Rev. Estud. Ling., Belo Horizonte, v. 29, n. 3, p. 2041-2078, 2021

Compilação, reciclagem e padronização de um *Corpus* Colaborativo de Linguística: percursos metodológicos

Compilation, recycling and standardization in a Collaborative Corpus of Linguistics: methodological approaches

Guilherme Fromm
Universidade Federal de Uberlândia (UFU), Uberlândia, Minas Gerais / Brasil
guifromm@ufu.br
http://orcid.org/0000-0001-5654-0135

Márcio Issamu Yamamoto
Universidade Federal de Jataí (UFJ), Jataí, Goiás / Brasil
marcioiy@ufg.br
http://orcid.org/0000-0001-9792-8187

Resumo: O objetivo deste texto é relatar uma experiência de trabalho com um *corpus* colaborativo, desenvolvida nos últimos 10 anos (2010-2020) com alunos de diversas turmas de graduação, pós-graduação e alunos de Iniciação Científica na Universidade Federal de Uberlândia (já parcialmente descrita por Fromm, 2013, e Fromm e Yamamoto, 2013). O trabalho parte da metodologia de elaboração do *corpus* (incluindo seu histórico) como um passo para um tipo de pesquisa específica (terminográfica), passa por uma análise para indicar e solucionar problemas de compilação e termina na sua adequação e padronização para reuso. O resultado é um *corpus* robusto, bem balanceado, bilíngue (inglês/português) e que poderá ser usado em inúmeras outras pesquisas na área de Linguística.

Palavras-chave: Linguística; Linguística de *Corpus*; *Corpus* colaborativo; Árvore de Domínio; Terminografia.

Abstract: This text describes a work experience with a collaborative corpus, developed during the last 10 years (2010-2020) with students in several undergraduate and graduate classes and students working on scientific undergraduate research projects at Universidade Federal de Uberlândia (already partially described by Fromm, 2013, and Fromm and Yamamoto, 2013). The work starts from the corpus elaboration methodology

eISSN: 2237-2083
DOI: 10.17851/2237-2083.29.3.2041-2078

(including its history) as a step towards a specific type of research (terminographic), goes through an analysis to point out and solve compilation problems and ends in its adequacy and standardization for reuse. The result is a robust, well-balanced, bilingual (English/Portuguese) corpus that can be used in numerous other studies in the area of Linguistics.

Keywords: Linguistics; Corpus Linguistics; Collaborative corpus; Domain Tree; Terminography.

Recebido em 19 de outubro de 2020

Aceito em 24 de fevereiro de 2021

1 Apresentação

As experiências com a Linguística de *Corpus* (doravante LC) fazem parte de todo o percurso na pós-graduação dos autores e dirigem nossas pesquisas há mais de 15 anos. Pretendemos, com este artigo, apresentar dois trabalhos acadêmicos sequenciais compartilhados pelos pesquisadores na elaboração e na reelaboração de um mesmo *corpus* num período de dez anos, dando maior ênfase aos processos metodológicos aplicados aos *corpora* compilados na pesquisa: (1) projeto de elaboração de *corpus* acadêmico bilíngue na área de Linguística e (2) projeto de reelaboração e construção de um vocabulário bilíngue de Linguística (YAMAMOTO, 2018).

Partimos do pressuposto de que uma pesquisa de caráter empírico, descritivo, estatístico e probabilístico que envolva *corpora* deva considerar que o pesquisador vá trabalhar com a metodologia da LC.[1] Há duas formas básicas para o trabalho com *corpus*: compilar o próprio *corpus* (oral ou escrito) ou usar *corpora* já prontos (normalmente disponíveis na internet). No primeiro caso, os *corpora*, elaborados sob medida para pesquisas, são geralmente pequenos, devido ao trabalho envolvido e à restrição de tempo; no segundo caso, nem sempre o conteúdo desses *corpora* é totalmente aproveitável em pesquisas propostas.

Os *corpora* bilíngues a serem aqui descritos foram compilados a várias mãos. Eles começaram como um mero exercício em sala de aula, com *corpora* descartáveis, evoluindo, em projetos posteriores,

[1] Reconhecemos que nem toda pesquisa que trabalha com um *corpus* (= conjunto de textos para análise) tem esse caráter descrito nem exige a utilização da LC.

para tornarem-se dois *corpora*, que podem ser reaproveitados por outros pesquisadores.

Para o desenvolvimento do artigo, além desta apresentação, pretendemos mostrar o histórico que nos levou ao trabalho com esses projetos (1 e 2), alguns pressupostos teóricos adotados na elaboração de *corpus*, a descrição da metodologia de trabalho em si, o percurso, do começo ao fim, entre o *corpus* inicial (do primeiro projeto) e o *corpus* final (do segundo projeto) e nossas considerações finais sobre todo o projeto.

2 Histórico

Nosso projeto partiu de uma ideia de reaproveitamento, em sala de aula, de *corpora* que os alunos elaboravam para a disciplina de graduação "Língua Inglesa: Estudos Descritivos e Linguística de Corpus 1" (de 60 horas, no curso de Inglês e Literaturas de Língua Inglesa), ministrada por um dos autores. A disciplina tinha dois objetivos de trabalho práticos: compilação e análise de *corpora* (com o consequente aprendizado de como trabalhar com ferramentas de análise lexical) e o uso deles para construção de vocabulários especializados por meio de um trabalho terminográfico a partir de um *software* desenvolvido por um dos autores (FROMM, 2007).[2] Em determinado momento, no primeiro ano em que a disciplina foi ministrada (2010), surgiu uma dúvida metodológica: deveríamos desenvolver novos *corpora* para cada turma, os chamados *corpora* descartáveis (VARANTOLA, 2002), ou poderíamos trabalhar num projeto maior que reaproveitasse esses *corpora*? Desdobrando um trabalho secundário desenvolvido na tese de um dos autores (FROMM, 2007), optamos por trabalhar com a compilação de um *corpus* que pudesse dialogar com cada aluno, por se tratar de sua própria área de atuação: Linguística.

A partir de um protótipo de Árvore de Domínio (taxonomia) da área de Linguística (FROMM, 2007), começamos os trabalhos com os alunos: cada grupo, composto por dois estudantes, deveria trabalhar com uma subárea da Linguística (teórica ou aplicada) de sua escolha, de forma

[2] VoTec. O resultado dos trabalhos dos alunos pode ser acessado em: http://treino.votec.ileel.ufu.br, área de Linguística. O vocabulário de Logística, por exemplo, foi uma das tentativas de trabalho colaborativo dos alunos que não prosperou (houve desinteresse pelo tema; percebemos, naquele momento, que o assunto a ser analisado é um passo metodológico importante quando da elaboração de *corpora* por parte dos alunos).

bilíngue (português e inglês), compilando *corpora* comparáveis (mesmo assunto) de 500 mil palavras daquela subárea e na língua de trabalho escolhida (a dupla trabalhava uma subárea, como *Etimologia*, e cada aluno se dedicava a uma língua). Como resultado geral, obtivemos, para cada dupla, um *subcorpus* de aproximadamente um milhão de palavras que poderia ser usado para vários tipos de pesquisas linguísticas.

3 Pressupostos teóricos

Apresentaremos, nos próximos subtópicos, alguns conceitos fundamentais sobre o que é um *corpus*, como ele se torna colaborativo e o que é um trabalho terminográfico (para o qual os *corpora* foram compilados).

3.1 Tamanho, balanceamento e representatividade

O **tamanho**, o **balanceamento** e a **representatividade** são conceitos e princípios inerentes à LC que nortearam a elaboração dos *corpora* (e posterior exploração dos mesmos para trabalhos terminográficos).

Em primeiro lugar, quanto ao **tamanho** de um *corpus*, Berber Sardinha (2004, p. 26-27) propõe três tipos de abordagens: (a) impressionística, (b) histórica e (c) estatística, e alguns parâmetros para o dimensionamento do tamanho do *corpus*: (1) ser representativo da comunidade de fala pesquisada; (2) atender o objetivo da pesquisa. Na (a) abordagem impressionística, o autor coloca como salvaguarda, já que não há maneira de saber o tamanho que representaria uma amostra ideal do recorte que está sendo feito, tornar seu tamanho o maior possível. A (b) abordagem histórica compara a frequência de palavras de acordo com a época da publicação das obras e descreve como essa frequência pode mudar com o tempo. Por fim, a (c) abordagem estatística baseia-se em dados matemáticos para definir a adequação do tamanho de um *corpus* e sua representatividade. Em conclusão, a questão de categorizar tamanho de *corpora* atualmente, tendo em vista os mega *corpora* (como os da BYU[3]), é um tanto complicada dada a velocidade com a qual a tecnologia de coleta (e análise) se desenvolve.

[3] Disponíveis em: https://www.english-corpora.org. Acesso em: 18 maio 2020.

Em segundo lugar, há o **balanceamento**: "processo pelo qual se garante que dois *corpora* sejam construídos de maneira similar quanto à origem, gênero, extensão, período de produção dos textos..." (TAGNIN; BEVILACQUA, 2013, p. 215), relacionado ao número de palavras e textos constituintes dos *corpora*.

Por fim, a **representatividade** está ligada ao princípio de que o *corpus* cumpre uma função representativa da língua (ou variedade linguística) e de uma comunidade representada. Logo, deve responder às perguntas: representativo de quê (amostragem) e para quem, considerados dados como frequência, número de palavras e aspecto probabilístico da linguagem (BERBER SARDINHA, 2004, p. 19-25).

3.2 Tipologia de *corpus*

A tipologia de *corpus,* em LC, busca descrever o desenho de um *corpus* e seu propósito de uso. A tipologia dos *corpora* desta pesquisa é baseada em Berber Sardinha (2004) e em Teixeira (2008), detalhada na seção 4.3.

Citamos algumas características apontadas por Berber Sardinha (2004, p. 20-21) relacionadas ao perfil dos *corpora* desta pesquisa. O autor inicia a descrição de tipologia de *corpus* com o critério escrito ou falado, sendo que o escrito pode ser impresso ou não. O segundo critério, o tempo, pode ser subdividido em: sincrônico, diacrônico, contemporâneo ou histórico.[4] O próximo critério, a seleção, se subdivide entre amostragem (estático, em oposição ao dinâmico), monitor (subdividido em dinâmico ou orgânico), e equilibrado.[5] O quarto critério, conteúdo, trata do texto

[4] O sincrônico abarca um período de tempo, enquanto o diacrônico abarca vários; o contemporâneo descreve o *corpus* de um tempo presente e o histórico o de um período de tempo passado.

[5] O *corpus* de amostragem envolve textos ou fragmentos de várias tipologias e gêneros textuais, como uma amostra finita da linguagem, em oposição ao *corpus* monitor, que descreve o estado atual da língua; o *corpus* dinâmico ou orgânico é flexível, pode ser aumentado ou diminuído diferentemente do estático (monitor); e o equilibrado é composto por textos em quantidades ou qualidades (tipologia, por exemplo) semelhantes. Neste projeto, o critério dinâmico ou orgânico descreve bem nossos *corpora*, no sentido de que podem ser aumentados com a adição de novas subáreas da Linguística; já critério equilíbrio é baseado no número de *tokens* que compõe os *corpora*. Por nossos corpora serem destinados a uma pesquisa terminográfica na área de Linguística, compilamos,

regional ou dialetal, especializado e multilíngue, o que nesta pesquisa significa *corpora* especializados da área de Linguística e multilíngues por abranger a Língua Portuguesa (doravante LP) e a Língua Inglesa (doravante LI).

Quanto à autoria, pode ser de aprendiz ou de língua nativa, isto é, conter textos de falantes não nativos ou nativos. O critério disposição interna trata se os textos do *corpus* são comparáveis (ex.: original e traduzido) ou se é alinhado (com linhas de tradução abaixo do texto original). A finalidade do *corpus* pode ser de (a) estudo, de (b) referência, de (c) treinamento ou teste. O objetivo do *corpus* de (a) estudo é descrever a língua; o de (b) referência contrasta com outros *corpora*; e o de (c) treinamento ou teste serve como subsídio para teste de ferramentas lexicais. Em nossos projetos, a finalidade do *corpus* é de estudo, já que os dados serão extraídos para definição terminológica.

Em comparação com a tipologia de Berber Sardinha (2004), a tipologia de Teixeira (2008) difere quanto ao item nível de codificação, o qual abrange metadados como cabeçalho e etiquetagem. Em ambos os projetos descritos neste artigo, a etiquetagem não se fez necessária, pois a ferramenta de palavras-chave do WST cumpriu a tarefa de isolar os termos, objeto das pesquisas terminográficas. Quanto aos cabeçalhos, eles foram inseridos manualmente e neles constam o título do artigo, a data de coleta e o *site* de origem.

3.3 O que é um *corpus* colaborativo?

Acreditamos que, em primeiro lugar, devamos situar o que consideramos um *corpus* colaborativo. É importante especificar, pois a ideia pode advir de algo generalista: Gardner, Krowne e Xiong (2010), por exemplo, consideram a Wikipedia como um grande *corpus* colaborativo.[6]

A proposta de compilação de *corpora* também pode abranger vários níveis de colaboração, como explicitada por Mello (2014), que

---

especificamente, textos especializados nessa área. Assim sendo, nosso foco foi a compilação de textos do gênero acadêmico, subdivididos em artigos científicos, resenhas e material instrucional, como manuais, aulas, livros, apostilas e resenhas.

[6] Em nenhum momento, no texto referido, os autores citam a Wikipedia como ideia de um *corpus* com a mesma conotação dada pela Linguística de *Corpus*. A proposta aqui é mais genérica, trabalhando com a definição de *corpus* como conjunto de textos que, no caso, são construídos colaborativamente.

trabalha com *corpora* orais (espontâneos de fala) no projeto C-ORAL-BRASIL, envolvendo a descrição de textos das gravações, anotações e a inserção de metadados sociolinguísticos, por exemplo; todo o trabalho é pensado coletivamente e deve ser gerido por um ou mais grupos de estudo, em nível nacional ou internacional. Nessa empreitada de compilação, há vários estágios a serem desenvolvidos pelos grupos: planejamento da arquitetura do *corpus* (com uma grande preocupação em relação às questões de tamanho e balanceamento), gravações (e todos os aparatos técnicos necessários), transcrições (obedecendo a critérios rigorosíssimos dentro de um projeto maior, no qual se insere), revisões das transcrições (inclusive com análises estatísticas), segmentação e alinhamento dos formatos de saída do *corpus*.

Nossa proposta, aqui, é muito mais restrita e voltada para a sala de aula. Bowker (2003) desenvolveu projetos semelhantes com alunos de tradução.[7] Embora não tivéssemos conhecimento de seu texto na época que começamos a trabalhar com nosso *corpus* colaborativo, percebemos que suas preocupações foram semelhantes às nossas: devemos usar um *corpus* já existente ou montamos um *corpus* específico para nossos experimentos? Os alunos têm fácil acesso aos computadores para compilar *corpora*? Um trabalho coletivo não poderia gerar um *corpus* maior e mais representativo? Tudo o que está disponível na internet pode virar *corpus* de estudo? Nossos *corpora* precisam de etiquetagem? Todas essas questões levam a uma pergunta básica: qual a metodologia a ser adotada para o trabalho colaborativo? Discutiremos, mais adiante, quais foram as nossas propostas metodológicas.

Diante do exposto, definimos *corpus* colaborativo como um *corpus* compilado por vários pesquisadores, em momentos diferentes, construído especificamente para atender à construção de um vocabulário bilíngue de Linguística. Como resultado, obtivemos um *corpus* mais representativo das subáreas da Linguística, sem necessidade de etiquetagem, já que o foco da pesquisa foram os termos da área.

---

[7] Diferente de nós, que trabalhamos com *corpora* comparáveis (textos de línguas diferentes, pertencentes à mesma área ou subárea de conhecimento, porém não se constituem como traduções), a autora trabalhou com seus alunos *corpora* paralelos (no par francês-inglês, tradução do mesmo texto de uma língua à outra). O tamanho adotado para cada *corpus* também foi diferente: enquanto partimos do pressuposto de um *corpus* com, pelo menos, 500 mil palavras em cada subárea, a autora trabalhou com *corpora* entre 20 e 50 mil palavras.

3.4 Trabalho terminográfico

Existem várias teorias que podem ser usadas para trabalhos terminográficos. Durante todo o projeto relatado aqui, pautamo-nos por duas delas. A primeira diz respeito à tipologia para obras de consulta ao léxico, sistematizada por Barbosa (2001), que divide a elaboração de obras lexicográficas e terminográficas entre (a) dicionário, (b) vocabulário e (c) glossário. De acordo com a autora (2001, p. 36), o (a) dicionário busca compilar e registrar as palavras de frequência regular situadas no campo de várias normas; o (b) vocabulário registra vocábulos delimitados por uma situação comunicativa, podendo estar relacionados a usos em espaços geográficos e grupos sociais diferentes e pertencentes ao mesmo universo de discurso; e o (c) glossário registra palavras provenientes de um texto ou discurso específico relacionadas ao mesmo tempo, espaço geográfico e contextos de uso da palavra. O vocabulário pode se subdividir em: (a) vocabulário fundamental e (b) vocabulário técnico-científico e especializados. O vocabulário fundamental abarca os vocábulos frequentes e regularmente distribuídos de uma comunidade ou de um- grupo social específico. Em nosso contexto, o vocabulário técnico-científico e especializados, que é um recorte de língua (de grupos profissionais), foi o escolhido para o trabalho com os alunos. Este vocabulário trabalha com uma área de especialidade, sua unidade é o termo (com significado restrito e alta frequência), o verbete pode apresentar mais de uma acepção (mas apenas dentro da área escolhida) e a perspectiva desse vocabulário é sincrônica e sinfásica, ou seja, relativo ao estilo de língua, tais quais familiar, formal ou literário.

A segunda, Teoria Comunicativa da Terminologia, de Cabré (2000), elabora o conceito de "termo". Entre as várias inovações propostas em sua teoria em relação à Teoria Geral da Terminologia de Wüster (publicada originalmente em 1931), de caráter prescritivo e que a precede, temos: (1) caráter descritivo que perpassa todo um projeto; (2) uso de *corpora* na coleta dos termos; (3) concepção que o termo é, basicamente um substantivo (embora outras palavras lexicais, como adjetivos e verbos, também possam aparecer como características de uma determinada área); (4) um termo **está** termo dentro de determinada área (geralmente com significado monossêmico), mas pode, ao mesmo tempo, pertencer à língua geral, com significados polissêmicos.

3.5 A importância da árvore de domínio

A construção da árvore de domínio ou conceitual[8] é um passo metodológico da Terminologia, como parte do percurso semasiológico,[9] que permite o levantamento e a delimitação conceitual dos termos a serem pesquisados e definidos numa obra terminográfica. Segundo Aubert (2001), a árvore de domínio, com suas áreas e subáreas definidas, contribui para evitar o que o autor denomina riscos de "ruído" e de silêncio, isto é, a presença de termos não relevantes a uma subárea ou a ausência de termos essenciais a uma área ou subárea específica. A árvore pode servir tanto ao pesquisador quanto aos leitores da obra desenvolvida, proporcionando-lhes uma visão da abrangência do tema e do trabalho desenvolvido.

A árvore de domínio contribui para a sistematização dos termos mais frequentes (procedimento este possível graças à LC) e para a categorização por subáreas numa pesquisa (ALVES *et al*., 2010; BARROS, 2004; FROMM, 2015, 2018). Ao adotarmos a LC como abordagem e metodologia para análise dos dados linguísticos, essa sistematização se dará a partir da lista de palavras-chave (candidatos a termos que podem ser classificados, posteriormente, nos subdomínios da árvore de proposta). Em seguida, por exemplo, analisando os cotextos e contextos em linhas de concordância, podemos identificar os semas ou traços semânticos que servirão à elaboração de um verbete.

3.6 Definição terminográfica baseada em *corpus*

Existem várias maneiras de construir uma definição, seja para fins lexicográficos, seja para fins terminográficos (como bem explicado no trabalho de CARDOSO, 2017). De modo geral, e para nosso projeto,

---

[8] A árvore de domínio é um esquema representativo de como se dividem as principais subáreas de uma grande área de conhecimento (como a árvore do CNPq – Conselho Nacional de Desenvolvimento Científico e Tecnológico –, por exemplo). Pode ser em formato de diagrama ou disposta em formato topicalizado. Ela pode ser levantada através de pesquisa com especialistas na área e/ou (como propomos) através da análise de *corpus* especializado.

[9] Percurso semasiológico, no contexto da Terminologia e da Terminografia, é quando partimos do termo para chegarmos ao seu conceito e definição, contrário ao percurso onomasiológico, que parte do conceito para chegar ao termo. Definir o que é Biolinguística, por exemplo, faz parte do percurso semasiológico.

em específico, usa-se o modelo de definição GPDE (Gênero Próximo, Diferenças Específicas) (FROMM; YAMAMOTO, 2020), também conhecido como Aristotélico ou por compreensão. O que Fromm (2007) propôs, no entanto, é que a definição, cujo modelo foi aplicado no *software* que elaborou para sua tese, fosse elaborada única e exclusivamente por meio de uma Análise Componencial (ILARI, 2003) de semas e palavras advindos de excertos coletados no *corpus* de pesquisa; ou seja, a definição terminográfica, no nosso projeto, deve ser baseada em *corpus*. Para os excertos, procuramos levantar aqueles que contivessem mais contextos definitórios e explicativos. Segundo Aubert (2001, p. 69), o contexto definitório descreve "o conjunto completo dos traços conceptuais distintivos do termo" e o contexto explicativo descreve alguns "traços conceptuais pertinentes específicos do termo", geralmente relacionados à "materialidade, finalidade, funcionamento e similares". Em relação à validação dos termos por parte de especialistas, preferimos usar a abordagem proposta por Tagnin (2012): não necessariamente precisamos da validação de especialistas para termos extraídos a partir de *corpora* sólidos e bem montados e que possam ser considerados altamente confiáveis. Respaldados em Tagnin (2012, p.169) consideramos que o maior especialista de uma ciência que possa validar verbetes sobre a área é o *corpus* que montamos sobre ela, não os profissionais que dela fazem parte, pois é difícil algum pesquisador/usuário ter noção de sua área como um todo.

3.7 Reciclagem de *corpus*

Na LC, o termo **reciclar**[10] pode ser usado com mais de um significado. Ele pode ser utilizado, por exemplo, para descrever a análise da padronização lexical, nas linhas de concordância, aplicada ao ensino e aprendizagem de língua estrangeira e de segunda língua (PÉREZ-PAREDES; SÁNCHEZ-TORNEL; CALERO, 2012; SINCLAIR, 2003). Sinclair (2003, p. XIV-XIX) subdivide esse procedimento de análise em sete outros passos: início, interpretação, consolidação, relato, reciclagem, resultado e repetição. Nesse contexto, o resultado, segundo o autor, significa elaborar uma lista de hipóteses finais sobre as informações provenientes das linhas de concordâncias e associá-las ao relatório final do nó da análise inicial.

[10] *Recycle* ou *reuse* em inglês.

Outro conceito existente na LC aplicado à reciclagem de *corpora* consiste na reutilização de *corpora* para atender objetivos específicos, distintos daqueles de seu projeto original. Por exemplo, os passos metodológicos que fazem parte do processo de reciclagem em contexto de tradução com *corpora* paralelos são: a coleta de material, a extração dos dados reutilizáveis e o refinamento e a aplicação dos dados extraídos (TIEDEMANN, 2003). A coleta de material trata do levantamento dos *corpora* paralelos utilizados para análise das possíveis traduções. Esses *corpora* etiquetados e alinhados podem ser aplicados, como recurso computacional, na lexicografia e na tradução automática.

O objetivo da reciclagem do *corpus* de Linguística (do primeiro projeto) que propomos é, primeiramente, não perder o produto resultante de um trabalho prévio. Em segundo lugar, reduzir tempo de uma nova pesquisa, caso *corpora* prévios possam ser reutilizados para análises linguísticas de um novo projeto ou para dar continuidade a um projeto preexistente, como foi nosso caso.

Quando os *corpora* compilados são grandes e provenientes de um período longo de coleta (abrangendo anos), é necessário considerar algumas variáveis: (1) **disponibilidade** atual do *corpus*; (2) **limpeza** a ser feita; e (3) **tipologia**, quando se trata da fidelidade ao tipo ou gênero textual.

A **disponibilidade** de um *corpus* trata da questão da disponibilidade, na Internet, dos arquivos que o compõem; isto é, devido à volatilidade da rede mundial, alguns textos podem ou não continuar disponíveis por muito tempo. Caberá aos pesquisadores se os textos já coletados servirão ou não aos objetivos de suas pesquisas e, então, definirem se vão mantê-los ou excluí-los de seus *corpora*. Essa indisponibilidade na Internet se deve a fatores como: a mudança do endereço na web, a retirada do arquivo da Internet ou a alocação do arquivo em outro *site* ou base de dados.

A **limpeza** é relacionada à remoção de dados não relevantes ao conteúdo linguístico que se objetiva analisar em um *corpus* (elementos pré e pós-textuais, referências bibliográficas, dados de quadros e tabelas que não estejam inseridos como figuras, elementos textuais em língua estrangeira etc.). Utilizamos os comandos de *Control F* e *Substituir* para localizar e deletar esses dados de forma manual.

A padronização e o balanceamento da **tipologia** ou o **gênero dos textos** são também fatores importantes ao se considerar a composição de *corpora*: o desenho dos *corpora* deve prever quais tipos e quantos

textos (ou *tokens*,[11] isto é, o número de itens ou ocorrências de um vocábulo que comporão o *corpus*) farão parte desses *corpora*, sob pena de, se mal planejados, apresentar resultados de análise que podem ser desacreditados.

Por fim, a partir de um *corpus* inicial, após ser limpo e equilibrado, é possível obter um novo *corpus* alterado a fim de atender aos propósitos do projeto de pesquisa. Em nosso caso, o *corpus* inicial serviu à construção do *corpus* final (reciclado) e à construção do vocabulário bilíngue de Linguística (YAMAMOTO, 2018).

3.8 Padronização de *corpora*

A ideia de padronização se refere ao ato de equilibrar o tamanho, o gênero e o subgênero textual e as demais características dos *corpora*, que devem ser previstas no desenho dos mesmos. O tamanho de cada *corpus* (de cada projeto) e dos sub*corpora* que o compõem, no nosso caso, deveu-se a uma pesquisa prévia (FROMM, 2007).[12] Também foram feitas pesquisas (notadamente com o buscador Google) para descobrir, a partir da área pesquisada (Linguística), a prevalência de gêneros (textos acadêmicos e instrucionais) e subgêneros (teses, dissertações, artigos científicos, resenhas e manuais).

4 Metodologia

Destacamos, nos próximos subitens, os passos seguidos na elaboração do *corpus* colaborativo de Linguística (primeiro projeto) e sua posterior reciclagem (segundo projeto). Para os *corpora* aqui descritos, adotamos as abordagens impressionística, mencionada por Berber

[11] Na frase "O maracujá e o abacate são frutas" há 7 *tokens*.

[12] Em sua tese (FROMM, 2007), um dos autores notou, em suas conclusões, que para a elaboração de trabalhos terminográficos baseados em *corpora*, um *corpus* de 30 mil palavras para cada subárea da computação (*corpus* escolhido para testar o *software* desenvolvido para a tese) não era suficiente para a elaboração de definições terminológicas. A partir dessa constatação, tomou-se como ponto de partida, para a compilação dos *corpora* de Linguística aqui descritos, um tamanho cerca de 16 vezes superior para cada subárea, de modo que os requisitos de elaboração de um vocabulário criado única e exclusivamente a partir de *corpora* (especialmente no tocante à elaboração das definições) pudessem ser contemplados.

Sardinha (2004), e histórica, baseada na experiência da compilação de *corpus* para um trabalho terminográfico anterior (FROMM, 2007).

4.1 A elaboração da árvore de domínio

Existem várias maneiras de elaborar a taxonomia de um domínio da ciência. Para os nossos *corpora*, foi pensada uma pergunta básica para a escolha das subáreas que compõem a árvore: qual seu objeto de estudo? Por exemplo, a Morfologia analisa as classes de palavras, os morfemas, as unidades mínimas de sentido etc. Uma pergunta básica como essa norteou todo o design dos nossos *corpora*. Não entraram aqui, por exemplo, teorias associadas às diferentes áreas da Linguística[13] ou metodologias[14] usadas em pesquisas na área. Linguística de *Corpus* e a Linguística Computacional, por exemplo, são muito mais abordagens e metodologias do que áreas específicas.

Voltando à pergunta mencionada no parágrafo anterior (qual o objeto de estudo?), a LC, ao nosso ver, não tem **um** objeto específico de pesquisa, já que pode ser usada por quase todas as áreas elencadas na árvore, por isso a consideramos uma abordagem (computacional) e metodologia (muito bem delimitada) de pesquisa.

Inicialmente a árvore foi concebida por Fromm (2007) e a metodologia adotada para a elaboração foi a consulta a livros da área de Linguística. O autor decidiu pela divisão da grande área entre Linguística Teórica e Linguística Aplicada, inserindo nelas subáreas: 25 subordinadas à Linguística Teórica e 6 à Linguística Aplicada.

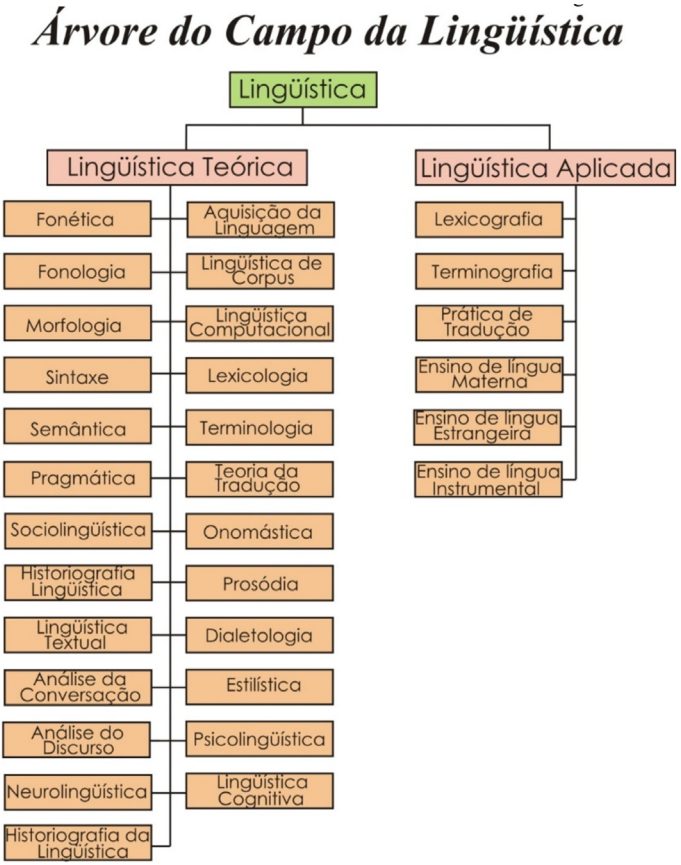
Nas pesquisas terminográficas tradicionais, a elaboração da árvore de domínio, com ajuda de especialistas da área a ser estudada, é essencial para que os termos sejam classificados. Nesta pesquisa, essa árvore possui certa plasticidade, pois pode ser alterada à medida que o *corpus* é compilado. Na Figura 1, observamos a árvore prototípica que deu início ao projeto (FROMM, 2007, p. 39).

---

[13] Note-se que uma mesma teoria (como o Gerativismo ou a Linguística Sistêmico-Funcional) pode ser encontrada como base teórica em várias subáreas da nossa árvore; ou, ainda, uma subárea pode ser analisada por diferentes teorias linguísticas.

[14] Essa é a percepção atual dos autores, diferente do que é apresentado na figura 1, quando os questionamentos sobre teoria, metodologia e objeto de estudo ainda estavam sendo discutidos.

FIGURA 1 – Primeira versão da Árvore de Domínio da Linguística



Fonte: Fromm (2007, p. 39).

Com a compilação dos *corpora* (LP e LI) e o desenvolvimento da pesquisa nas diversas turmas que cursaram a disciplina, o formato da árvore foi mudando. Um passo adotado quanto à escolha de qual área a dupla de alunos trabalharia pode ser verificado na Figura 2. Conforme as áreas eram pesquisadas, as cores nos retângulos correspondentes a cada subárea mudavam:

a) vermelho para as subáreas consideradas completas (com, no mínimo, 500 mil *tokens*), impedindo que novos grupos escolhessem aquelas subáreas;

b) amarelo para as subáreas que ainda não estavam completas, pois ainda faltavam *tokens* em uma ou ambas as línguas[15] – no caso, uma nova dupla trabalharia para completar o que estava faltando (os textos já coletados eram repassados ao novo grupo);

c) verde para as áreas ainda não trabalhadas e que poderiam ser livremente escolhidas pelos alunos.

FIGURA 2 – Versão intermediária da Árvore de Domínio da Linguística

Fonte: Elaborada pelos autores

[15] O desenrolar da compilação do *corpus* nos mostrou que determinadas subáreas possuíam grande quantidade de textos em língua inglesa, mas poucos textos em língua portuguesa, indicando um desbalanço no estado da arte de língua para língua. No caso específico da Ecolinguística, foram necessários 4 anos para que a quantidade mínima fosse atingida em língua portuguesa.

A plasticidade da árvore, já citada anteriormente, pode ser notada entre as três figuras aqui apresentadas. Nota-se, por exemplo, que **Pesquisa Narrativa** aparece somente na Figura 2; na época, incluída por sugestão de uma especialista. Posteriormente, decidimos que certas áreas, como Pesquisa Narrativa, seriam consideradas metodologias/abordagens e não áreas com objetos de pesquisa bem delimitados, sendo retiradas das árvores (e seus respectivos sub*corpora* do *corpus* de LA de cada língua).

A elaboração da árvore a partir das informações advindas do *corpus* nem sempre fica clara. Por exemplo, quando da condução dos estudos de mestrado de um dos autores (YAMAMOTO, 2015), observamos que, apesar de semelhantes em relação à abordagem diacrônica da língua, as subáreas da Etimologia, Filologia e Linguística Histórica haviam se consolidado como disciplinas em épocas diferentes (já constando da FIGURA 2 também), e eram distintas quanto ao objeto de estudo. Essas subáreas foram fundidas ou desmembradas, entrando e saindo da árvore (e dos *corpora*) várias vezes.

A plasticidade na construção da árvore de domínio da Linguística fez parte de uma correspondência biunívoca entre a montagem da taxonomia da área e o *corpus* dessa área. A partir da árvore inicial (FIGURA 1) e passando pelas diversas fases de sua elaboração (como na FIGURA 2), conforme o *corpus* era elaborado, os pesquisadores percebiam, nas fontes consultadas, que novas subáreas da Linguística (como a Biolinguística e a Linguística Diacrônica, já presentes na FIGURA 3) iam surgindo com o tempo (num período de 10 anos). Conforme o *corpus* demonstrava a existência de novas subáreas, as mesmas iam sendo in*corpora*das à árvore e novos levantamentos foram feitos para completar os 500 mil *tokens* em cada uma delas. Essa metodologia de elaboração de árvore e compilação de *corpus,* acreditamos, poderia ser considerada uma complementação à definição de pesquisa direcionada pelo *corpus* (*corpus driven*) proposta por Viana e Tagnin (2015, p. 323): "[...] estudo que se desenvolve conforme dados apresentados pelo *corpus*, sem pressuposições teóricas".

FIGURA 3 – Versão atual da Árvore de Domínio da Linguística

Fonte: Elaborada pelos autores.

Algumas decisões sobre como nomear áreas e subáreas na árvore, assim como a profundidade de análise, foram tomadas como um consenso entre os autores deste texto. Um exemplo é a questão de quantos níveis deveríamos descer na estrutura. Algumas áreas, como a Onomástica, por exemplo, ainda podem conter subníveis (Toponímia e Antroponímia, no caso). Devido à complexidade do trabalho e à indisponibilidade de tempo, decidiu-se por uma árvore de apenas 3 níveis.

No segundo nível da árvore, podemos notar uma renomeação de Linguística Teórica para Linguística e, finalmente, para Linguística Descritiva (LD). Para fazermos nossa escolha, consideramos o caráter descritivo da Linguística como ciência, e não o caráter prescritivo da Gramática. No intuito de corroborar nossa escolha, fizemos a busca das definições nos *corpora* de estudo em LP e encontramos a definição de LD em contraste com a LA, conforme Albuquerque (2017, p. 227):

> [...] as duas principais subáreas da linguística de acordo com a necessidade de pesquisa e contribuições científicas são a linguística descritiva e a linguística aplicada. A primeira por sua clara importância documental e analítica das línguas do mundo. A segunda por suas contribuições significativas no campo educacional, que procuram criar uma ponte entre teoria e prática, entre universidade e comunidade.

Temos aí uma citação que compara as duas Linguísticas e as contrasta, de forma que os argumentos são claros e objetivos, atendendo nossa proposta. Com exceção da terminologia adotada na Figura 2 (que pode ser considerada um erro, pois a subárea de segundo nível e a área principal, no primeiro nível, possuíam a mesma designação), a preferência pela nomeação Linguística Descritiva pode ser considerada como uma decisão política dos pesquisadores.

É importante notar, no entanto, que a árvore proposta na Figura 3[16] (derivada da pesquisa de doutorado de um dos autores) não é definitiva;[17] nenhuma árvore de domínio pode ser considerada finalizada, pois subáreas aparecem (e, por que não, desaparecem), conforme o desenvolvimento científico. E, já que esses projetos tratam da compilação de *corpora* a partir de uma correspondência biunívoca,[18] temos que reconhecer as limitações desses projetos: nem todas as possíveis subáreas da Linguística podem estar, em termos de quantidade de textos, igualmente representadas na internet. Um exemplo dessa questão foi a área de **Etimologia** (assim

[16] Há outras versões da árvore, como as propostas em Fromm e Yamamoto (2013) e Fromm (2018).

[17] Especialmente porque, conforme mencionado, não continuamos a pesquisa para um possível quarto nível.

[18] Dados levantados a partir do *corpus* sugerem a (re)elaboração dos subcampos da árvore ⟷ os subcampos da árvore (pesquisados pelos autores e sugeridos por colegas pesquisadores) indicam os *subcorpora* a serem levantados.

como a de Ecolinguística, já citada), cujo *corpus* demorou vários anos para ser compilado durante os projetos; nossa interpretação do fato é que, como novas subáreas aparecem (e, com elas, paulatinamente, novos textos vão sendo escritos), outras podem estar com uma baixa produção textual, ao menos no momento de compilação do *corpus*;[19] logo, a compilação de seus respectivos *corpora* precisaria passar por uma metodologia de coleta que não simplesmente o *download* de textos disponíveis na rede (como, por exemplo, escanear textos impressos, passá-los por um OCR[20] e disponibilizá-los eletronicamente).

4.2 Os gêneros adotados

Desde o início, o gênero adotado para a compilação das diversas fases dos *corpora* foi o acadêmico, subdividido pelos subgêneros: tese, dissertação e artigo científico. Posteriormente, na pesquisa de doutorado de um dos autores, por sugestão de um dos membros de sua banca de mestrado (YAMAMOTO, 2015), foram incluídos materiais instrucionais (manuais).[21]

4.3 O *corpus* colaborativo

A tipologia (de acordo com Teixeira (2008)) do *corpus* colaborativo de Linguística (primeiro projeto), compilado pelos alunos, é descrita no Quadro 1.

---

[19] O *corpus* desta pesquisa abarca produções compiladas entre os anos de 2010 e 2020, cujos textos podem ter sido publicados antes de 2010 até 2018, sendo, portanto, um *corpus* contemporâneo.

[20] Em inglês, *Optical Character Recognition*, que, em português, significa Reconhecimento óptico de caracteres é uma tecnologia que permite a conversão de imagens, textos em PDF, documentos escaneados, entre outros, em textos legíveis e editáveis.

[21] Manuais, aqui, são livros gerais de Linguística voltados aos alunos, pois eles contêm traços semânticos mais fáceis de serem identificados (= apresentam mais contextos definitórios) e usados em pesquisas terminológicas ou terminográficas.

QUADRO 1 – Tipologia do *corpus* colaborativo do primeiro projeto

Língua	Bilíngue (inglês e português)
Modo	Escrito (textos acadêmicos: artigos científicos, dissertações e teses)
Data de publicação	Sincrônico (levantamento realizado entre 2010 e 2018), fechado[22]
Seleção	Estático
Conteúdo	Especializado (Linguística)
Autoria	Falantes nativos/não nativos (inglês e português), individual/coletivo
Disposição Interna	Comparável
Uso na pesquisa	Estudo (análise terminológica/terminográfica)
Tamanho	Grande (mais de 10 milhões de palavras)
Nível de Codificação	Com cabeçalhos,[23] sem etiquetas[24]

Fonte: Elaborado pelos autores.

Em relação ao tamanho do *corpus* desse primeiro projeto, podemos verificar os dados na Tabela 1.

[22] Mesmo com algumas subáreas incompletas, foi fechado por causa do encerramento do trabalho com os alunos neste *corpus* de Linguística.

[23] Informações de coleta delimitadas por um sinal de menor e maior, que permite ignorá-las nas buscas feitas nas ferramentas de análise (ver FIGURA 4).

[24] Nunca houve, por parte dos pesquisadores, preocupação em etiquetar as diferentes versões do corpus. Como o objetivo da compilação era o trabalho terminográfico, as ferramentas básicas (lista de palavras, palavras-chave e concordanciador) dos programas de análise lexical (conforme o nome já sugere, os mesmos foram desenvolvidos para trabalhos a partir do léxico) não necessitavam de etiquetagem para o trabalho de elaboração de um vocabulário na área de Linguística. Embora a teoria de Cabré indique um destaque para substantivos como candidatos a termos, é muito mais fácil, para os pesquisadores, usarem a ferramenta palavras-chave do que etiquetar o corpus (e ter que revisar todas as etiquetas, além do problema de achar um bom etiquetador em língua portuguesa).

TABELA 1 – Tamanho do *corpus* colaborativo

	Português	Inglês
Número de textos	1.873	1.911
Tokens	26.856.704	29.369.816
Types[25]	353.700	429.384
Type/token ratio (TTR)[26]	1,32%	1,46%
Número Total de *Tokens*	56.226.520	

Fonte: Elaborada pelos autores.

4.4 Armazenamento de *corpora*

A organização e o arquivamento de *corpora* em uma máquina são passos muito importantes, dos quais depende o acesso eficaz aos arquivos do projeto, isto é, o *corpus*, as listas de palavras, as listas de palavras-chave e de outros documentos necessários à produção das obras terminográficas.

4.4.1 *Corpus* colaborativo

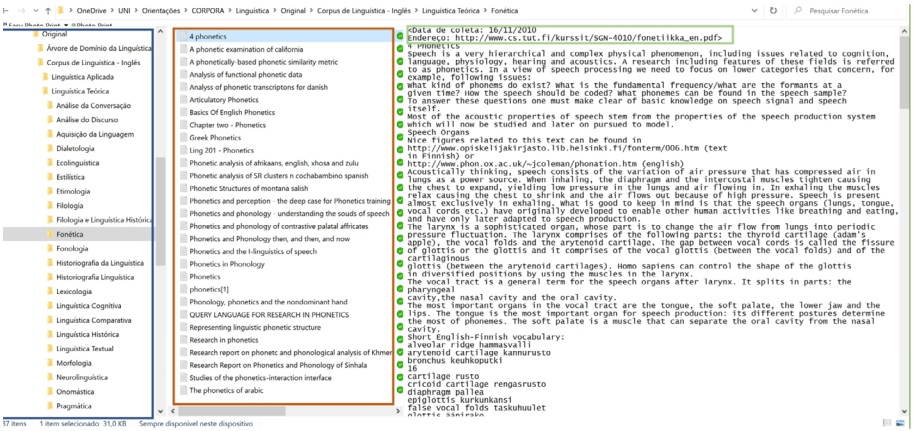
Podemos descrever a estruturação de compilação do projeto inicial através da Figura 4. Não houve, no começo do projeto, uma preocupação em salvar os arquivos com algum título, tipo de arquivo ou codificação específicos; os alunos apenas salvavam os artigos que baixavam em formato .txt com seus nomes completos (retângulo vermelho). Foram inseridos, a partir do segundo ano, pequenos cabeçalhos (no começo dos textos), nos quais constavam data e endereço de coleta (retângulo verde).

[25] *Types* são o total de formas, ou o número de vocábulos presentes em um dado *corpus*, ou seja, na frase "O maracujá e o abacate são frutas" há 6 *types*, já que o artigo **o** repete duas vezes.

[26] *Type/token ratio* (*TTR*) é o resultado proveniente da divisão das formas (*types*) pelos itens (*tokens*) de um *corpus* expressa em porcentagem. Esta razão indica a riqueza lexical de um *corpus*, de forma que quanto maior for esta razão, maior será o número de palavras diferentes contidas naquele *corpus*.

Para montar a estrutura de compilação, simplesmente reproduzimos a estrutura da árvore de domínio da Linguística (quadro azul).[27]

FIGURA 4 – Estruturação inicial do projeto



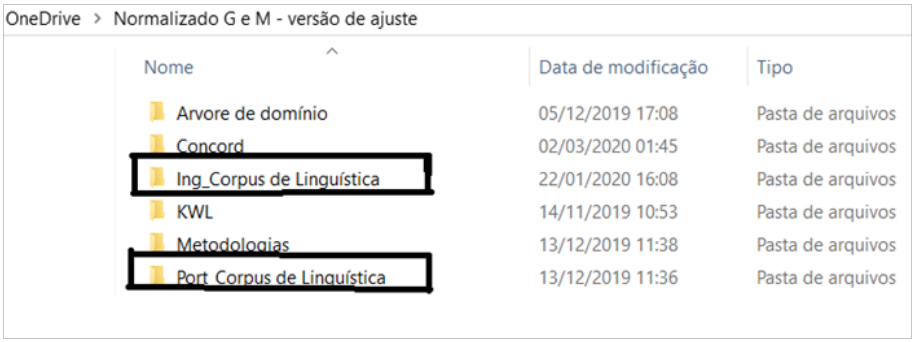
Fonte: Elaborada pelos autores.

4.4.2 *Corpus* reciclado

Para a execução do segundo projeto, organizamos uma pasta em um disco rígido onde foram arquivados os *corpora* de Linguística, em LP e LI, e essa pasta foi compartilhada entre os pesquisadores por meio do OneDrive, ferramenta usada para a disponibilização de dados nas nuvens. Na Figura 5, podemos ver as pastas destes *corpora*.

[27] Obviamente a estruturação das pastas foi sendo alterada de acordo com as revisões da árvore de domínio. Nesta figura, podemos perceber que a estrutura está mais próxima da árvore mostrada na Figura 2.

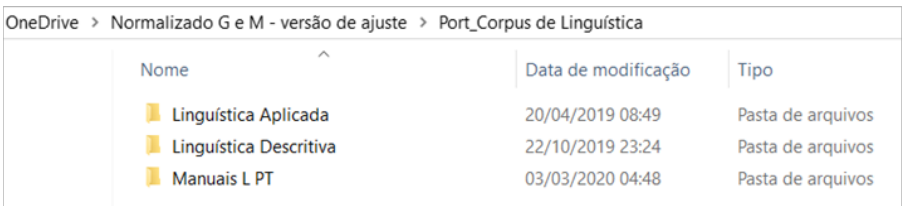
FIGURA 5 – Pasta para arquivamento dos *corpora* de Linguística



Fonte: Elaborada pelos autores.

A Figura 5 mostra que os *corpora* de Linguística foram arquivados numa pasta RAIZ, nomeada **Normalizado G e M** (iniciais dos pesquisadores). Dentro dessa pasta, há as subdivisões nas duas línguas do projeto, ou seja, em LP, Port_*Corpus* de Linguística, e em LI, Ing_*Corpus* de Linguística. Ao abrir essas pastas, encontramos três subpastas: as duas subáreas da Linguística e a pasta com o *corpus* de manuais de Linguística. Veja a Figura 6.

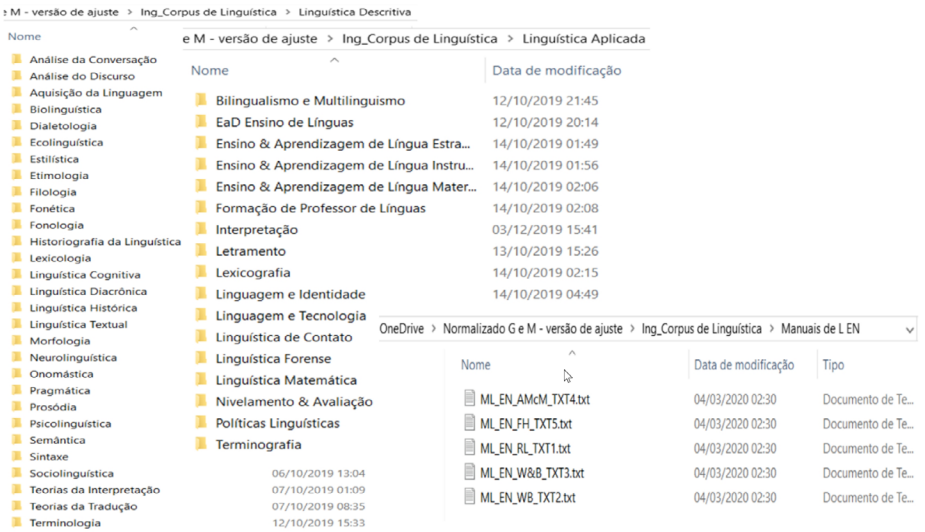
FIGURA 6 – Subpastas: LD, LA e manuais de Linguística



Fonte: Elaborada pelos autores.

Na Figura 6, encontramos a pasta de Linguística Aplicada, com suas 18 subáreas, a pasta de Linguística Descritiva, com suas 29 subáreas e o *corpus* com os 13 manuais de Linguística em LP. O mesmo acontece para a LI e cinco manuais de Linguística, que podem ser vistos na Figura 7.

FIGURA 7 – Subpastas de LD, LA e manuais de Linguística com arquivos codificados em LI

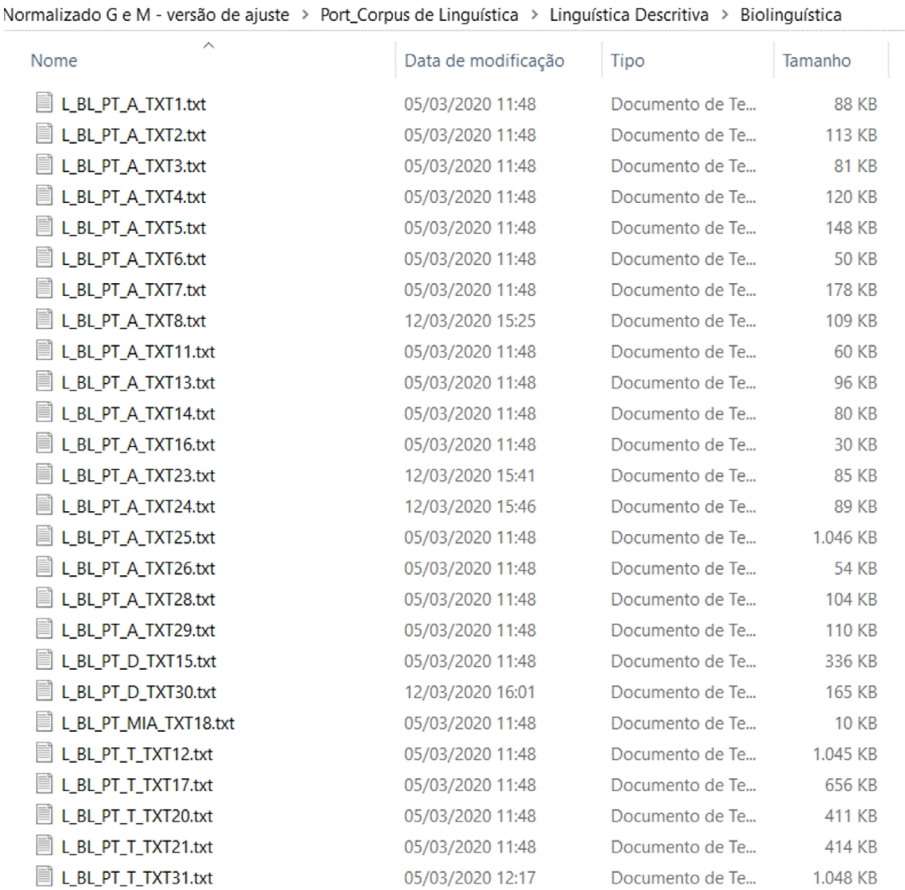


Fonte: Elaborada pelos autores.

Na Figura 7, é possível visualizar a pasta de Linguística Descritiva e suas subáreas: da Análise da Conversação à Terminologia; de Linguística Aplicada: das subáreas Bilinguismo e Multilinguismo à Terminografia; e, finalmente, o *corpus* de manuais de Linguística em LI, codificados (ML – manual de Linguística, língua: EN de English, abreviatura dos autores e o número do arquivo). Nesses passos, buscamos mostrar de forma objetiva uma possibilidade de organização de grandes *corpora* para um trabalho terminográfico.

Com o objetivo de facilitar o processo de explicação e compreensão dos procedimentos, expomos o processo de nomeação dos arquivos da pasta de Biolinguística em LP na Figura 8.

FIGURA 8 – *Subcorpus* de Biolinguística: nomeação dos arquivos



Fonte: Elaborada pelos autores.

O primeiro arquivo traz a codificação **L_BL_PT_A_TXT1**. A letra **L** significa que o arquivo pertence à subárea da Linguística Descritiva. Essa codificação foi criada antes que o binômio Linguística Descritiva fosse adotado pelos pesquisadores. Caso esse mesmo procedimento fosse adotado hoje, seria mais adequado usar as letras LD, que melhor representariam primeira subdivisão da árvore, ou mesmo um sistema de renomeação automática como a proposta pela plataforma do ToGatherUp de Oliveira (2019). Em contraste à essa codificação, há a abreviação **LA** para a subárea da Linguística Aplicada. Em seguida, há a abreviação **BL**, adotada para a

subárea da Biolinguística, existente no terceiro nível da árvore de domínio da Linguística. Depois, apresentamos a língua do *subcorpus* – neste caso, o português –, codificado por **PT**. No caso do inglês, a codificação foi **EN**, como mencionado previamente. Depois da língua, expomos o gênero ou subgênero, de quatro formas diferentes: **A**, **D**, **MIA** e **T**. A letra **A** foi usada para **artigos científicos**; **D** para **dissertação** de mestrado; **MIA** para **Material instrucional**; e **T** para **teses** de doutorado. Finalmente, há o número do texto, a partir do número 1, apesar de, em geral, não seguir a sequência correta, já que arquivos foram removidos para que houvesse o balanceamento dos *corpora*. Não consideramos necessária a indicação do ano de publicação dos textos, pois não era um trabalho de perspectiva histórica ou diacrônica, mas caso esta perspectiva fosse adotada, o ano de publicação teria sido necessário. O mesmo procedimento pode se aplicar ao quesito de autoria das obras, caso seja relevante, como foi no segundo projeto: o dado de autoria facilitou a organização do *corpus* de manuais em sua primeira fase, o que não persistiu quando do carregamento dos textos na plataforma do ToGatherUp.

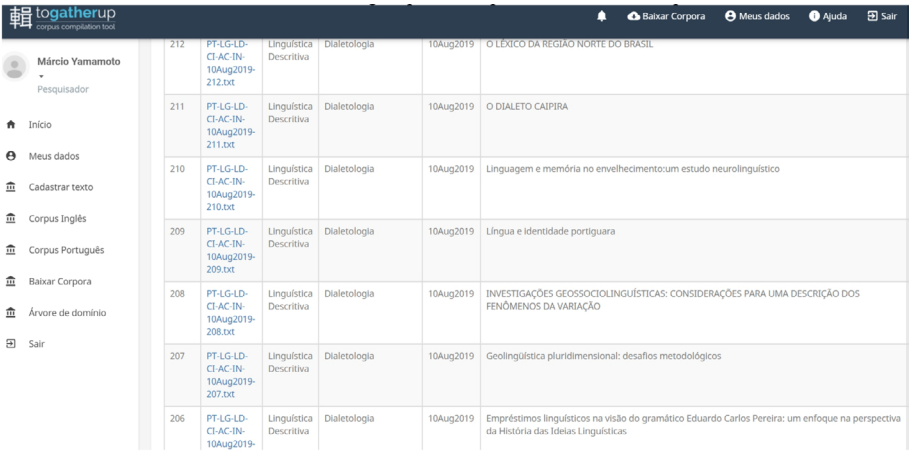
Esses passos metodológicos foram adotados para a organização e o arquivamento dos *corpora* de Linguística, cujo dimensionamento final totalizou 49,89 milhões de palavras. O próximo questionamento foi a forma de arquivar com segurança os *corpora.* Apesar de haver opções on-line, como o Google Drive, da empresa Google, e o OneDrive, da Microsoft, podendo ser gratuitas ou pagas, optamos pela plataforma ToGatherUp (OLIVEIRA, 2019).[28]

Essa plataforma permite que o trabalho terminográfico seja mais prático e simplificado, devido a algumas funções automáticas, tais como, organização, codificação/nomeação dos arquivos, arquivamento de textos em diretórios (personalizado para os projetos), visualização do dimensionamento do *corpus* (quantidade de textos e palavras) e a inserção de cabeçalho e de metadados nos textos. Outras vantagens são a segurança dos dados, a sustentabilidade (economia de *hardware*, economia de espaço), a inovação e a flexibilidade. Há outras funções e vantagens que não trataremos aqui devido ao foco deste artigo, mas que podem ser consultadas de acordo com a referência bibliográfica. Como a escolha dessa plataforma foi posterior ao processo de reciclagem

[28] Plataforma de acesso restrito a pesquisadores, disponível em: http://web-dev.ileel.ufu.br/togatherup/projetos/computacao/mod_login.php. Acesso: 20 maio 2020.

dos *corpora* desta pesquisa, já havíamos desenvolvido nossa própria codificação para os arquivos. Para que o leitor possa ter uma visão dessa plataforma, trazemos as Figuras 9 e 10 e, em seguida, explicaremos seu funcionamento.

FIGURA 9 – Imagem parcial da plataforma ToGatherUp



Fonte: Selecionada pelos autores.

FIGURA 10 – Imagem parcial da plataforma ToGatherUp

O LÉXICO DA REGIÃO NORTE DO BRASIL	http://www.sbpcnet.org.br/livro/61ra/simposios/si_...	8695	00:05:00
O DIALETO CAIPIRA	http://www.letras.ufscar.br/linguasagem/edicao21/p...	263475	00:05:00
Linguagem e memória no envelhecimento:um estudo neurolinguístico	http://www.revistainvestigacoes.com.br/volume-25-N...	29330	00:05:00
Língua e identidade portiguara	http://www.revistainvestigacoes.com.br/volume-22-N...	25176	00:05:00
INVESTIGAÇÕES GEOSSOCIOLINGUÍSTICAS: CONSIDERAÇÕES PARA UMA DESCRIÇÃO DOS FENÔMENOS DA VARIAÇÃO	http://e-revista.unioeste.br/index.php/linguaselet...	35401	00:05:00
Geolingüística pluridimensional: desafios metodológicos	http://celsul.org.br/Encontros/08/geolinguistica_p...	11848	00:05:00
Empréstimos linguísticos na visão do gramático Eduardo Carlos Pereira: um enfoque na perspectiva da História das Ideias Linguísticas	http://www.revistainvestigacoes.com.br/volume-25-N...	29039	00:05:00

Fonte: Selecionada pelos autores.

Objetivando a concisão, descreveremos somente os dados relativos ao registro e arquivamento dos textos. A primeira coluna, à esquerda, de 212-206, contém os números (parciais) dos arquivos carregados no ToGatherUp. Caso não haja a exclusão de nenhum deles, os números são correspondentes à quantidade de textos carregados na plataforma. A segunda coluna mostra a codificação dos arquivos .txt com dados da língua, do nível na árvore de domínio, a data de carregamento etc.; na terceira e na quarta colunas, há a classificação dos textos em Linguística Descritiva ou Aplicada (nível 2), e na subárea do próximo nível da taxonomia, neste caso, a Dialetologia; a quinta coluna registra a data de carregamento do arquivo: 10Aug2019, correspondendo ao dia, mês e ano; a sexta coluna registra o título do texto; a sétima, o endereço da Internet onde o texto se encontra disponível; a oitava coluna registra a quantidade de palavras-ocorrência (*tokens*) do texto (leitura automatizada pela plataforma); e a última coluna registra o tempo necessário utilizado para a busca e o processamento do texto, por parte do pesquisador, antes do carregamento na plataforma.[29]

4.5 A reciclagem

A reciclagem do *corpus* de Linguística do primeiro projeto significa que o pesquisador tem em mãos, como foi no nosso caso, sub*corpora* de **dimensionamento** maior ou menor que o tamanho predeterminado como objetivo da pesquisa no segundo projeto. A partir desse *corpus* existente, o pesquisador o recicla de forma a aumentar ou reduzir os sub*corpora*, até que atinja o dimensionamento adequado ao seu projeto de pesquisa terminográfico em questão, neste caso, de 500 mil *tokens*.

No que diz respeito à **limpeza**, objetivamos, em nosso trabalho, explorar os contextos definitórios e explicativos dos termos, provenientes dos textos acadêmicos na área de Linguística (portanto, vários dados apresentados nos textos não eram relevantes para nossa pesquisa). No tocante à **tipologia** ou ao **gênero dos textos**, observamos que, na análise da composição dos gêneros textuais, havia aqueles que não atendiam ao

[29] Uma das características da plataforma é calcular o tempo gasto nas várias fases de compilação, pré-análise e nomeação dos arquivos. Consideramos importante o pesquisador apresentar, em seu trabalho, uma quantificação de tempo gasto numa compilação de *corpus*, de forma a deixar clara ao leitor a quantificação do tempo para a execução de projetos dessa natureza.

gênero acadêmico, o que os desqualificava para comporem os *corpora*, como, por exemplo, páginas genéricas da Internet, páginas de *blogs* e textos descritivos que pecavam na cientificidade da informação.[30] Apresentamos, na sequência, dados mais detalhados da reciclagem dos *corpora* do primeiro projeto.

4.5.1 Redimensionamento de *corpus*

Em nossa pesquisa terminográfica, o dimensionamento estabelecido como padrão ou ideal foi de 500 mil *tokens* para cada subárea da Linguística; o planejado, no entanto, não se concretizou de maneira uniforme quanto à quantidade de *tokens* levantados pelos alunos: algumas duplas não conseguiram chegar ao número mínimo de *tokens*, já outras se entusiasmaram na coleta e extrapolaram o valor. Essa questão (além do próprio redimensionamento após a limpeza) explica o surgimento de sub*corpora* pequenos ou grandes. Classificamos os sub*corpora* compilados no projeto inicial em quatro tipos diferentes, abaixo relacionados, e mostramos como trabalhamos com cada tipo, no sentido de adequá-los ao *corpus* reciclado de Linguística do novo projeto.

(1) Inexistente: fizemos a compilação dos *tokens* a partir do zero.[31]

(2) Pequeno (menor que 500 mil *tokens*): compilamos o *subcorpus* até que atingisse o dimensionamento de 500 mil *tokens*.

(3) Aproximado: foi mantido como estava ou redimensionado, se após a limpeza (de dados não relevantes) contivesse menos que 500 mil *tokens*.

(4) Grande (maior que 500 mil *tokens*): após a limpeza, fizemos a leitura no WordSmith Tools 7.0 (WST; SCOTT, 2015) e usamos a ferramenta PLOT para eliminar os textos com menor densidade terminológica.[32]

---

[30] Como especialistas na própria área em estudo, tivemos condições de fazer esse tipo de análise. Para o levantamento de *corpora* em outras áreas, por exemplo, teríamos que consultar especialistas nas referidas áreas para uma certificação das fontes de coleta mais relevantes.

[31] Isso significa que uma determinada subárea da Linguística estava prevista na árvore do primeiro projeto, mas não foi feita a compilação para ela ou se trata de uma nova subárea, inserida na árvore de domínio do segundo projeto.

[32] Densidade terminológica (BARROS, 2004) é usada no sentido de identificar textos nos quais havia uma maior ou menor ocorrência de palavras-chave de determinada subárea em específico. Usamos a ferramenta PLOT do WST para identificar os textos

4.5.2 Limpeza do *corpus*

Tendo em vista os objetivos de identificar termos e seus contextos, na limpeza do *corpus* do segundo projeto, eliminamos dados textuais e metadados (como marcação XML adicionada aos textos) que não atendiam ao nosso objetivo: extrair os traços conceituais úteis à construção das definições terminológicas e enciclopédicas do VoBLing[33] a partir dos contextos definitórios e dos contextos explicativos dos termos.

Em relação aos dados textuais eliminados dos artigos, temos: o nome da revista, o volume, a edição, o nome de autores, os metadados dos autores, as notas de rodapé, a bibliografia e os anexos. Quanto às dissertações, teses e trabalhos de conclusão de curso, eliminamos as informações pré-textuais e pós-textuais, tais como: os dados institucionais, os dados do autor, as abreviaturas, o sumário, os agradecimentos, a bibliografia ou as referências bibliográficas, os anexos e os apêndices. Outros dados eliminados foram as citações em língua estrangeira, as notas de rodapé, quadros e tabelas. Esses dados foram eliminados (manualmente), pois interfeririam na contagem de *tokens*, de forma que o dimensionamento dos *corpora* ficaria comprometido e o balanceamento dos *corpora* seria prejudicado.

4.6 O *corpus* reciclado

O dimensionamento dos sub*corpora* de LD e LA foi de aproximadamente 47,7 milhões de *tokens*, provenientes dos gêneros científicos, sem incluir os manuais, pois eles compuseram um *corpus* à parte. Vale ressaltar que o *subcorpus* de Linguística Matemática (LM) foi menor que o das outras subáreas; para esta subárea, na LI, foi possível compilar os 500 mil *tokens*, o que não ocorreu da mesma forma na LP, limitada a 220 mil *tokens*. Logo, os *corpora* de LM totalizaram 440 mil *tokens*, em LP e LI, e não 1 milhão de *tokens*, padrão estabelecido

---

nos quais os 10 termos mais frequentes daquela subárea apresentavam menor frequência. Em seguida, fomos eliminando esses textos até que o número de *tokens* almejado fosse atingido (500 mil).

[33] VoBLing, projeto de vocabulário bilíngue de Linguística, português-inglês. O mesmo se valer de uma nova versão da plataforma já usada (VoTec), com adição de vários campos novos, como nota (que contém uma definição enciclopédica do termo), etimologia, som etc. Disponível em: http://vobling.votec.ileel.ufu.br/

como ideal. Já o dimensionamento dos *corpora* de manuais foi de 2,2 milhões de *tokens* nas duas línguas, já balanceados. Nosso balanceamento não englobou gêneros textuais diversos, senão o acadêmico. Logo, o balanceamento quanto aos gêneros textuais não se fez necessário e não fez parte do desenho dos projetos à exceção do gênero manuais de Linguística. O número de *tokens* no corpus de manuais foi quantificado, pois fez parte do desenho posterior, já do projeto final a fim de que obtivéssemos um maior número de contextos definitórios e explicativos para a extração de traços semânticos úteis à construção das definições dos termos.

A tipologia do *corpus* reciclado é demonstrada no Quadro 2.

QUADRO 2 – Tipologia do *corpus* reciclado

Língua	Bilíngue (inglês e português)
Modo	Escrito (textos acadêmicos: artigos científicos, dissertações e teses; textos instrucionais: manuais)
Data de publicação	Sincrônico (readequação e novos levantamentos realizados entre 2018 e 2020), fechado
Seleção	Amostragem, estático
Conteúdo	Especializado (Linguística)
Autoria	Falantes nativos/não nativos (inglês e português), individual/coletivo
Disposição interna	Comparável
Uso na pesquisa	Estudo (análise terminológica/terminográfica)
Tamanho	Grande (mais de 10 milhões de palavras)
Nível de Codificação	Com cabeçalhos, sem etiquetas

Fonte: Elaborado pelos autores.

Em relação ao tamanho do *corpus* reciclado, podemos verificar os dados na Tabela 2:

TABELA 2 – Tamanho do *corpus* reciclado

	Português	Inglês
Número de textos	1665	1911
Tokens	24.597.756	25.294.698
Types	356.989	356.911
Type/token ratio (TTR)	1,45%	1,41%
TOTAL *tokens*	49.892.454	

Fonte: Elaborada pelos autores.

Percebemos, comparando as Tabelas 1 e 2, que houve uma diminuição na quantidade de *tokens* entre os *corpora* do primeiro e do segundo projetos, embora o *corpus* reciclado do segundo projeto contenha mais subáreas na árvore de domínio (e, consequentemente, no *corpus*) e contenha textos de uma tipologia diversa do primeiro projeto (os manuais). Também percebemos que houve um melhor equacionamento nas TTRs no *corpus* do segundo projeto entre LP e LI, garantindo um melhor balanceamento. Acreditamos que a padronização levou a esse fenômeno.

5 Do *corpus* inicial ao *corpus* final

Nesta seção, compartilhamos algumas reflexões provenientes do percurso da reciclagem dos *corpora*, desde a coleta manual até ao arquivamento. Trataremos de aspectos metodológicos da coleta e do processamento dos *corpora* que optamos por adotar.

Como ponto de partida, é importante refletir sobre os gêneros textuais e sua constituição para um projeto terminográfico. Isso porque, na pesquisa de mestrado de um dos autores, observamos que trabalhar somente com artigos, dissertações e teses não foi suficiente para obtermos contextos definitórios e explicativos com traços semânticos suficientes para a construção de definições terminológicas.

Em segundo lugar, quando se trata da busca do gênero acadêmico na Internet, há arquivos como dissertações e teses de acesso público, cujos direitos autorais são cedidos para fins educacionais ou para o progresso da ciência, contudo encontram-se bloqueados em formato PDF. Nesse

caso, foi necessário usar programas (*software*) ou *sites* da Internet[34] que permitiam que os arquivos fossem salvos em formato txt.

Ainda tratando-se do gênero textual e sua disponibilidade, é importante considerar se esse gênero está disponível em formato digital. Do contrário, será necessário partir de um formato físico impresso até chegar no formato digitalizado. No segundo projeto, foi necessário escanear manuais de Linguística, converter os textos e corrigi-los. O escaneamento ainda é um procedimento moroso, que demanda uma quantidade considerável de tempo. Em seguida, como os manuais foram salvos em formato de imagem, foi necessário decodificá-lo para o formato de textos, usando programas específicos[35] e, em geral, pagos. Após a decodificação para textos, restou a correção textual, já que a maioria dos programas de OCR não decodifica a acentuação e diacríticos da língua portuguesa corretamente. Em nossa pesquisa, utilizamos o Microsoft Word para correção textual e salvamos os arquivos em formato .txt para o processamento pelo WST.[36]

Finalmente, há que considerar a codificação do arquivo txt, pois o WST requer que todos os arquivos sejam da mesma codificação para o correto processamento textual. Em um trabalho colaborativo, pode ocorrer que os membros salvem os arquivos em formato .txt, porém com codificações diferentes como UTF-8, UTF-16 ou ANSI, ocasionando a incompatibilidade de leitura. Para solucionar esse problema, utilizamos o utilitário *Text Converter* do WST para padronização dos arquivos em formato UTF-16 LE (segundo recomendação do manual do programa).

---

[34] Exemplo de *sites* para manuseio de arquivos PDF: https://freemypdf.com/, https://www.ilovepdf.com/.

[35] Neste projeto, usamos o OCR OmniPage (https://www.kofax.com/Products/omnipage), a versão paga do Adobe Acrobat (https://www.adobe.com/br) e o ABBYY Screenshot Reader (https://www.abbyy.com/pt-br/).

[36] Na fase de coleta do *corpus* inicial, no primeiro projeto, tanto o WST quanto o AntConc foram usados pelos alunos para verificar o tamanho dos sub*corpora* levantados e, na sequência dos trabalhos, para elaborar um projeto terminográfico. A partir do segundo projeto, de reciclagem do *corpus*, apenas o WST foi usado, devido à robustez computacional e à possibilidade de salvar os resultados.

6 Considerações finais

Dez anos de trabalhos contínuos na elaboração e na reelaboração de *corpora* na área de Linguística nos trouxeram uma rica experiência. Do começo, quando éramos novatos na universidade e a LC era coisa para *nerds*, até o presente, quando encontramos mais colegas trabalhando com a LC e os alunos escolhendo nossas disciplinas de graduação e pós-graduação por causa da LC, o panorama foi se alterando.

Nesse período, formamos (os autores e colegas da universidade) toda uma geração de alunos que consegue pensar em trabalhos de pós-graduação usando a abordagem e a metodologia da Linguística de *Corpus*. Uma geração que já começa a pensar em não mais apenas consumir *software* de análise lexical, mas também desenvolver *software* específicos para suas pesquisas e que trabalha, ao mesmo tempo, com descrição e aplicação (em sala de aula).

A elaboração de uma árvore de domínio da Linguística, baseada em *corpus*, nos trouxe um conhecimento geral da área que, até hoje, poucos colegas têm; o conhecimento de retrabalhar continuamente a compilação das fontes textuais de uma ciência também nos proporcionou autoridade para comentar sobre a taxonomia dessa ciência.

Para análises linguísticas, notamos que nosso *corpus* colaborativo (que se constituiu como estático),[37] assim como *corpora* advindos de qualquer área da ciência, deveria ter um caráter monitor: esse *corpus* colaborativo pode ter sua compilação (e, consequentemente, sua árvore de domínio) sempre reavaliada e ampliada. Novas funcionalidades, como as etiquetagens (começando com a morfossintática) e a disponibilização desse *corpus* em plataforma eletrônica (com as ferramentas básicas da LC) para pesquisas online, devem ser agregadas para que esse *corpus* seja sempre útil para atuais e futuros pesquisadores.

Por fim, gostaríamos de agradecer, imensamente, às dezenas de alunos de graduação, pós-graduação e iniciação científica que nos ajudaram a tornar o *corpus* do primeiro projeto uma experiência didática e de pesquisa muito importante para aprimorar nossa formação como professores e pesquisadores.

[37] Estático porque seu desenho não pressupunha um projeto contínuo. Cada vez que o *corpus* foi trabalhado, o mesmo era reaberto e encerrado ao final do período, sem expectativa de nova atualização.

Contribuição dos autores

O primeiro autor (FROMM) foi o responsável pela descrição do *corpus* original, o segundo autor (YAMAMOTO) foi o responsável pela descrição do segundo *corpus*. O restante do texto foi escrito e revisto pelos dois autores.

Referências

ALBUQUERQUE, D. B. de. Múltiplos olhares em Linguística e Linguística Aplicada, 2016. *Ecolinguística*: Revista Brasileira de Ecologia e Linguagem (ECO-REBEL), Brasília, v. 3, n. 1, p. 227-237, 2017. Disponível em: https://periodicos.unb.br/index.php/erbel/article/view/26263/23021 Acesso em: 26 mar. 2020.

ALVES, I. M. *et al*. (org.). *Estudos lexicais em diferentes perspectivas*. São Paulo: FFLCH/USP, 2010.

AUBERT, F. H. *Introdução à metodologia da pesquisa terminológica bilíngue*. 2. ed. São Paulo: FFLCH/CITRAT, 2001.

BARBOSA, M. A. Dicionário, vocabulário, glossário: concepções. *In:* ALVES, I. M. (org.). *A constituição da normalização terminológica no Brasil*. 2. ed. São Paulo: FFLCH/CITRAT, 2001. p. 23-45.

BARROS, L. A. *Curso básico de terminologia*. São Paulo: EDUSP, 2004.

BERBER SARDINHA, T. *Linguística de Corpus*. São Paulo: Manole, 2004.

BOWKER, L. Towards a Collaborative Approach to Corpus Building in the Translation Classroom. *In:* BAER, J. B.; KOBY, G. S. (ed.). *Beyond the Ivory Tower*: Rethinking Translation Pedagogy. Amsterdam; Philadelphia: John Benjamins, 2003. p. 193-210.

CABRÉ, M. T. Hacia una teoría comunicativa de la terminología: aspectos metodológicos. *In:* CABRÉ, M. T. *La Terminologia*: representación y comunicación. Barcelona: IULA, 2000. p. 129-150.

CARDOSO, S. A. F. *TermosTeo*: a elaboração de vocabulários monolíngues de termos da Teologia em um estudo conduzido por *corpus*. 2017. 340f. Tese (Doutorado em Estudos Linguísticos) – Instituto de Letras e Linguística, Universidade Federal de Uberlândia, Uberlândia, 2017.

FROMM, G. Vocabulário de linguística: treinamento em terminografia bilíngue, uso de corpora e ambiente de gestão terminológica. *In*: ISQUERDO, A. N.; DAL CORNO, G. O. M. (org.). *As ciências do léxico*: lexicologia, lexicografia, terminologia. Campo Grande: Editora UFMS, 2018. p. 309-328.

FROMM, G. Vocabulário de Linguística: treinamento em Terminografia Bilíngue, uso de corpora e ambiente de gestão terminológica. *In:* ENCONTRO INTERMEDIÁRIO DO GT DE LEXICOLOGIA, LEXICOGRAFIA E TERMINOLOGIA DA ANPOLL, 10., 2015, Rio de Janeiro. *Anais* [...]. Rio de Janeiro: ANPOLL, 2015. p. 1-5.

FROMM, G. A questão da taxonomia num *corpus* colaborativo para construção de um vocabulário na área de linguística. In: SIMPÓSIO INTERNACIONAL DE LETRAS E LINGUÍSTICA - SILEL, 2013, Uberlândia. *Anais* [...]. Uberlândia: EDUFU, 2013. p. 1-9.

FROMM, G. Ensino de terminologia: trabalhando com site e banco de dados. *Debate Terminológico*, Porto Alegre, v. 6, p. 2-22, 2010.

FROMM, G. VoTec: a construção de vocabulários eletrônicos para aprendizes de tradução. 2007. 215f. Tese (Doutorado em Estudos Linguísticos e Literários em inglês) – Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo, 2007. Disponível em: https://www.teses.usp.br/teses/disponiveis/8/8147/tde-08072008-150855/publico/TESE_GUILHERME_FROMM.pdf. Acesso em: 30 mar. 2020.

FROMM, G.; YAMAMOTO, M. I. Terminologia, Terminografia, Tradução e Linguística de *Corpus*: a criação de um vocabulário bilíngue sobre Linguística. *In*: TAGNIN, S.; BEVILACQUA, C. (org.). *Corpora na Terminologia*. São Paulo: Hub Editorial, 2013. p. 129-152.

FROMM, G.; YAMAMOTO, M. I. A microestrutura em verbetes da área da Linguística. *Revista Estudos da Linguagem*, Belo Horizonte, v. 28, n. 1, p. 205-234, 2020. DOI: http://dx.doi.org/10.17851/2237-2083.28.1.205-234. Disponível em: http://www.periodicos.letras.ufmg.br/index.php/relin/article/view/15255. Acesso em: 23 mar. 2020.

GARDNER, J.; KROWNE, A.; XIONG, L. Automatic Invocation Linking for Collaborative Web-Based Corpora. *In*: CHBEIR, R.; BADR, Y.; ABRAHAM, A. (org.). *Emergent Web Intelligence*: Advanced Semantic Technologies. Londres: Springer-Verlag, 2010. p. 23-45.

ILARI, R. *Introdução ao estudo do léxico:* brincando com as palavras. 2. ed. São Paulo: Contexto, 2003.

MELLO, H. Methodological Issues for Spontaneous Speech *Corpora* Compilation: The Case of C-ORAL-BRASIL. *In:* RASO, T.; MELLO, H. (ed.). *Spoken Corpora and Linguistic Studies*. Amsterdam: John Benjamins Publishing, 2014. p. 27-68. DOI: https://doi.org/10.1075/scl.61.01mel

OLIVEIRA, F. P. de. *ToGatherUp*: um protótipo de ferramenta para a construção de *corpora*. 2019. 219f. Dissertação (Mestrado em Estudos Linguísticos) – Instituto de Letras e Linguística, Universidade Federal de Uberlândia, 2019. DOI: http://dx.doi.org/10.14393/ufu.di.2019.679.

PÉREZ-PAREDES, P. L.; SÁNCHEZ-TORNEL, M.; CALERO, J. A. M. Learners' Search Patterns During *Corpus*-Based Focus-on-Form Activities. *International Journal of Corpus Linguistics*, [*S.l.*], v. 17, n. 4, p. 482-515, 2012. DOI: https://doi.org/10.1075/ijcl.17.4.02par

SCOTT, M. *WordSmith Tools*. Version 7. Liverpool: Lexical Analysis Software, 2015.

SINCLAIR, J. *Reading Concordances*: An Introduction. London: Longman, 2003.

TAGNIN, S.; BEVILACQUA, C. *Corpora na Terminologia*. São Paulo: Hub Editorial, 2013.

TAGNIN, S. E. O. *Corpus*-Driven Terminology in Brazil. *In:* POUPET, A. L. B.; XATARA, C. (org.). *Cahiers de Lexicologie* – Dynamique de la Recherche en Lexicologie, Lexicographie et Terminologie au Brésil. Paris: Classiques Garnier, 2012. p. 169-182.

TEIXEIRA, E. D. *A Linguística de Corpus a serviço do tradutor*: proposta de um dicionário de culinária voltado para a produção textual. 2008. 439f. Tese (Doutorado, Programa de Pós-Graduação em Estudos Linguísticos e Literários em Inglês) – Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, 2008.

TIEDEMANN, J. *Recycling Translations*: Extraction of Lexical Data from Parallel *Corpora* and their Application in Natural Language Processing. Uppsala: Acta Universitatis Upsaliensis, 2003.

VARANTOLA, K. Disposable *Corpora* as Intelligent Tools in Translation. *Cadernos de Tradução*, Florianópolis, v. 1, n. 9, p. 171-189, 2002. Disponível em: https://periodicos.ufsc.br/index.php/traducao/article/view/5985/5689. Acesso em: 15 maio 2020.

VIANA, V.; TAGNIN, S. E. O. (org.). *Corpora na Tradução*. São Paulo: Hub Editorial, 2015.

YAMAMOTO, M. I. *Linguística histórica e Linguística de corpus:* caminhos que se cruzam para desvelar a história da linguagem: um vocabulário bilíngue português-inglês. 2015. 118f. Dissertação (Mestrado em Linguística Letras e Artes) – Instituto de Letras e Linguística, Universidade Federal de Uberlândia, 2015. Disponível em https://repositorio.ufu.br/handle/123456789/15483. Acesso em: 15 mar. 2020.

YAMAMOTO, M. I. Vocabulário bilíngue português/inglês de linguística geral. *Revista Philologus*,v. 24, p. 272-297, 2018. disponível em http://www.filologia.org.br/rph/ano24/70supl/023.pdf . Acesso em: 15 mar. 2020.