadas pela mais recente edição do PNLD Didático, nas categorias 1 e 2, para os anos iniciais do Ensino Fundamental, contamos com um total de 111 coleções de livros (Brasil, 2023). Assim, o extenso volume de material tem dificultado bastante a realização da meta inicial da pesquisa, que era a de compilar um *corpus* com todas as obras aprovadas pelo PNLD (Didático e Literário) para o segmento do Ensino Fundamental compreendido entre o 4º e o 5º ano. O formato de distribuição das obras (exclusivamente impresso) também tem sido um obstáculo, pois nos obriga a dispensar uma etapa da pesquisa para a digitalização dos materiais.

Pelos motivos citados acima, nosso *corpus* ainda apresenta um volume de dados bastante modesto. Ainda assim, esperamos que futuramente consigamos utilizar o material em fase de compilação para descobrir aspectos e peculiaridades dos textos escritos para crianças brasileiras do 4º e 5º ano do Ensino Fundamental e, com isso, sugerir importantes aprimoramentos aos dicionários escolares voltados para esse público. Os parâmetros aqui fixados representam, portanto, um primeiro passo em direção a pesquisas mais robustas de compilação e análise de *corpus*. Esperamos que essas e outras pesquisas, a serem divulgadas pelo nosso grupo futuramente, encorajem novos estudiosos a voltarem a sua atenção para essa nova possibilidade de contribuição do Programa Nacional do Livro e do Material Didático para a educação brasileira.

[12] Dados do ano de 2023.

Agradecimentos

Agradecemos à Escola Municipal de Ensino Fundamental Fioravante Webber, por emprestar itens de seu acervo, às editoras Globo Livros e Salamandra, por doarem materiais para a nossa pesquisa, e à Katherine, por nos ajudar na digitalização dos livros. Agradecemos também à Yasmin Ribas pela revisão atenta e cuidadosa do presente artigo.

Referências

ATKINS, B. T. S.; RUNDELL, M. *The Oxford Guide to Practical Lexicography.* Oxford: Oxford University Press, 2008.

BANERJI, N.; GUPTA, V.; KILGARRIFF, A.; TUGWELL, D. Oxford Children's Corpus: a Corpus of Children's Writing, Reading, and Education. *In*: Hardie, A.; Love, R. (Ed.). *Corpus Linguistics 2013*: Abstract Book. Lancaster: University Centre for Computer Corpus Research on Language, 2013. p. 315–318.

BIDERMAN, M. T. C. Os dicionários da contemporaneidade: arquitetura, métodos e técnicas. *In*: OLIVEIRA, A. M. P.; ISQUERDO, A. N. (Ed.). *As ciências do léxico*: lexicologia, lexicografia, terminologia. Campo Grande: Editora UFMS, 1998. p. 129–142.

BRANGEL, L. M. A lexicografia pedagógica no Reino Unido e no Brasil: subsídios da produção britânica para o aprimoramento das obras nacionais. *Caminhos em Linguística Aplicada*, v. 15, n. 2, p. 125–142, 2016a. Disponível em: http://periodicos.unitau.br/ojs/index.php/caminhoslinguistica/article/view/2068. Acesso em: 1 jun. 2023.

BRANGEL, L. M. Considerações sobre o Programa Constante de Informações de dicionários escolares de língua portuguesa voltados para o público infantil. *Caligrama*: Revista de Estudos Românicos, v. 18, n. 2, p. 155–177, 2013b. DOI: http://dx.doi.org/10.17851/2238-3824.18.2.155-177.

BRANGEL, L. M. Contribuições para a lexicografia pedagógica a partir de dados extraídos de livros didáticos. *Estudos da Língua(gem)*, v. 11, n. 2, p. 43–61, 2013a.

BRANGEL, L. M. Proposta de um dicionário intermediário de língua portuguesa para uso em sala de aula: características, público-alvo e função. *(Con)Textos Linguísticos*, v. 11, n. 20, p. 85–105, 2017. Disponível em: https://periodicos.ufes.br/contextoslinguisticos/article/view/16867. Acesso em: 1 jun. 2023.

BRANGEL, L. M. *Proposta teórico-metodológica para a geração de paráfrases explanatórias em dicionários voltados para crianças*: uma abordagem cognitiva. 2016. 209f. Tese (Doutorado em Letras) – Instituto de Letras, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2016b.

BRANGEL, L. M. Review of Oxford Primary Dictionary. *Eurasian Journal of Applied Linguistics*, v. 1, n. 2, p. 99–102, 2015. DOI: https://doi.org/10.32601/ejal.460626.

BRASIL. Ministério da Educação. *Guia digital PNLD 2023*: Obras didáticas – objeto 1. Código das coleções. Brasília: MEC/SEB/FNDE, 2023. Disponível em: https://pnld.nees.ufal.br/pnld_2023_anos_iniciais_ensino_fundamental_obras_didaticas/pnld_2023_anos_iniciais_ensino_fundamental_obras_didaticas_codigo_obras. Acesso em: 3 maio 2023.

BRASIL. Ministério da Educação. Secretaria de Educação Básica. *Com direito à palavra*: dicionários em sala de aula. Brasília: Ministério da Educação, Secretaria de Educação Básica, 2012.

BUGUEÑO MIRANDA, F. V. A estruturação de um dicionário. *In*: Bugueño Miranda, F. V.; Borba, L. C. de. (Org.). *Manual de (Meta)Lexicografia*. Goiás: Espaço Acadêmico, 2018. p. 14–29.

BUGUEÑO MIRANDA, F. V. Balanço e perspectivas da lexicografia. *Cadernos de Tradução*, Florianópolis, v. 2, n. 32, p. 15–37, 2013. DOI: https://doi.org/10.5007/2175-7968.2013v2n32p15.

BUGUEÑO MIRANDA, F. V. Da classificação de obras lexicográficas e seus problemas: proposta de uma taxonomia. *Alfa Revista de Linguística*, v. 58, n. 1, p. 215–23, 2014. DOI: https://doi.org/10.1590/S1981-57942014000100009.

BUGUEÑO MIRANDA, F. V. O que é macroestrutura no dicionário de língua? *In*: ALVES, I. M. A.; ISQUERDO, A. N. (Org.). *As ciências do léxico*: Lexicologia, lexicografia e terminologia. Campo Grande: Humanitas, 2007. p. 261–272. v. III.

BUGUEÑO MIRANDA, F. V.; BORBA, L. C. de. (Org.) *Manual de (Meta)Lexicografia*. Goiás: Espaço Acadêmico, 2018. 158 p.

BUGUEÑO MIRANDA, F. V.; FARIAS, V. S. Desenho da macroestrutura de um dicionário escolar de língua portuguesa. *In*: BEVILACQUA, C. R.; HUMBLÉ, P.; XATARA, C. M. (Org.). *Lexicografia Pedagógica*: pesquisas e perspectivas. Florianópolis: UFSC / NUT, 2008. p. 129–167.

BUGUEÑO MIRANDA, F. V.; FARIAS, V. S. Panorama crítico dos dicionários escolares brasileiros. *Lusorama*, Frankfurt am Main, v. 77/78, p. 29–78, 2009.

CIGNONI, L.; LANZETTA, E.; PECCHIA, L.; TURRINI, G. Children's Aid to a Children's Dictionary. *In*: Gellerstam, M. *et al.* (Ed.). *EURALEX '96 Proceedings I–II*. Papers Submitted to the Seventh EURALEX International Congress on Lexicography in Göteborg, Sweden. Gothenburg: Department of Swedish, Göteborg University, 1996. p. 659–666.

FARIAS, V. F. *Desenho de um dicionário escolar de língua portuguesa*. 2009. 285f. Dissertação (Mestrado em Letras) – Instituto de Letras, UFRGS, Porto Alegre, 2009.

FARIAS, V. F. *Sobre a definição lexicográfica e seus problemas*: fundamentos para uma teoria geral dos mecanismos explanatórios em dicionários semasiológicos. 2013. 398f. Tese (Doutorado em Letras) – Instituto de Letras, UFRGS, Porto Alegre, 2013.

HAENSCH, G.; WOLF, L.; ETTINGER, S.; WERNER, R. *La lexicografía*. De la lingüística teórica a la lexicografia práctica. Madrid: Gredos, 1982.

HARTMANN, R. R. K.; JAMES, G. *Dictionary of Lexicography*. London/ New York: Routledge, 2002.

HAUSMANN, F. J.; WIEGAND, H. E. Component Parts and Structures of General Monolingual Dictionaries: a Survey. *In*: HAUSMANN, F. J.; REICHMANN, O.; WIEGAND, H. E.; ZGUSTA, L. (Ed.). *Wörterbücher, dictionaries, dictionnaires*. Ein internationales Handbuch zur Lexikographie. Berlin/New York: Walter de Gruyter, 1989. v. 1, p. 328–360.

KILGARRIFF, A. Putting the *corpus* into the dictionary. *In*: OOI, V. B. Y.; PAKIR, A.; TALIB, I. S.; TAN, P. K. W. (Ed.). *Perspectives in Lexicography*: Asia and Beyond. Israel: K Dictionaries Ltd, 2009. p. 239–47.

LANDAU, S. I. *Dictionaries*: the Art and Craft of Lexicography. 2. ed. Cambridge: Cambridge University Press, 2001.

MATTOS, G. *Dicionário Júnior da língua portuguesa*. 3. ed. São Paulo: FTD, 2005.

MOON, R. Sinclair, lexicography, and the Cobuild Project: the Application of Theory. *International Journal of Corpus Linguistics*, v. 12, n. 2, p. 159–181, 2007. DOI: http://dx.doi.org/10.1075/ijcl.12.2.05moo.

MOON, R. What can a *corpus* tell us about lexis? *In*: MCCARTHY, M.; O'KEEFFE, A. (Ed.). *The Routledge Handbook of Corpus Linguistics*. London: Routledge, 2010. p. 197–211.

OATES, T. *Why Textbooks Count*: A Policy Paper. Cambridge: Cambridge Assessment, 2014. Disponível em: http://www.cambridgeassessment.org.uk/Images/181744-why-textbooks-count-tim-oates.pdf. Acesso em: 22 maio 2023.

OLIVEIRA, A. F. S. de. Taxonomia de dicionários monolíngues de inglês para falantes não nativos. *Signo*, v. 35, p. 224–241, 2010. Edição especial. DOI: https://doi.org/10.17058/signo.v35i0.1429.

Oxford Primary Dictionary. New York: Oxford University Press, 2011.

PIRES, J. A. *Contribuições para dicionários escolares destinados às séries iniciais*. 2012. 150f. Dissertação (Mestrado em Letras) – Instituto de Letras, UFRGS, Porto Alegre, 2012.

SARDINHA, T. B. Linguística de Corpus: histórico e problemática. *DELTA: Documentação de Estudos em Linguística Teórica e Aplicada*, v. 16, n. 2, p. 323–367, 2000. DOI: https://doi.org/10.1590/S0102-44502000000200005.

SAVAIRA. *Saraiva Júnior*: dicionário da língua portuguesa ilustrado. 3. ed. São Paulo: Saraiva, 2010.

SOARES, M. B. Um olhar sobre o livro didático. *Presença pedagógica*, v. 2, n. 12, p. 52–64, 1996.

SWANEPOEL, P. Dictionary Typologies: a Pragmatic Approach. *In*: STERKENBURG, P. *A Practical Guide to Lexicography*. Amsterdam/Philadelphia: John Benjamins Publishing Company, 2003. p. 44–69.

WELKER, H. A. *Dicionários*. Uma pequena introdução à lexicografia. 2. ed. Brasília: Thesaurus, 2004.

WILD, K.; KILGARRIFF, A.; TUGWELL, D. The Oxford Children's Corpus: Using a Children's Corpus in Lexicography. *International Journal of Lexicography*, v. 26, n. 2, p. 190–218, 2012. DOI: https://doi.org/10.1093/ijl/ecs017.