http://dx.doi.org/10.1590/1678-460X2020360209

O *Corpus* de Português Escrito em Periódicos - CoPEP

The Corpus *of Portuguese from Academic Journals*

Tanara Zingano Kuhn[1]
José Pedro Ferreira[2]

RESUMO

O presente estudo tem como objetivo descrever os desafios e soluções encontrados na compilação do Corpus de Português Escrito em Periódicos - CoPEP, que contém aproximadamente 40 milhões de palavras, é equilibrado entre as variedades português brasileiro e português europeu em número de palavras e cobre seis grandes áreas de conhecimento. Primeiramente, apresentaremos o contexto de criação do CoPEP, qual seja, a elaboração de um dicionário on-line de português para universitários, para o qual serviu como fonte primária de obtenção de evidências linguísticas. Assim, foram as características desse projeto lexicográfico que informaram os critérios de criação do desenho do CoPEP e as consequentes tomadas de decisão. A seguir, descreveremos a metodologia de aquisição de dados, com foco especial nos desafios enfrentados e nas soluções encontradas. Terminaremos com a descrição da

1. Centro de Estudos de Linguística Geral e Aplicada (CELGA-ILTEC), Universidade de Coimbra – Portugal. http://orcid.org/0000-0003-2640-5500. E-mail: tanarazingano@outlook.com.
2. Centro de Estudos de Linguística Geral e Aplicada (CELGA-ILTEC), Universidade de Coimbra – Portugal. https://orcid.org/0000-0003-0593-5043. E-mail: jpferreira@gmx.com.

fase final de compilação, na qual aplicamos uma série de procedimentos para obtenção de equilíbrio.

Palavras-chave: *corpus multivariedade, compilação de corpus, discurso acadêmico, língua portuguesa*

ABSTRACT

The present study aims to describe the challenges faced and solutions found in the compilation of the Corpus de Português Escrito em Periódicos - CoPEP, which contains approximately 40 million words, is balanced between the Brazilian Portuguese and European Portuguese varieties in number of words and covers six large areas of knowledge. Firstly, we will present the context of the creation of CoPEP, namely, the make of an on-line dictionary of Portuguese for university students, to which CoPEP served as the primary source for linguistic evidence extraction. Thus, it was the characteristics of this lexicographic project that informed the design criteria for CoPEP and the consequent decision-making process. Next, we will describe the methodology of data acquisition, with a special focus on the challenges that were faced, and the solutions found. We will conclude with the description of the final compilation phase, which involved procedures for obtaining balance.

Keywords: *multivariate corpus, corpus compilation, academic discourse, Portuguese language.*

Introdução

O objetivo deste estudo é descrever os desafios e soluções encontrados na compilação do *Corpus* de Português Escrito em Periódicos – CoPEP, que contém aproximadamente 40 milhões de palavras, é equilibrado entre as variedades português brasileiro e português europeu em número de palavras e cobre seis grandes áreas de conhecimento. Foi criado para servir como fonte primária para a elaboração do Dicionário *On-line* de Português para Universitários (doravante, DOPU) [3] (Kuhn, 2017), uma vez constatada a inadequação, para o desenvolvimento

3. Em construção.

DELTA
36.2
2020

desse projeto lexicográfico, dos *corpora* de português existentes, como se mostrará abaixo.

Tendo em vista que o processo de criação de um *corpus* está inteiramente condicionado pelo seu propósito, pois é este que define os critérios de compilação e justifica as decisões que serão tomadas (Summers 1993; Meyer 2002), é importante, antes de dar início à descrição da compilação do CoPEP, apresentar mais informações a respeito do contexto de sua origem.

No projeto DOPU, foi adotada, pela primeira vez para a língua portuguesa, a abordagem semiautomatizada para criação de dicionários (cf. Gantar et al. 2016). Esta consiste na extração automática de dados do *corpus* e importação para o sistema de escrita de dicionários. Com esse método, o ponto de partida do lexicógrafo para a escrita de verbetes não é mais o *corpus* e as linhas de concordâncias nem os resultados de *word sketches* (como adotado para a elaboração do *Macmillan English Dictionary for Advanced Learners*, por exemplo; ver Kilgarriff e Rundell 2002), mas sim as entradas, diretamente no sistema de escrita de dicionários, que se encontram parcialmente pré-preenchidas com informações extraídas de forma automática do *corpus*, cabendo ao lexicógrafo apenas fazer análises, alterações e edições.

Como se vê, o *corpus* tem um papel central em projetos lexicográficos que adotam a abordagem semiautomatizada. Quanto mais adequado às necessidades do projeto for o *corpus*, mais qualidade terá o dicionário. Nesse sentido, de forma a estabelecer os atributos indispensáveis a um *corpus* que atendesse as demandas do projeto DOPU da maneira mais completa possível, foi necessário explicitar as funções do DOPU e as características do perfil do seu potencial usuário.

O DOPU será um dicionário que deverá servir de auxílio pedagógico a alunos de graduação e pós-graduação de cursos de diferentes áreas em instituições cuja língua de instrução é o português, seja variedade brasileira ou portuguesa[4], usado como língua materna ou como língua adicional, em tarefas de compreensão e produção. Além disso, o DOPU será personalizável, satisfazendo as demandas de diferentes perfis em um único recurso digital.

4. Espera-se incorporar outras variedades no futuro.

Assim, concluímos que o *corpus* ideal para a elaboração do DOPU deveria ser composto por textos de gêneros acadêmicos, produzidos por autores experientes de diversas áreas de conhecimento e escritos em português brasileiro e europeu. Deveria também conter metadados com informações que permitissem buscas refinadas no *corpus* e descrições detalhadas das evidências lexicais obtidas.

O primeiro passo no desenvolvimento do projeto DOPU foi examinar os *corpora* portugueses existentes contendo textos acadêmicos e determinar sua adequação para esse projeto. De muitos *corpora* de português existentes, que abrangem diferentes variedades de língua, registros e gêneros, apenas alguns contêm textos acadêmicos. Entre estes, nenhum reúne todas as características mencionadas acima, como se vê na **Tabela 1**[5]. Consequentemente, foi decidido compilar um novo *corpus* de textos acadêmicos especialmente para esse projeto: o *Corpus* de Português Escrito em Periódicos - CoPEP.

5. Uma versão em inglês desta tabela foi publicada em Kuhn & Kosem (2016).

Tabela 1 - Análise de adequação de *corpora* de português com textos acadêmicos

***Corpus* e autor(es)**	**Tamanho**	**Características**	**Razões para não adequação aos propósitos do projeto DOPU**
Portuguese Web 2011 (ptTenTen, Palavras parsed) Autor: Equipe *Sketch Engine*	2.757.635.105 palavras*	Textos de *sites* de natureza acadêmica / científica (universidades, periódicos, governamentais, repositórios de teses, etc.). Etiquetado pelo *parser* PALAVRAS (Bick 2000).	Metadados cruciais como fonte (tipo de publicação: periódico, livro, tese, etc.), ano de publicação e área de conhecimento não estão disponíveis. Não há possibilidade de medir a qualidade da escrita e composição do *corpus*.
Portuguese Web 2011 (ptTenTen, Freeling v3) Autor: Equipe *Sketch Engine*	3.900.501.097 palavras	Textos de *sites* com natureza acadêmica / científica (universidades, periódicos, governamentais, repositórios de teses, etc.) Etiquetada por *Freeling 3.0* (Padró e Stanilovsky 2012)	Metadados cruciais como fonte (tipo de publicação), ano de publicação, área de conhecimento e variedade de língua não estão disponíveis. O país do *website* é equivalente à variedade do português, o que não é uma abordagem precisa para determinação de informações tão relevantes para o projeto. Não há possibilidade de medir a qualidade da escrita e composição do *corpus*.
Corpus Araneum Portugallicum Maius (Portuguese, 15.05) Autor: Vladimír Benko	862.134.902 palavras	Textos de *sites* de natureza acadêmica / científica (universidades, periódicos, governamentais, repositórios de teses, etc.). Para ser usado para linguística contrastiva e projetos lexicográficos bilíngues.	Metadados cruciais como fonte (tipo de publicação: periódico, livro, tese, etc.), ano de publicação e área de conhecimento não estão disponíveis. Não há possibilidade de medir a qualidade da escrita e composição do *corpus*.
Corpus Brasileiro Autor: Tony Berber Sardinha (coordenador)	1.133.416.757 *tokens*	*Corpus* geral do português brasileiro. O *subcorpus* acadêmico contém 258.585.002 *tokens* de artigos, 310.972.387 *tokens* de teses e dissertações e 6.947.244 *tokens* de anais.	Metadados cruciais, como ano de publicação e área de conhecimento, não estão publicamente disponíveis. Nenhuma informação sobre a qualidade dos textos que compõem o *subcorpus* acadêmico. Apenas português do Brasil.
Corpus do Português (Genre/ historical version) Autores: Mark Davies e Michael Ferreira	45 milhões de palavras	Textos dos anos 1300 aos 1900. Os textos dos anos 1900 perfazem 20 milhões de palavras, com equilíbrio entre os gêneros acadêmico, ficcional, falado e jornal. Seu *subcorpus* acadêmico é composto por 3.087.052 palavras de Portugal e 2.816.802 do Brasil.	O *subcorpus* acadêmico é composto por entradas retiradas de enciclopédias *on-line* brasileiras e portuguesas.
CPBA – Corpus do Português Brasileiro Acadêmico Autores: Grupo de pesquisa UPLA, coordenado por Cristina Becker Lopes Perna, na PUCRS	22.777.993 *tokens* (Peixoto 2015:44)	Livros e periódicos de seis diferentes áreas do conhecimento fornecidos por oito universidades brasileiras, compreendendo produções escritas de professores e alunos de graduação e pós-graduação.	Não disponível publicamente. Apenas português do Brasil.
CRPC - Corpus de Referência do Português Contemporâneo Autores: Desenvolvido no Centro de Linguística da Universidade de Lisboa (CLUL).	311 milhões de palavras (falado + escrito) Aproximadamente 310 milhões de palavras de textos escritos	*Corpus* de linguagem geral. Português europeu e outras variedades (Brasil, Angola, Cabo Verde, Guiné-Bissau, Moçambique, São Tomé e Príncipe, Goa, Macau e Timor-Leste). Diferentes tipos de texto, incluindo científicos. Textos da segunda metade do século XIX a 2008.	Metadados não estão consistentemente disponíveis.

* Nesta tabela, palavras, *tokens* ou ambos são usados ao fornecer informações sobre o tamanho do corpus, dependendo das informações disponíveis.

Este texto apresenta-se dividido em três partes. A primeira descreve o processo de compilação do CoPEP, que envolveu elaboração do seu desenho; levantamento de fontes; e extração, limpeza, nomeação dos arquivos e obtenção de equilíbrio entre os *subcorpora*. Aqui destacamos os desafios que enfrentamos no que tange às determinações de critérios e justificamos as decisões tomadas. Já na segunda parte, apresentamos o CoPEP e suas características. Encerramos este trabalho com algumas considerações finais.

1. Compilação de *corpora*: o que diz a Linguística de *Corpus*

De acordo com a metodologia da Linguística de *Corpus* (Biber et al. 1998; Meyer 2002; Wynne 2005; McEnery et al. 2006; Sinclair 2003a, 2003b), estes são os fatores a serem considerados na construção de um *corpus*: tamanho do *corpus*; equilíbrio e representatividade; limpeza dos dados; codificação de caracteres; marcação de *corpus* e anotação de *corpus*.

A definição do tamanho do *corpus* pode seguir duas perspectivas. Uma opção é estabelecer o número total de palavras requeridas para o *corpus* ser usado para determinado propósito. Assim, por exemplo, foi decidido que os *corpora* de variedades portuguesas deveriam totalizar pelo menos 30 milhões de *tokens* (Almeida et al. 2013) no projeto que criou o Vocabulário Ortográfico Comum da Língua Portuguesa (VOC). Outra abordagem consiste em estabelecer que o tamanho do *corpus* será definido consoante a quantidade de textos disponíveis representantes da fatia de linguagem de interesse.

O equilíbrio do *corpus* refere-se à distribuição de textos de acordo com o tipo de textos que compõem o *corpus*. Em um *corpus* de vários gêneros, por exemplo, pode ser decidido que a distribuição de textos deve ser equilibrada entre eles. Outra possibilidade, em *corpora* que cobrem várias variedades de um idioma, é determinar um número igual de *tokens* ou de textos em cada *subcorpus* de variedade.

A representatividade diz respeito à seleção de amostras de textos que representam a fatia da linguagem em estudo. Por exemplo, o

DELTA
36.2
2020

British National Corpus (BNC) foi construído para representar a língua inglesa contemporânea usada na Grã-Bretanha. Assim, amostras de textos foram coletadas de periódicos acadêmicos, jornais, revistas, livros de literatura, folhetos, entre outros, para representar o inglês escrito. Para o inglês falado, muitas horas de entrevistas, conversas telefônicas, palestras, etc. foram transcritas e incluídas no *corpus*. Embora o BNC não seja equilibrado (por exemplo, apenas 10% são textos orais), certamente tentou ser representativo.

Após esse primeiro estágio, e para preparar o próximo (anotação de *corpus*), o *corpus* precisa ser automaticamente e / ou manualmente limpo. A extensão dos elementos a serem limpos, mais uma vez, depende da finalidade do *corpus*. Para fins lexicográficos, como a criação de dicionários, é comum manter apenas informações textuais, portanto, figuras, tabelas e gráficos geralmente são removidos.

É aconselhável trabalhar na codificação de caracteres antes da limpeza, pois qualquer transformação problemática pode ser localizada e, se for generalizada, a conversão de codificação pode ser revisada. Atualmente, a codificação UTF-8 tem sido amplamente utilizada em uma série de ferramentas de *corpus*.

A próxima fase é a marcação de *corpus* e implica a adição de marcas aos textos, a fim de contribuir para a análise avançada do *corpus*. Para tanto, é comum os metadados serem incluídos nos textos através de cabeçalhos. A totalidade e o tipo de informação contida nos cabeçalhos dependem de quantos dados externos foram registrados ao se coletarem os textos.

Finalmente, a anotação de *corpus* refere-se à adição de etiquetas a todas as palavras do *corpus*, a fim de permitir várias formas de análise. Esta fase consiste em vários processos de identificação e transformação, nomeadamente, *tokenização*, lematização e etiquetação.

Um *tokenizador* (*tokenizer*) percorre o texto e o separa em *tokens*, onde um *token* corresponde a qualquer sequência de caracteres entre dois espaços. Isso significa que os *tokens* incluem não só palavras, mas também, por exemplo, algarismos, símbolos e pontuação.

Outra ferramenta, o lematizador (*lemmatizer*), analisa os *tokens*, identifica sua forma e os transforma em formas de citação de palavras.

Por exemplo, o *token estudos* (substantivo) está no plural; o lematizador o transforma em *estudo*. O mesmo se aplica aos verbos e adjetivos flexionados, sendo o primeiro convertido na forma infinitiva, enquanto o último é apresentado nas formas masculina e singular.

Os etiquetadores de partes do discurso (*part-of-speech tagger* ou *POS-tagger*) são ferramentas que identificam *tokens* e atribuem etiquetas com informações morfossintáticas. Deve-se atentar para o fato de que essas etiquetas não apenas informam a classe da palavra, por exemplo, substantivo, verbo, advérbio, etc., mas também mostram o tipo de flexão da palavra em particular. As etiquetas assumem a forma de códigos e, para interpretá-las, os lexicógrafos devem consultar o conjunto de etiquetas do etiquetador usado para a anotação do *corpus*.

No projeto DOPU, no qual a ferramenta de *corpus Sketch Engine* (Kilgarriff et al. 2004) desempenhou um papel central, a anotação do CoPEP foi realizada no próprio *Sketch Engine* pelo etiquetador padrão disponível para os *corpora* de língua portuguesa, que é o *Freeling v3* (Padró e Stanilovsky 2012).

É possível também atribuírem-se etiquetas conforme as relações sintáticas entre as palavras em uma frase; nesse caso, a ferramenta empregada é um *parser*.

2. Estabelecimento de critérios de compilação

Conforme informado anteriormente, o CoPEP deveria fornecer informação lexical para o desenvolvimento do DOPU. A **Tabela 2** apresenta as características linguísticas que o DOPU deverá cobrir e os critérios de compilação do *corpus* que permitem atingir esses objetivos:

DELTA
36.2
2020

Tabela 2 – Estabelecimento de critérios de compilação

Informação lexical no DOPU	**Critérios de compilação do *corpus***
Modelo do uso típico do português em contextos acadêmicos escritos	Fonte: textos acadêmicos de qualidade
Português brasileiro (PB) e português europeu (PE)	Equilíbrio: 50% de textos portugueses e 50% de textos brasileiros
Representante de várias disciplinas	Cobertura: diferentes áreas do conhecimento
Uso da linguagem corrente	Período: sincrônico
Nomenclatura exaustiva Verbetes completos	Tamanho: o maior possível

Certas condições do nosso contexto de trabalho também contribuíram para a definição de critérios de ordem administrativa. Por exemplo, foi preciso encontrar alternativas para se superarem as seguintes restrições: a) falta de uma equipe de profissionais para o desenvolvimento do projeto DOPU (conforme Klosa (2013), projetos lexicográficos requerem a participação de cientistas computacionais, linguistas de *corpus*, lexicógrafos); b) inexistência de verba para contrato de pessoal, aquisição de tecnologias computacionais (*hardware* e *software*) e compra de direitos autorais; c) tempo limitado para aquisição de dados, manipulação de problemas de direitos autorais e digitalização - caso este último método fosse adotado.

Foi encontrada uma solução que pareceu servir tanto para os critérios de compilação textual do *corpus* como para as condições de trabalho de criação do *corpus*: extrair textos de periódicos *on-line* gratuitos e revisados por pares publicados no Brasil e em Portugal. Uma fonte conhecida e confiável que atende a todas essas condições é a plataforma SciELO (*Scientific Eletronic Library On-line*), por isso foi decidido que todos os textos em português de todas as revistas em cada coleção nacional da SciELO seriam extraídos.

É importante mencionar, no entanto, que logo se descobriu que a coleção brasileira SciELO é muito maior que a coleção portuguesa. A principal consequência de tal diferença de tamanho foi que, como um dos critérios para a construção do *corpus* foi o equilíbrio entre PB e PE, o tamanho total do *corpus* foi determinado pelo menor grupo de textos, isto é, de Portugal. Como McEnery et al. (2006:71) explicaram: “Na

construção de um *corpus* equilibrado de acordo com proporções fixas (...) a falta de dados para um tipo de texto pode restringir o tamanho das amostras de outros tipos de texto" (tradução nossa).

3. A fonte dos textos

A *Scientific Electronic Library On-line* (SciELO) é uma base de dados de periódicos científicos de acesso aberto com coleções nacionais de 15 países, incluindo o Brasil e Portugal.

Dado o papel fundamental da fonte dos textos na elaboração de um dicionário orientado por *corpus*, a SciELO foi considerada ideal para este projeto devido ao completo cumprimento de todos os requisitos mencionados no início deste texto. Por um lado, a SciELO abriga uma grande quantidade de revistas científicas atuais e arbitradas do Brasil e de Portugal, cobrindo diferentes áreas temáticas, fornecendo, assim, grandes quantidades de amostras de escrita especializada de ambos os países em assuntos variados. Por outro lado, é gratuita, *on-line*, de acesso aberto e possui uma estrutura organizacional comum a todas as coleções nacionais, facilitando a identificação e extração de metadados textuais e balanceamento de *corpus*.

A seguir, nos debruçamos sobre cada um dos critérios para a criação deste *corpus* e explicamos em que medida a SciELO atende a esses requisitos, destacando os pontos fortes dessa fonte.

a) O *corpus* deve ser composto por textos acadêmicos escritos que retratam um uso exemplar da linguagem: O modelo SciELO possui critérios muito rigorosos para admissão e retenção de periódicos científicos em suas coleções nacionais, como conteúdo científico, processo de revisão por pares, uso dos periódicos e fator de impacto, o que, em teoria, implica na publicação de documentos muito bem escritos e de alta qualidade. Parece plausível concluir que esses textos retratam um uso exemplar da linguagem.

b) O *corpus* deve ser equilibrado em termos de variedades do português - 50% em português brasileiro, 50% em português europeu - e cobrir diferentes áreas do conhecimento: Como mencionado anteriormente, a SciELO hospeda coleções brasileiras e portuguesas

DELTA
36.2
2020

de periódicos, que consequentemente estão sujeitas aos mesmos critérios rígidos de admissão e retenção[6] explicados acima. Além disso, a infraestrutura organizacional comum seguida pela rede SciELO tem implicações importantes para o desenho do *corpus* pelo menos de duas maneiras. Em primeiro lugar, os periódicos em todas as coleções seguem a mesma classificação de áreas temáticas, o que, por um lado, confere um caráter objetivo à delicada questão da classificação das áreas do conhecimento[7] e, por outro, facilita a construção de um *corpus* equilibrado, não apenas entre variedades da língua portuguesa, mas também no que diz respeito à distribuição igual entre as áreas. Em segundo lugar, uma estrutura comum para a organização da rede SciELO implica a adoção de marcações de texto e estrutura semelhantes em cada página nacional. Assim, as operações computacionais necessárias para a identificação de metadados de texto, extração de textos, conversão, homogeneização e limpeza devem ser aplicáveis tanto para a coleção de periódicos SciELO brasileira (www.scielo.br, doravante, SciELO-Br), quanto para a portuguesa (www.scielo.mec.pt, doravante SciELO-Pt), agilizando o processo de compilação de *corpus*.

c) Sincrônico: Outra vantagem é o caráter contemporâneo da SciELO. Na SciELO-Pt, as datas de publicação dos periódicos variam de 1997 até o presente, enquanto na SciELO-Br, que hospeda alguns conjuntos completos de coleções de periódicos, as datas vão de 1909 até o presente, com a maior parte das publicações de 1998 em diante. Dada essa característica, a SciELO possui as condições ideais para fornecer textos para se fazer um *corpus* sincrônico.

d) Grande em tamanho: Além de todos os benefícios acima, uma vez que a SciELO-Br tem 345 periódicos e a SciELO-Pt, 55[8], é possível prever a construção de um *corpus* grande. Apesar da redução do

6. Ver SciELO-Pt: http://www.scielo.mec.pt/avaliacao/avaliacao_en.htm e SciELO-Br: http://www.scielo.br/avaliacao/avaliacao_en.htm.

7. A categorização de áreas de conhecimento normalmente varia entre as plataformas de indexação, bibliotecas e universidades, tornando muito difícil para os compiladores de *corpora* decidir como alocar textos nessas áreas. O fato de a SciELO adotar um único conjunto de assuntos sob os quais os periódicos são integrados minimiza a necessidade de emprego de subjetividade no processo de classificação de textos em nosso novo *corpus*, reduzindo assim o risco de problemas futuros com a determinação de etiquetas de áreas de conhecimento para itens lexicais no DOPU.

8. Na época da compilação, de fevereiro a agosto de 2016.

número total de periódicos utilizáveis para aproximadamente 110 em função do requisito de equilíbrio entre as variedades, uma contagem inicial aproximada de palavras estimou o tamanho do *corpus* como contendo 45 milhões de palavras. Quando comparado com os *corpora* em geral, pode-se dizer que este é um número pequeno. No entanto, como a SciELO atende plenamente os requisitos de desenho do *corpus*, que como se viu, são muito rigorosos, optamos por usar as duas coleções nacionais como as únicas fontes de textos, o que limitou o tamanho do *corpus*, mas garantiu sua qualidade.

4. Descrição do processo de compilação

O processo aqui descrito ocorreu entre fevereiro e agosto de 2016. O passo inicial foi obter informações detalhadas sobre os periódicos e textos nas duas coleções nacionais da SciELO para que pudessem ser tomadas decisões sobre como continuar o processo de compilação de acordo com as etapas descritas na seção 1 acima.

4.1. Examinando as fontes

A primeira etapa envolveu a identificação dos periódicos de cada coleção nacional. Assim, a extração automática dos títulos de cada periódico, identidades únicas ISSN (*International Standard Serial Number*)[9], número de edições por periódico e idioma de publicação (disponíveis como metadados) nos permitiu fazer uma estimativa por alto das áreas de conhecimento existentes e do tamanho dos *subcorpora*.

Em primeiro lugar, a cada revista foi atribuída uma área científica, seguindo a classificação da CAPES[10] de áreas do conhecimento em

9. O número ISSN - *International Standard Serial Number* - foi utilizado como variável no processo de extração de textos, garantindo a não repetição de periódicos no corpus. Além disso, esse número único de identidade foi incluído nos nomes dos arquivos, facilitando a interoperabilidade com a plataforma SciELO e, com isso, acesso imediato à fonte original.

10. CAPES significa Coordenação de Aperfeiçoamento de Pessoal de Nível Superior e é uma fundação do Ministério da Educação no Brasil. Ela criou a Tabela das Áreas de Conhecimento do Ensino Superior, com quatro níveis hierárquicos, desde o mais geral - grande área - até o mais específico - especialidade. A fim de facilitar as atividades de

36.2
2020

Colégios de Ciências da Vida (CV), Colégio de Humanidades (HU) e Colégio de Ciências Exatas, Tecnológicas e Multidisciplinar (CE), como mostrado nas **Figuras 1 a 3**[11]. Essa foi uma tentativa de obter uma visão ampla da distribuição de periódicos por área em cada coleção SciELO sem entrar em subdivisões mais especializadas de domínios.

COLÉGIO DE CIÊNCIAS DA VIDA

CIÊNCIAS AGRÁRIAS	CIÊNCIAS BIOLÓGICAS	CIÊNCIAS DA SAÚDE
Ciência de Alimentos	Biodiversidade	Educação Física
Ciências Agrárias I	Ciências Biológicas I	Enfermagem
Medicina Veterinária	Ciências Biológicas II	Farmácia
Zootecnia / Recursos Pesqueiros	Ciências Biológicas III	Medicina I
		Medicina II
		Medicina III
		Nutrição
		Odontologia
		Saúde Coletiva

Figura 1 – Colégio de Ciências da Vida

avaliação, que é uma de suas linhas de ação, a CAPES adotou uma categorização mais ampla, com grandes áreas agrupadas em colégios, adotando "afinidade" como o principal critério de agrupamento. No CoPEP, essa categorização mais ampla foi adotada. Para mais informações, consulte: http://www.capes.gov.br.

11. CAPES- Tabela de áreas de conhecimento.

COLÉGIO DE HUMANIDADES

CIÊNCIAS HUMANAS	CIÊNCIAS SOCIAIS APLICADAS	LINGUÍSTICA, LETRAS E ARTES
Antropologia / Arqueologia	Administração, Ciências Contábeis e Turismo	Artes / Música
Ciência Política e Relações Internacionais	Arquitetura e Urbanismo	Letras / Linguística
Educação	Ciências Sociais Aplicadas	
Filosofia / Teologia	Direito	
Geografia	Economia	
História	Planejamento Urbano e Regional / Demografia	
Psicologia	Serviço Social	
Sociologia		

Figura 2 – Colégio de Humanidades

COLÉGIO DE CIÊNCIAS EXATAS, TECNOLÓGICAS E MULTIDISCIPLINAR

CIÊNCIAS EXATAS E DA TERRA	ENGENHARIAS	MULTIDISCIPLINAR
Astronomia / Física	Engenharias I	Biotecnologia
Ciência da Computação	Engenharias II	Ciências Ambientais
Geociências	Engenharias III	Ensino
Matemática / Probabilidade e Estatística	Engenharias IV	Interdisciplinar
Química		Materiais

Figura 3 – Colégio de Ciências Exatas, Tecnológicas e Multidisciplinar

Como esperado, dada a grande diferença na cobertura de tempo de cada coleção nacional, confirmamos que a SciELO-Pt é muito menor do que a SciELO-Br. Em fevereiro de 2016, havia 55 periódicos na SciELO-Pt. O número de artigos variou de 1 a 75 por periódico, totalizando 965 edições com a seguinte distribuição nos colégios: 457 (HU), 426 (CV) e 82 (CE), enquanto a SciELO-Br continha 345 periódicos em sua página. O número de artigos por periódico variou de dois a

DELTA
36.2
2020

440, totalizando 18.270 artigos no total. Aqui, o Colégio de Ciências da Vida é predominante: 11.454 artigos, seguido de 9.054 artigos das Humanidades e 2.209 do Colégio de Ciências Exatas, Tecnológicas e Multidisciplinar. Consequentemente, a configuração final do *corpus* em termos de tamanho e equilíbrio foi determinada pela menor coleção, que é a SciELO-Pt.

Antes de avançarmos para a extração dos textos, foi tomada a decisão de que, inicialmente, criaríamos *subcorpora* referentes ao país de origem das publicações, isto é, o *subcorpus* BR para os textos extraídos da SciELO-Br e o *subcorpus* PT para os textos extraídos da SciELO-Pt, assumindo, por ora e para simplificação nesta fase da compilação, que a variedade do português em cada *subcorpus* seria, por padrão, aquela do país de publicação[12]. A identificação da variedade de língua usada em cada texto, que permitiu definir que os *subcorpora* inicialmente referentes à origem da fonte poderiam ser considerados *subcorpora* de variedades na versão final do *corpus*, foi realizada apenas no final do processo, como será mostrado mais adiante.

4.2. Construindo o corpus

A primeira parte desta seção descreve o processo inicial de construção do *corpus* baseado na extração de arquivos XML das coleções nacionais SciELO. Contudo, devido a sérios problemas de inconsistência do XML na SciELO, o CoPEP acabou por ser construído a partir da extração de arquivos HTML. Assim, a segunda parte desta seção exibe esse processo que, apesar de não ter sido o planejado, se beneficiou de aprimoramentos aprendidos durante a primeira extração (XML).

12. Esse é um ponto-chave porque, embora se pudesse supor que contribuições em cada uma das coleções nacionais deveriam ter sido escritas na variedade do respectivo país de publicação, sabemos que essa não é a realidade. Por um lado, não só é bastante comum que os acadêmicos portugueses publiquem no Brasil e vice-versa, como também o é que artigos resultantes de trabalho em equipe, composta por investigadores brasileiros e portugueses, sejam publicados em ambos os países. Por outro lado, nosso mundo globalizado deu lugar a uma internacionalização crescente de universidades e centros de pesquisa, com estudantes e pesquisadores publicando em seu país de origem, mas também de residência. Isso significa que as coleções nacionais não podem ser consideradas como representativas da variedade de língua.

4.2.1. Extração de XML

Uma vez determinado o intervalo de tempo do *corpus* (2000 a 2015), designados os colégios de conhecimento (CV, HU e CE) para cada periódico, definida a SciELO-Pt como a primeira coleção a ser extraída e identificado que as URL de cada artigo na SciELO contém um identificador único baseado no padrão "ISSN-ISSNANOEDICAO-NUMERO", os procedimentos computacionais foram iniciados. Estes foram os passos dados:

a) Tomando como variáveis o padrão acima mencionado, o intervalo de tempo definido, os colégios de conhecimento atribuídos e a definição de que apenas os arquivos XML escritos em português deviam ser extraídos, foram sobregerados *links* usando expressões regulares.

b) Esses *links* foram baixados usando um aplicativo *weget* em um *bash loop* baseado neste tipo de comando: wget -nv -nc -O <pasta_destino / arquivo_destino.xml> <URL>

c) Os *links* foram colocados na pasta *subcorpus* PT e nas pastas CV, HU ou CE referentes aos colégios de conhecimento.

d) Os artigos XML brutos da SciELO-Pt foram extraídos usando XSLT (*engine xsltproc* no Debian) e o *translational* XSL principal do W3C.

e) Cada artigo foi salvo automaticamente como um arquivo cujo nome continha o ISSN, ano, número da edição, número do artigo e colégio de conhecimento. O *script* de extração da SciELO-Pt teve vários problemas, parando muitas vezes e fazendo com que a extração demorasse muito tempo. Em simultâneo, foram selecionados manualmente periódicos e número de artigos por periódico a serem extraídos da SciELO-Br, visando manter o equilíbrio com a SciELO-Pt em termos de equivalência entre colégios e temas.

A grande maioria dos periódicos apresentou correlação direta, por exemplo, Psicologia (SciELO-Pt) e Fractal: Revista de Psicologia (SciELO-Br). No entanto, outros não. Nesse caso, analisamos manualmente a descrição da revista na SciELO-Pt e tentamos encontrar um equivalente na SciELO-Br. A **Tabela 3** mostra um exemplo como ilustração do processo de busca por correspondência, apontando a(s) revista(s) escolhida(s) e a justificativa.

DELTA
36.2
2020

Tabela 3 – Busca de correspondência entre revistas das duas coleções

Colégio	**SciELO-Pt**	**SciELO-Br**	**Justificativa**
CE	Revista de Gestão Costeira Integrada (16 números)	Revista Brasileira de Oceanografia (5 números) Revista Ambiente & Água (11 números)	Nenhuma correspondência direta em termos de tópico. Dado o número insuficiente de números da revista com o tópico mais próximo (Revista Brasileira de Oceanografia), selecionamos um segundo periódico com um tópico genérico mais amplo (Revista Ambiente & Água).

Quando a extração dos artigos SciELO-Pt foi realizada, percebemos que os metadados XML continham informações sobre grandes áreas de conhecimento para cada periódico, seguindo a classificação da CAPES apresentada na seção 2 acima. Assim, essa informação também foi extraída de ambas as coleções nacionais. A **Tabela 4** apresenta as abreviaturas usadas para os colégios e grandes áreas em nosso *corpus*.

Tabela 4 – Colégios e grandes áreas no CoPEP

Colégios	***Colégio de Humanidades (HU)***		***Colégio de Ciências da Vida (CV)***		***Colégio de Ciências Exatas, da Terra e Multidisciplinar (CE)***	
Grandes áreas	Ciências Humanas (Hu)	Ciências Socias Aplicadas (Ap)	Ciências da Saúde (He)	Ciências Agrícolas (Ag)	Engenharia (En)	Ciências Exatas e da Terra (Ex)

Como algumas revistas estavam classificadas na SciELO em mais de uma grande área, foi feita uma tentativa de fornecer critérios objetivos para determinar apenas uma disciplina por revista. Então seguimos estes passos:

i. Como a Fundação para a Ciência e a Tecnologia (FCT) em Portugal corresponde aproximadamente à CAPES no Brasil, verificamos a quais grandes áreas essas revistas foram designadas pela FCT. Por exemplo, a Revista Portuguesa de Ciências do Desporto (classifica-

ção SciELO-Pt: Ciências da Saúde, Ciências Humanas) pertence às Ciências da Saúde na FCT, assim, no nosso *corpus*, essa revista foi classificada como grande área: He; colégio: CV.

ii. Se os periódicos não constavam na FCT, verificamos a classificação da *ISI Web of Knowledge*. Por exemplo, a revista Nascer e Crescer é classificada na SciELO-Pt como Ciências Sociais Aplicadas, Ciências Biológicas, Ciências da Saúde, Ciências Humanas. Segundo o *ISI Web of Knowledge*, essa revista pertence à pediatria, que é uma especialização em Ciências da Saúde. Assim, no CoPEP, a Nascer e Crescer foi atribuída à grande área He.

iii. Se nem a FCT nem a ISI fornecessem uma classificação clara, então o colégio da revista correspondente na outra coleção seria adotado. Por exemplo, como correspondentes aos Cadernos de Estudos Africanos da SciELO-Pt (classificação SciELO: Ciências Sociais Aplicadas, Ciências Humanas, Linguística, Artes e Humanidades), foram atribuídos dois periódicos na SciELO-Br, o Afro-Ásia e os Estudos Afro-Asiáticos, que pertencem às Ciências Humanas. Logo, os Cadernos de Estudos Africanos foram definidos como pertencentes à grande área de Ciências Humanas. Outro resultado inesperado da extração de artigos SciELO-Pt é que alguns periódicos, que haviam sido inicialmente marcados como escritos em inglês, em função de seus títulos estarem em inglês, estavam escritos em português. Estes foram adicionados ao *corpus*, e novos artigos correspondentes tiveram que ser encontrados na SciELO-Br, seguindo a mesma análise manual dos periódicos descrita acima. No final, apenas um periódico estava escrito exclusivamente em inglês, sendo assim descartado. Por fim, resultados adicionais da extração mostraram que em muitos casos em que a língua havia sido determinada como portuguesa, os textos estavam apenas disponíveis em formato PDF. Isso foi descoberto com a identificação no XML da linha * Texto completo disponível apenas em PDF. *. Então, optamos pela exclusão desses documentos, mantendo, no entanto, um registro dos nomes dos arquivos (1.650 arquivos no total) em uma pasta diferente. A decisão pelo descarte desses arquivos se deveu porque a conversão de arquivos PDF em formato TXT requer revisão manual de cada arquivo, em função dos resultados recorrentemente problemáticos (por exemplo, artigos de duas colunas por página perdem completamente sua estrutura interna). Após essa série de ações

DELTA
36.2
2020

manuais, o *script* de extração dos textos da SciELO-Br passou a ser executado. Deve-se mencionar que, assim como na SciELO-Pt, houve muitos eventos de servidor inativo e interrupções intermitentes. Assim que os *subcorpora* BR e PT estavam completos, foram realizados outros procedimentos computacionais adicionais:

f) Limpeza automática de arquivos XML extraídos: exclusão de informações bibliográficas do autor (nome, afiliação, e-mail, cargo), resumos, palavras-chave, imagens, tabelas, figuras, gráficos, referências, informações extras (informações de direitos autorais, endereço do editor, agradecimentos, recebido em, aceito em).

g) Substituição da adoção de pastas de colégios pela nomeação dos arquivos com informações sobre colégio e grande área. Assim, cada arquivo foi renomeado em lote com base em códigos de posição fixa, mantendo o identificador exclusivo usado nas URL da SciELO para referência futura. Por exemplo, no nome do arquivo BRCVHe0104-42302008000100023, temos:

BR: domínio de primeiro nível do país para a variedade do português do artigo
CV: colégio de conhecimento
He: grande área de conhecimento
0104-4230: ISSN
2008: ano de publicação
0001: número (edição)
00023: número (posição) do artigo na edição

h) Realização da homogeneização da codificação. Os arquivos na codificação Latin1/ ISO-8859-1 foram convertidos em UTF-8. Ao analisarmos as transformações resultantes da limpeza, percebemos que muitos arquivos não eram XML válidos, sugerindo inconsistência da estrutura organizacional da SciELO. Enquanto experimentávamos soluções alternativas, selecionamos uma amostra dos 46.935 arquivos limpos e convertidos para analisarmos manualmente e avaliarmos a extensão dos problemas. Deve-se mencionar que o propósito dessa análise não foi gerar estatísticas, mas apenas verificar se os problemas atingiam todos os arquivos. O processo de análise manual foi realizado conforme descrito abaixo.

i. Descrição dos aspectos inspecionados (identificados durante a verificação dos resultados da transformação):

• Limpeza inadequada: presença de marcadores XML.

• Arquivos vazios: sem conteúdo textual.

• Arquivos inúteis: nenhuma informação textual qualificada, por exemplo resumos em português e outras línguas (espanhol ou inglês), dados biográficos, referências, palavras-chave, etc.

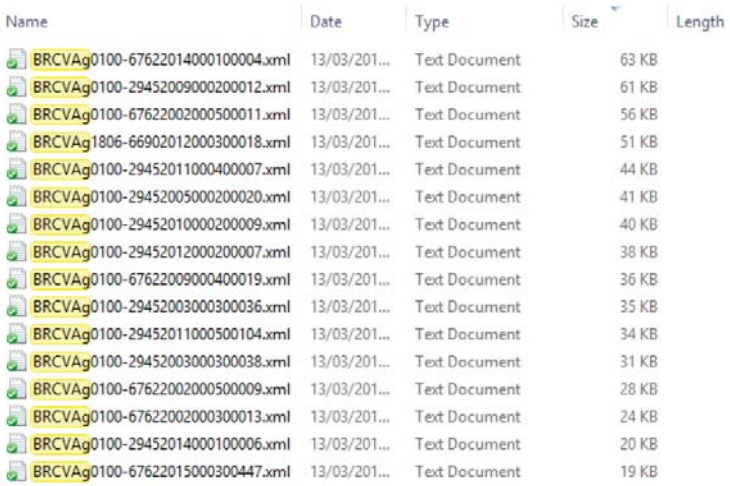
ii. Organização dos dados:

1. Os arquivos foram agrupados de acordo com a variedade de língua, colégio e grande área, e o número total de arquivos para cada grupo foi registrado:

Tabela 5 – Número de arquivos por *subcorpus* de variedade e grandes áreas de conhecimento

***Subcorpus* BR (número de arquivos)**	***Subcorpus* PT (número de arquivos)**
CVAg (4.115)	CVAg (758)
CVHe (14.966)	CVHe (3.321)
CEEn (1.345)	CEEn (178)
CEEx (4.083)	CEEx (189)
HUAp (2.959)	HUAp (1.131)
HUHu (9.733)	HUHu (3.788)

2. No Explorador de Arquivos (*File Explorer*), os arquivos de cada grupo foram ordenados por tamanho, com o maior arquivo no topo. Um arquivo foi selecionado aproximadamente a cada 10 KB de diferença de tamanho, até o tamanho mínimo de 10 KB. A determinação desse valor resultou da análise manual de arquivos de tamanho menor, que eram apenas lixo, enquanto arquivos de 10 KB às vezes correspondiam a editoriais.



Name	Date	Type	Size	Length
BRCVAg0100-67622014000100004.xml	13/03/201...	Text Document	63 KB	
BRCVAg0100-29452009000200012.xml	13/03/201...	Text Document	61 KB	
BRCVAg0100-67622002000500011.xml	13/03/201...	Text Document	56 KB	
BRCVAg1806-66902012000300018.xml	13/03/201...	Text Document	51 KB	
BRCVAg0100-29452011000400007.xml	13/03/201...	Text Document	44 KB	
BRCVAg0100-29452005000200020.xml	13/03/201...	Text Document	41 KB	
BRCVAg0100-29452010000200009.xml	13/03/201...	Text Document	40 KB	
BRCVAg0100-29452012000200007.xml	13/03/201...	Text Document	38 KB	
BRCVAg0100-67622009000400019.xml	13/03/201...	Text Document	36 KB	
BRCVAg0100-29452003000300036.xml	13/03/201...	Text Document	35 KB	
BRCVAg0100-29452011000500104.xml	13/03/201...	Text Document	34 KB	
BRCVAg0100-29452003000300038.xml	13/03/201...	Text Document	31 KB	
BRCVAg0100-67622002000500009.xml	13/03/201...	Text Document	28 KB	
BRCVAg0100-67622002000300013.xml	13/03/201...	Text Document	24 KB	
BRCVAg0100-29452014000100006.xml	13/03/201...	Text Document	20 KB	
BRCVAg0100-67622015000300447.xml	13/03/201...	Text Document	19 KB	

Figura 4 – Arquivos ordenados para seleção

3. Os arquivos da amostra foram agrupados em pastas. Cada pasta foi carregada no programa *CountAnything*[13] para contagens de palavras e caracteres. Os resultados foram salvos como arquivos de texto;

4. Os resultados foram importados para uma planilha do Excel, ordenada por tamanho.

Tabela 6 – Arquivos selecionados no Excel

	Size	**Words**	**Chars**
BRCVAg (4.115 files)			
BRCVAg\BRCVAg1806-66902016	10KB	1.190	9.738
BRCVAg\BRCVAg0100-67622015	15KB	1.667	14.205
BRCVAg\BRCVAg0100-29452014	20KB	2.525	17.681
BRCVAg\BRCVAg0100-29452003	31KB	3.836	26.676

13. http://ginstrom.com/CountAnything/.

iii. Análise qualitativa:

1. Abrimos um arquivo de cada vez e examinamos:
- conversão de codificação.
- presença de marcadores XML.
- presença de corpo de texto (nenhum, incompleto, completo).

1.1. Quanto ao corpo do texto, as anotações foram:
- se nenhum corpo de texto estava presente: inútil.
- se o corpo completo do texto estava presente: artigo completo.
- se o corpo incompleto do texto estava presente: incompleto.

1.2. Se a seção do periódico era identificável, o nome era registrado entre parênteses.

iv. Resultados: Presença de erros de codificação em quase todos os arquivos, apesar de termos feito conversão prévia da codificação original ISO-8859-1 para UTF-8, e de marcadores XML em todos os arquivos. Além disso, os textos completos não estavam limpos, com a presença de título, resumo(s), palavras-chave, informações sobre o autor, referências, entre outros. A **Figura 5** ilustra o conteúdo dos textos que foram analisados:

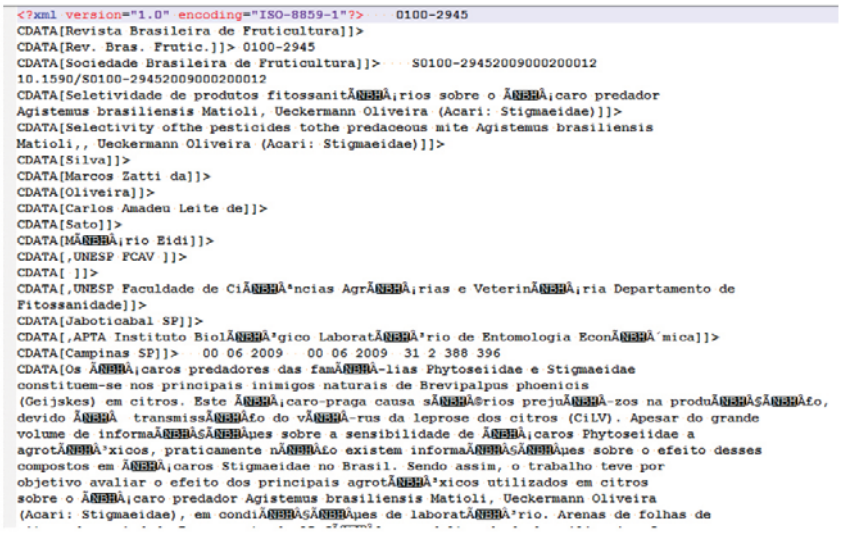


Figura 5 – Exemplo do resultado de uma extração de XML após limpeza

DELTA
36.2
2020

O número total de arquivos neste primeiro *corpus* provisório era de 46.935 arquivos. Havia 44.544 arquivos com tamanhos entre 0 e 10 KB e 2.011 arquivos com tamanho entre 11 KB e 21 KB. No exame, 23 de 27 arquivos do tipo "inútil" eram menores que 21KB, sugerindo que apenas cerca de 380 artigos poderiam conter textos.

v. Conclusões: Essa análise mostrou que a limpeza tinha sido ineficaz e que faltava conteúdo textual nos arquivos em todo o *corpus*, indicando que a extração de XML não era viável e que uma nova solução deveria ser encontrada. Foi assim decidido que deveríamos fazer a extração de textos em formato HTML.

4.2.2. Extração de arquivos HTML

Após contato com o informático responsável pela SciELO, foi confirmado que a estrutura organizacional da plataforma não era consistente, com arquivos XML problemáticos[14]. A decisão foi então extrair arquivos HTML.

Apesar desse revés inesperado, a extração HTML beneficiou-se das lições aprendidas no primeiro processo, que aprimoraram e ajudaram a acelerar o processo de construção do *corpus*. Deve-se ressaltar que, apesar de conseguirmos obter arquivos válidos desta vez, o servidor SciELO ainda apresentava problemas. Assim, vários eventos de servidor inativo causaram inconvenientes interrupções de execuções de *scripts*, causando atrasos.

A extração decorreu da seguinte forma, primeiramente da SciELO-Pt, e depois da SciELO-Br:

14. Santos e Packer (2014:83) explicaram que, até 2013, "os arquivos recebidos dos editores da revista foram convertidos em texto simples codificado em formato HTML (HyperText Markup Language) para serem marcados de acordo com a estrutura SciELO SGML, e então armazenados em um banco de dados para publicação e distribuição *on-line*" (tradução nossa). Em 2013, o formato XML substituiu a estrutura do SGML. Isso sugere que essa transição de um formato para outro pode ter sido a razão para essa estrutura organizacional problemática. No entanto, pode-se dizer que esse problema deverá ser superado quando o processo de transformação estiver concluído.

a) Os arquivos foram obtidos usando um *script* PHP, fazendo-se automaticamente sua nomeação e pré-limpeza e normalizando-os usando um analisador DOM.

Utilizaram-se os mesmos parâmetros: anos 2000 a 2015, ISSN, número da edição, número do artigo, fonte da variedade de língua, colégio e grande área.

b) Alguns arquivos foram colocados de lado (os que indicavam claramente a disponibilidade somente em PDF) usando um *script bash* baseado em *grep*.

Nesta etapa, 1.650 arquivos cujos textos estavam disponíveis apenas em PDF foram excluídos.

c) Comandos *sed* baseados em expressões regulares foram usados para recortar e aparar os arquivos; os ponteiros foram variações de: limite superior - "palavras-chave"; limite inferior - "referências".

d) A limpeza final do HTML foi feita usando um *asciinator* (*html2text*).

e) Finalmente, as restantes entidades de caracteres foram recodificadas usando o *recode* GNU. Como na extração anterior, selecionamos uma amostra dos arquivos extraídos, desta vez de entre os 7.740 arquivos do *subcorpus* PT, para uma análise manual dos textos, de forma a avaliarmos os resultados desse novo procedimento. Enquanto isso, o *script* de extração ainda estava sendo executado na SciELO-Br.

A análise da amostra de arquivos foi realizada desta forma:

i. Arquivos com tamanhos de 1KB, 2KB e 3KB foram todos abertos. Vimos que arquivos de tamanho 1KB eram lixo. Os ficheiros com 2KB apresentavam por vezes textos curtos que deveriam ser mantidos, por exemplo, "carta ao Director", "nota introdutória", "pôsteres", "notícias", "editorial" e, por vezes, conteúdo inútil. Esses últimos tipos de arquivos geralmente continham endereço de correspondência, título, nome do autor, afiliação, resumo, e-mail, entre outras informações irrelevantes para este *corpus*, juntamente com a frase "Texto completo disponível apenas em formato PDF". Alguns arquivos com 3 KB consistiam em informações relevantes, enquanto outros eram textos

DELTA
36.2
2020

incompletos ou simplesmente resumos em outros idiomas com outras informações.

ii. No Explorador de Arquivos (*File Explorer*), os arquivos foram ordenados por nome. Depois, selecionávamos cinco arquivos, descíamos na lista 200 arquivos e selecionávamos outros cinco arquivos, repetindo esse processo até chegar ao final da lista e evitando arquivos menores que 3KB. No final, a amostra continha 195 arquivos.

iii. Abrimos os arquivos um por um e tomamos nota dos diferentes tipos de informações indesejadas encontradas ou casos de corte incorreto de texto (falta de início ou fim do texto).

iv. Resultados: A extração de HTML foi muito bem-sucedida, pois todos os arquivos eram textos simples, sem qualquer tipo de marcação HTML. A limpeza foi 100% efetiva em 8 de 195 arquivos, com diferentes níveis de eficiência nos restantes, variando desde alguns tipos comuns e esperados de informações indesejadas até aqueles que poderiam ser eliminados através de limpeza automática adicional (diferentemente da extração XML, na qual os textos eram apenas lixo). A codificação da conversão para UTF-8 estava correta, com apenas uma transformação com falha no caso de alguns símbolos. Outro resultado significativo da análise é que encontramos um padrão que poderia ser usado na nova limpeza automática: a presença de *[Creative_Commons_License]* no final do texto, que às vezes é seguido por informações desnecessárias, como instituição, endereço da instituição, um espaço reservado para a imagem e um endereço de e-mail. Por exemplo:

> [Creative_Commons_License] Todo o conteúdo deste periódico, exceto onde está identificado, está licenciado sob uma licença_Creative_Commons
> Instituto de Ciências Sociais da Universidade de Lisboa
> Av. Professor Aníbal de Bettencourt, 9
> 1600-189 Lisboa
> [/img/pt/e-mailt.gif]

Independentemente de ser seguida por algumas sequências extras de caracteres ou não, a linha *Creative Commons* foi considerada um ponto de corte eficaz.

v. Conclusões: Decidiu-se que arquivos do tamanho de 1KB poderiam ser eliminados com segurança, já que eram lixo eletrônico. Além disso, tornou-se evidente que alguma limpeza extra deveria ser realizada, na qual observações sobre tamanhos e padrões (PDF, *Creative Commons License*) poderiam ser incorporadas como parâmetros.

4.2.2.1. Segunda extração HTML

A partir dos resultados da análise, um novo roteiro de limpeza automática foi implementado. Os novos parâmetros de exclusão foram:

- arquivos menores que 2,5 KB.
- Padrão *Creative Commons.*
- arquivos menores que 5KB contendo "in? PDF? <" ou "em? PDF?".
- maior intervalo de tempo: 1997 até o presente (abril de 2016).

A decisão de prolongar o período de tempo para o ano da publicação mais antiga na SciELO-Pt resultou da redução significativa do tamanho do *subcorpus* PT devido ao número considerável de arquivos "somente PDF". Uma segunda limpeza foi realizada nos *subcorpora* PT e BR (este último, com extração já finalizada nessa altura). A **Tabela 7** mostra os números totais para os dados tratados, ou seja, após os parâmetros acima terem sido aplicados. Deve-se notar que outros 706 arquivos com indicação de disponibilidade somente em PDF foram excluídos do *subcorpus* PT, totalizando 2.356 arquivos deste *subcorpus* que, em princípio, não poderiam ser utilizados (como mencionado na seção 2 acima).

DELTA
36.2
2020

Tabela 7 – Número de *tokens* por *subcorpus* de variedade de idioma e grande área de conhecimento

	***Tokens* no PT**	***Tokens* no BR**
HUHu	12.408.219	44.433.173
HUAp	2.959.848	11.686.493
HU	15.368.067	56.870.606
CVHe	6.163.331	20.744.523
CVAg	1.196.919	11.330.446
CV	7.360.250	32.134.231
CEEn	332.449	7.549.102
CEEx	602.955	14.509.119
CE	935.404	23.821.008
Total	23.663.721 6.632 arquivos	114.334.616 25.619 arquivos

Como se pode ver, o *subcorpus* BR é muito maior do que o seu equivalente, exigindo um cuidadoso processo de balanceamento. Com base no número de *tokens* por grande área, os arquivos (em ordem de tamanho) foram selecionados primeiro até atingir uma contagem total aproximada daquela do *subcorpus* PT. Esta primeira versão do *corpus* tinha 47.341.296 *tokens*: PT: 23.663.721 e BR: 23.677.575. No entanto, deve ser lembrado que estes *subcorpora* se referem à fonte de publicação, e não à variedade de língua.

Enquanto o *corpus* estava sendo compilado, uma *sketch grammar*[15] para o português também estava sendo desenvolvida (Kuhn & Kosem, 2016). Conforme erros resultantes de problemas na limpeza iam sendo encontrados (por exemplo, a presença de um parágrafo inteiro em outro idioma, palavras-chave, resumos, agradecimentos, etc.) fomos registrando os nomes dos arquivos em que ocorriam e os procurando no *corpus*. Revisamos manualmente todos os textos, corrigindo outros erros que ocasionalmente surgiam, a saber, segundos títulos em inglês, nomes de autores, palavras "coladas" ou simplesmente descartamos

15. Tal recurso faz parte do *Sketch Engine* e serve para dar instruções ao sistema sobre como encontrar *word sketches* em um *corpus*.

arquivos com textos incompletos e arquivos adicionais "somente em PDF".

Além da limpeza manual, também tomamos nota dos arquivos escritos em uma variedade do português contrária à da publicação para futuras modificações de nomes de arquivos, por exemplo:

#artigos em PB encontrados no *subcorpus* PT
PTCEEx1646-88722014000100005.html.txt
PTCVAg0871-018X2014000300005.html.txt
#artigos em PE encontrados no *subcorpus* BR
BRCVHe0034-71672014000300360.html.txt
BRCVHe0034-71672014000600913.html.txt
BRHUHu0002-05912014000100002.html.txt

O procedimento para identificação da variedade deveria ter sido realizado já nesta versão mais limpa do *corpus*. No entanto, como o número total de *tokens* foi reduzido ainda mais devido aos resultados da avaliação manual dos textos, decidimos que deveríamos experimentar fazer conversão de PDF, uma vez que 2.356 arquivos poderiam fornecer muito mais dados.

4.2.3. Conversão de PDF

Idealmente, os PDF seriam convertidos automaticamente, e teríamos que apenas analisar manualmente os resultados. No entanto, devido a alguns atrasos no procedimento computacional, iniciamos o processo de conversão manualmente. Isso implicava baixar os PDF um a um, depois enviá-los para um conversor de PDF para TXT *on-line* gratuito e salvar os resultados, ou seja, salvar os arquivos em formato TXT.

Tendo uma lista dos *links* para as versões *on-line* em PDF de 1.650 artigos da SciELO-Pt que haviam sido descartados na primeira extração de HTML, abrimos cada um deles para verificar se seria possível fazer a conversão. Como testes preliminares de conversão de textos organizados em duas colunas por página indicaram que essa estrutura se perdeu na versão TXT, criamos um critério de fluxo de trabalho no qual apenas documentos com uma coluna eram candidatos à conversão, evitando assim trabalhos desnecessários.

DELTA
36.2
2020

Seguimos então o procedimento mencionado acima, o que resultou em 69 arquivos TXT, que também foram limpos manualmente, e incluíam a eliminação de títulos em outros idiomas, resumos (em português e outros idiomas), informações do autor, palavras-chave, agradecimentos, referências, tabelas, imagens, datas de submissão e publicação, licenças *Creative Commons* e alguns outros tipos de informações irrelevantes. Além disso, outros elementos típicos do formato PDF, ou seja, marcadores de *layout* da revista, como cabeçalhos e rodapés com nomes de autores ou títulos de periódicos, números de página e notas de rodapé, também foram eliminados manualmente.

Embora esses textos tenham sido publicados na SciELO-Pt, uma análise manual mostrou que 18 foram de fato escritos em português brasileiro. Esses nomes de arquivos foram registrados e foi criada uma lista para extração com os *links* dos 733 PDF restantes. Esses PDF foram o número total restante após a conversão manual de 69 textos e a eliminação de 848 documentos com coluna dupla.

Ao mesmo tempo, a *sketch grammar* estava sendo aprimorada, então tivemos a oportunidade de realizar ainda mais limpeza manual no *corpus* (da mesma maneira descrita anteriormente), descartando textos adicionais escritos em outros idiomas e eliminando títulos duplos, referências e a indicação de textos somente em PDF dos textos. Esta nova versão modificada do *corpus* foi salva para ser tratada posteriormente.

Em seguida, tomamos os 638 arquivos TXT convertidos automaticamente para fazermos a limpeza manual. Inicialmente, colocamos muitos deles de lado devido à estrutura dos textos, que ficou desordenada após a conversão, e demos prioridade a arquivos pertencentes a grandes áreas com menos textos, como Ciências Agrárias, Engenharia, Ciências Exatas e Ciências da Vida. No entanto, entre aqueles textos do tipo "quebra-cabeça", que havíamos descartado, havia uma quantidade considerável de Ciências Exatas e Ciências da Vida. Assim, apesar do trabalho árduo, optamos por incluí-los na revisão manual e reorganizar sua estrutura interna tomando como referência a versão original em PDF.

No final, analisamos e limpamos manualmente 213 arquivos, 34 deles escritos em português brasileiro e 179 em português europeu.

Tendo uma versão nova e mais limpa do *corpus*, e um adicional de 213 arquivos, poderíamos então passar para a fase final: a criação de *subcorpora* de variedade do português, nomeação dos novos arquivos e balanceamento.

4.2.4. Fase final da construção do corpus

A fase final consistiu na criação de um *subcorpus* português brasileiro e um *subcorpus* português europeu, renomeação dos arquivos com a adição de um código para variedade de idioma, e equilíbrio do *corpus* de acordo com a quantidade total de palavras por *subcorpora* e por grande área.

4.2.4.1. Distribuição dos textos nos subcorpora

Até o momento, o *corpus* estava dividido em dois *subcorpora* de acordo com as fontes de publicação dos textos, ou seja, *subcorpus* BR para textos da SciELO-Br e *subcorpus* PT para textos da SciELO-Pt. Como explicado anteriormente, uma correspondência direta entre o local de publicação e a variedade de língua escrita não pode ser feita no caso da língua portuguesa. Assim, decidimos que cada um desses *subcorpora* seria analisado para confirmarmos a variedade na qual os textos foram escritos. Arquivos com divergência entre fonte e variedade foram movidos para o *subcorpus* da variedade identificada. No final, obtivemos dois *subcorpora* conforme a variedade do português usado nos textos.

Para a análise de cada *subcorpora*, adotamos um procedimento de identificação da variedade de língua dos textos, que foi realizado em duas fases para cada *subcorpus*: (1) identificação automática das variedades PB e PE em textos escritos em PE e PB, respectivamente; (2) confirmação manual da correspondência entre a variedade identificada e aquela escrita.

Deve-se salientar que a identificação automática da variedade do português não é um método infalível, devido a várias razões. Em primeiro lugar, há casos em que os textos escritos em coautoria por bra-

DELTA
36.2
2020

sileiros e portugueses exibem as duas variedades simultaneamente. Em segundo lugar, alguns autores aplicam a ortografia inconsistentemente, variando aleatoriamente entre PB e PE. Finalmente, às vezes as grafias supostamente consideradas desviantes em determinada variedade são, de fato, formas aceitas. Diante disso, os resultados da identificação automática das variedades devem ser considerados apenas indicativos, não uma conclusão definitiva. É por isso que o procedimento seguido neste trabalho envolveu uma segunda fase, na qual revisamos manualmente os textos para confirmar a indicação automática da variedade.

É importante lembrar que, neste momento da compilação, os 213 arquivos convertidos em PDF e manualmente analisados, conforme descrito na seção anterior, não haviam sido adicionados aos *subcorpora*, pois a identificação da variedade já havia sido feita durante a avaliação manual dos resultados da conversão para TXT. Eles foram adicionados ao *corpus* apenas após os novos *subcorpora* terem sido definidos.

Começando com o *subcorpus* BR, estas foram as etapas seguidas:

a) Uso de uma *stoplist* com variantes do PE: Foi utilizada uma *stoplist* com variantes PE (por exemplo, *excepção*, *facto*). Anotamos em uma lista os nomes dos arquivos de textos contendo pelo menos uma ocorrência de qualquer uma das variantes PE. No final, a lista de textos candidatos a PE continha 82 textos, variando de 40 a 1 ocorrência por texto.

b) Uso de uma *stoplist* com terminações fonológicas do PE: Em seguida, um *script* com terminações fonológicas do PE (por exemplo, *-émico*, *-ómico*) foi executado. Como acima, os nomes dos arquivos contendo tais terminações foram salvos, totalizando 101 arquivos, variando de 44 a 1 ocorrência.

c) Mesclagem das listas resultantes das duas *stoplists*: O resultado foi uma lista de candidatos contendo 149 nomes de arquivos.

d) Revisão manual: O objetivo da revisão manual era confirmar a correspondência entre a variedade identificada e aquela realmente usada na escrita. Classificamos os 149 nomes de arquivos em ordem de frequência e lemos cada texto até chegar à posição 50, correspondendo ao último texto com três ocorrências de variantes PE. Revisamos então mais 16 textos com duas ocorrências e concluímos que eram

todos falsos positivos, ou seja, os textos haviam sido de fato escritos em português brasileiro. Assim, os arquivos restantes foram determinados como sendo PB. A revisão manual se baseou no fato de um dos compiladores ser um falante de português brasileiro como língua materna que conhece as características distintivas do português europeu, especificamente no contexto do português acadêmico escrito. Assim, o exame dos textos procurou identificar elementos característicos de uma determinada variedade ou que pertencem a uma ou outra cultura, como (entre muitas outras):

i) Para confirmação de textos escritos em português europeu:
- *facto*
- *registo*
- acento agudo com verbos no passado simples (*habituámo-nos*);
- *acção* antes de 2009
- *ideia* antes de 2009
- colocação *deitar abaixo*

ii) Para confirmação de textos escritos em português brasileiro:
- *aspecto* e *ruptura* após 2009
- *européia*, *idéia* (ditongo acentuado)
- *de fato*
- *tese de doutorado*
- *bolsista CAPES*
- trema (*ü*)

No final, dos 149 textos candidatos a PE, apenas 20 haviam sido escritos em português europeu. Um achado interessante resultante desta revisão foi a identificação de textos publicados após 2009 que empregavam simultaneamente a ortografia do Acordo Ortográfico de 1990 e do Acordo Ortográfico de 1945. Outra observação reveladora refere-se à presença de títulos "enganosos", que foram transformados em PB, enquanto o artigo estava, de fato, escrito em PE (por exemplo, o título: *Índices plaquetários em indivíduos com doença hepática alcoólica crônica*). As etapas foram seguidas da mesma forma com o *subcorpus* PT, sendo a única diferença a realização de um método ligeiramente modificado para inspeção manual dos resultados, conforme será explicado a seguir.

DELTA
36.2
2020

Chama a atenção que a identificação automática de variantes de PB no *subcorpus* PT forneceu uma lista com 921 nomes de arquivos. Em comparação com a lista de 149 textos acima, uma diferença tão grande sugere mais brasileiros publicando em Portugal do que o contrário.

No contexto de uma análise manual de cada texto, trata-se de uma lista de tamanho descomunal. Assim, decidimos definir cinco ocorrências como o ponto de corte, isto é, todos os arquivos com seis ocorrências ou mais de variantes PB foram determinados como sendo PB (328 arquivos no total) enquanto textos com cinco ocorrências ou menos de variantes PB foram considerados PE (593 arquivos). No entanto, como cinco é um limiar bastante alto, entendemos que o ideal seria uma análise manual adicional, a qual, devido a restrições de tempo, não pôde ser realizada.

Por fim, vale ressaltar que, enquanto os processos acima estavam sendo realizados, encontramos algumas ocorrências esparsas de casos como este: *“ocidentalização da Amazônia&rdquo*. Então criamos um *script* de limpeza para conversão de entidades ausentes (com base nas tabelas do W3C e *WebStandards*), o que corrigiu os problemas imediatamente.

4.2.4.2. Renomeação de arquivos

Como os resultados da seção acima confirmaram que a fonte não poderia ser considerada equivalente à variedade do português, os arquivos tiveram que ser renomeados para incorporar essa nova informação. A solução foi incluir duas letras extras no início do nome, correspondendo à variedade do português, deixando as duas seguintes para representar a fonte da publicação. No final, houve quatro combinações possíveis (negrito e itálico são usados aqui apenas para destacar a motivação dos códigos):

Br*BR*: Português do **Br**asil publicado na SciELO-*Br*
Eu*BR*: Português **Eu**ropeu publicado na SciELO-*Br*
Eu*PT*: Português **Eu**ropeu publicado na SciELO-*Pt*
Br*PT*: Português do **Br**asil publicado na SciELO-*Pt*

Os nomes de arquivos no *corpus* contêm todos os metadados necessários para uma pesquisa avançada de *corpus*:

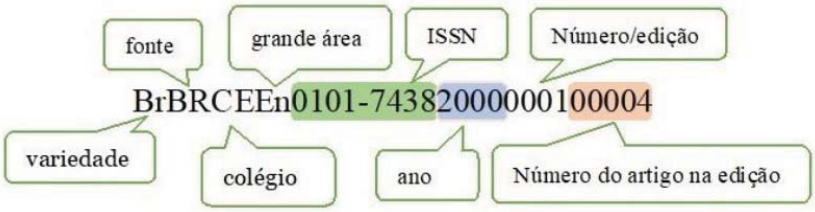


Figura 6 – Codificação no nome do arquivo CoPEP

Neste estágio da compilação, os 213 arquivos manualmente convertidos de PDF foram adicionados ao *corpus* e classificados em *subcorpora* de variedade de acordo com a identificação manual que já havia sido feita anteriormente.

4.2.4.3. Equilíbrio

Primeira versão do *corpus*: Como se sabe, o *subcorpus* PE era muito menor que o *subcorpus* PB. Para balanceamento, os seguintes critérios de exclusão foram seguidos:

- textos anteriores ao ano 2000.
- textos cuja variedade era diferente da sua fonte.
- arquivos menores.

Finalmente, depois de separar os textos dos *subcorpora*, renomear os arquivos e proceder com o processo de balanceamento, estava pronta a primeira versão do *Corpus* de Português Escrito em Periódicos - CoPEP (CoPEP_v.1), com 10.491 textos e 44.072.203 palavras, sendo 22.036.438 de PB e 22.035.765 de PE.

DELTA
36.2
2020

5. Aprimorando o *corpus*

A decisão de melhorar ainda mais o *corpus* levou à análise manual dos arquivos BrPT no CoPEP_v1. Quatro dos 113 arquivos BrPT foram identificados como sendo realmente escritos em PE, que foram então renomeados. Aproveitamos e fizemos mais algumas limpezas manuais, removendo títulos duplos, resumos, palavras-chave, dados do autor, entre outras informações indesejadas.

Além disso, mais algumas análises foram realizadas nos arquivos que haviam sido automaticamente determinados como PE, de acordo com o critério "ocorrência de cinco ou menos variantes de PE no *sub-corpus* PT" (cf. 4.2.4.1 acima). O objetivo foi confirmar manualmente uma correspondência entre a variedade identificada e aquela escrita. A primeira rodada de análises resultou na identificação de 51 arquivos escritos de fato em PB, que foram então renomeados. Em outros, para acelerar esse processo, adotamos um procedimento de identificação de variedade "invertido". Desta vez, os textos PE foram inspecionados com *stoplists* de variantes PE e terminações fonológicas PE, de forma a confirmar correspondência entre a variedade identificada e a fonte.

Depois que o balanceamento foi executado novamente, um *script* foi executado para extrair informações do nome do arquivo e convertê-lo em cabeçalhos de metadados nos textos, permitindo buscas avançadas em ferramentas de *corpus* com tal função, a partir de critérios como colégio, grande área, ano, etc., como se vê na **Figura 7**.

```
<doc variety="Br" source="BR" school="Ex-Tech-Multi Sciences" great area="Engineering"
issn="0101-7438" year="2000" issue="0001" article_num="00004">
Algoritmo de programação de máquinas individuais com penalidades distintas de
adiantamento e atraso


1. INTRODUÇÃO
Desde o início da difusão de princípios do JIT (Just-In-Time) ' que pode ser
considerado como um sistema de administração industrial relativo ao estoque ' ,
tem crescido a importância da diminuição do estoque no processamento de
produtos. Estamos fazendo mais um esforço nesse sentido, considerando como
característica-chave de nosso trabalho um dos elementos mais importantes do JIT
```

Figura 7 – Cabeçalho com metadados

O procedimento meticuloso de compilação de *corpus* aqui descrito gerou um *corpus* de trabalho, o CoPEP_1.3, que cumpriu plenamente

o propósito de servir de base para a proposta do desenho do DOPU (Kuhn, 2017). Contudo, de forma a tornar público o nosso *corpus*, decidimos fazer ainda duas pequenas melhorias, tendo sempre em mente que os *corpora* podem ser continuamente melhorados e que, portanto, é preciso definir um limite. Assim, definimos o CoPEP_1.4 como a versão final.

5.1. CoPEP_1.4

As novidades dessa versão são: (1) Acréscimo de 41 textos do colégio das Ciências Exatas, da Terra e Multidisciplinar[16], correspondente ao período de 2016 a 2018 — todo o processo descrito anteriormente foi seguido para a compilação desse pequeno *corpus*; (2) uso de marcadores especiais de posição textual (*placeholders*) para fórmulas, imagem, tabela, etc., por exemplo: <img/> — ferramentas de *corpus* como *Sketch Engine* e *WordSmith Tools* ignoram marcadores como esse, o que permite que, ao interrogarmos o *corpus*, as sequências de caracteres entre < > não apareçam nos resultados nem sejam contabilizadas para caracterização estatística do *corpus*.

6. O *Corpus* de Português Escrito em Periódicos: CoPEP

O CoPEP contém 9.900 textos distribuídos em seis grandes áreas agrupadas em três colégios de conhecimento, totalizando 40.424.598 palavras.

A **Tabela 8** apresenta as informações estatísticas sobre o conteúdo do *corpus*. Como podemos ver, os dois *subcorpora*, compostos por textos escritos em português brasileiro e português europeu, respectivamente, são quase do mesmo tamanho. Equilíbrio semelhante entre as variedades também foi mantido em termos de grandes áreas e colégios.

16. Agradecemos à Márcia Dornelles, que colaborou nessa compilação durante um estágio de estudos junto ao CELGA-ILTEC, em outubro e novembro de 2018.

Tabela 8 – CoPEP em números

		Corpus	***Português brasileiro***	***Português europeu***
	Textos	9.900	3.811	6.089
	Palavras	40.424.598	20.250.823	20.173.775
	Tokens *(também para os dados abaixo)*	48.840.337	24.427.255	24.413.082
Colégio de Humanidades		**30.988.552**	**15.460.402**	**15.528.150**
	Ciências Humanas	25.595.789	12.763.135	12.832.654
	Ciências Sociais Aplicadas	5.392.763	2.697.267	2.695.496
Colégio de Ciências da Vida		**16.151.841**	**8.112.981**	**8.038.860**
	Ciências da Saúde	13.540.819	6.797.058	6.743.761
	Ciências Agrícolas	2.611.022	1.315.923	1.295.099
Colégio de Ciências Exatas, da Terra e Multidisciplinar		**1.699.944**	**853.872**	**846.072**
	Ciências Exatas e da Terra	829.983	409.500	420.483
	Engenharia	869.961	444.372	425.589

Cabe ressaltar que o desequilíbrio entre colégios e grandes áreas é reflexo da distribuição de documentos publicados por grande área em cada coleção nacional na plataforma SciELO, como se vê nas **Figuras 8 e 9**[17]. Além disso, é preciso lembrar que todos os textos escritos em inglês foram descartados. Em Ciências da Vida, por exemplo, há diversos periódicos escritos exclusivamente em inglês.

17. Scielo Brazil Analytics; Scielo Portugal Analytics, respectivamente.

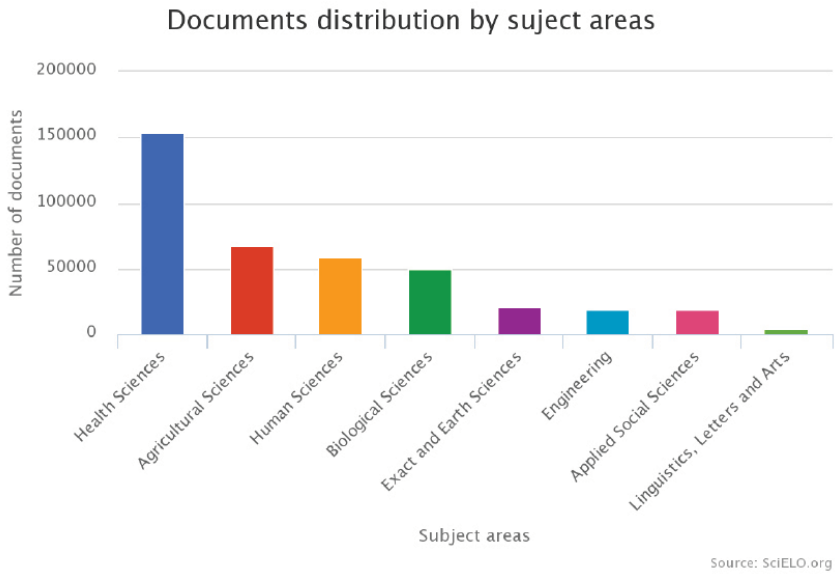


Figura 8 – Distribuição de documentos por grandes áreas - SciELO Brasil

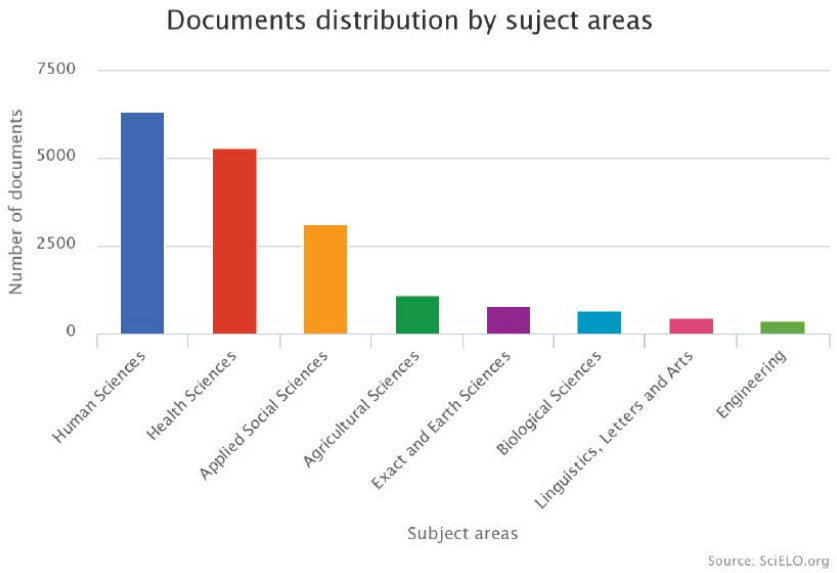


Figura 9 – Distribuição de documentos por grandes áreas - SciELO Portugal

DELTA
36.2
2020

7. Considerações finais

Embora o CoPEP tenha sido especialmente compilado para a elaboração de um dicionário *on-line* de português para universitários (Kuhn, 2017), não temos dúvidas de que esse *corpus* se constitui não apenas em um importante complemento às ofertas existentes de recursos para pesquisas sobre a linguagem acadêmica em português, como também em uma fonte de consulta de qualidade para professores, alunos, linguistas, lexicógrafos e outros interessados pela língua portuguesa. Assim, decidimos disponibilizá-lo publicamente nos programas de análise de *corpus Sketch Engine*[18] e TEITOK (Janssen 2016). Neste último, além da interrogação do *corpus*, é possível também baixá-lo em bruto, ou seja, apenas em formato TXT, sem etiquetação morfossintática, permitindo, assim, que aqueles interessados usem os anotadores que julgarem mais apropriados.

Agradecimentos

Agradecemos especialmente ao Dr. Iztok Kosem pelas conversas muito inspiradoras sobre o desenho do *corpus*. Tanara Zingano Kuhn também lhe agradece pela orientação em uma Missão Científica de Curta Duração (*Short-term Scientific Mission*) (Ação COST *European Network of e-Lexicography*; bolsa COST-STSM-IS1305-210216-071459) na Universidade de Liubliana. Tanara Zingano Kuhn recebeu bolsa CAPES de doutorado pleno no exterior.

Referências bibliográficas

ALMEIDA, Gladis Maria de Barcellos; FERREIRA, José Pedro; CORREIA, Margarita; OLIVEIRA, Gilvan Müller de. 2013. Vocabulário Ortográfico Comum (VOC): constituição de uma base lexical para a língua portuguesa. *Estudos Linguísticos*, 42/1: 204–215.

BENKO, Vladimír. 2015. *Araneum Portugallicum Maius, verze 15.05. Ústav Českého národního korpusu FF UK, Praha 2015.* Disponível em <https://

18. https://www.sketchengine.eu/copep-corpus-of-portuguese-from-academic-journals/

kontext.korpus.cz/first_form?corpname=aranea%2Faranport_pt_ar13__b_a#> Acesso em 23 nov. 2016.

BIBER, Douglas; CONRAD, Susan; REPPEN, Randy. 1998. *Corpus linguistics. Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

BICK, Eckhard. 2000. *The parsing system Palavras, Automatic grammatical analysis of Portuguese in a constraint grammar framework*. Doctoral dissertation, Aarhus University.

CAPES – Tabela de áreas do conhecimento. Disponível em <http://www.capes.gov.br/avaliacao/instrumentos-de-apoio/tabela-de-areas-do-conhecimento-avaliacao> Acesso em 4 fev. 2016.

CAPES. Disponível em <www.capes.gov.br> Acesso em 4 fev. 2016.

Centro de Linguística da Universidade de Lisboa (CLUL). Recursos *on-line*. Disponível em <http://clul.ul.pt/en/resources> Acesso em 20 nov. 2016.

Centro de Linguística da Universidade de Lisboa (CLUL). *CRPC - Corpus de Referência do Português Contemporâneo*. Disponível em <http://alfclul.clul.ul.pt/CQPweb/crpcfg16/> Acesso em 23 nov. 2016.

Centro de Pesquisas, Recursos e Informação de Linguagem (CEPRIL), Programa de Pós-Graduação em Linguística Aplicada (LAEL) da PUCSP. *Corpus Brasileiro*. Disponível em <http://corpusbrasileiro.pucsp.br/cb/Acesso.html>. Acesso em 20 nov. 2016.

Corpus BNC (British National Corpus) - Disponível em <http://www.natcorp.ox.ac.uk/> Acesso em 23 nov. 2016.

Corpus do Português: genre/historical. Disponível em <www.corpusdoportugues.org/hist-gen/> Acesso em 20 nov. 2016.

GANTAR, Polona; KOSEM, Iztok; KREK, Simon. 2016. Discovering automated lexicography: The case of the Slovene lexical database. *International Journal of Lexicography*, 29/2: 200–225.

JANSSEN, Maarten. 2016. TEITOK: Text-Faithful Annotated Corpora. *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC 2016), Portorož, Slovenia.

KILGARRIFF, Adam; RUNDELL, Michael. 2002. Lexical profiling software and its lexicographic applications: A Case study. In BRAASCH, Anna; POVLSEN, Claus (Eds.). *Proceedings of the 10th EURALEX International Congress*. København: Center for Sprogteknologi.

KILGARRIFF, Adam; RYCHLÝ, Pavel; SMRZ, Pavel; TUGWELL, David. 2004. The Sketch Engine. In: WILLIAMS, Geoffrey; VESSIER, Sandra (Eds.). *Proceedings of the 11th EURALEX*

DELTA
36.2
2020

International Congress. Lorient : Université de Bretagne-Sud, Faculté des lettres et des sciences humaines.

KLOSA, Annette. 2013. The lexicographical process (with special focus on *on-line* dictionaries). In: GOUWS, Rufus. H. ; HEID, Ulrich; SCHWEICKARD, Wolfgang; WIEGAND, Herbert. E. (Eds.). *Dictionaries. An International Encyclopedia of Lexicography. Supplement Volume: Recent Developments with Focus on Electronic and Computational Lexicography*. Berlin, Boston: De Gruyter.

KUHN, Tanara Zingano; KOSEM, Iztok. 2016. Devising a sketch grammar for academic Portuguese. *Slovenščina 2.0,* 4/1: 124-161.

KUHN, Tanara Zingano. 2017. A design proposal of an *on-line* corpus driven dictionary of Portuguese for university students. Tese de doutorado. Universidade de Lisboa.

MCENERY, Tony; XIAO, Richard; TONO, Yukio. 2006. *Corpus-Based Language Studies: An Advanced Resource Book*. Abingdon: Routledge.

MEYER, Charles F. 2002. *English Corpus Linguistics: An Introduction*. Cambridge: Cambridge University.

PADRÓ, Lluís; STANILOVSKY, Evgeny. 2012. FreeLing 3.0: Towards wider multilinguality. *Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA*, 1–7.

PEIXOTO, Rafael Marcos Tort. 2015. *O Fenômeno (De)Queísta No Corpus do Português Brasileiro Acadêmico*. Dissertação de Mestrado. Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS).

RUNDELL, Michael (Ed.). 2002. *The Macmillan English Dictionary for Advanced Learners*. Oxford: Macmillan Publishers Limited.

SANTOS, Solange. M.; PACKER, Abel L. 2014. Production of SciELO collections and journals. In: PACKER, Abel L.; COP, Nicholas; LUCCISANO, Adriana; RAMALHO, Amanda; SPINAK, Ernesto (Eds.). *SciELO: 15 Years of Open Access - An Analytic Study of Open Access and Scholarly Communication*. UNESCO.

Scielo Brazil Analytics. Disponível em < http://analytics.scielo.org/w/publication/article?collection=scl > Acesso em 24 nov. 2016.

Scielo Brazil. Disponível em <www.scielo.br> Acesso em 15 fev. 2016.

Scielo Portugal Analytics. Disponível em <http://analytics.scielo.org/w/publication/article?collection=prt> Acesso em 24 nov. 2016.

Scielo Portugal. Disponível em <www.scielo.mec.pt> Acesso em 1 fev. 2016.

Scielo. Disponível em <www.scielo.org> Acesso em 23 nov. 2016.

SINCLAIR, John. 2003a. Corpora for lexicography. In: STERKENBURG, Piet van (Ed.). *A Practical Guide to Lexicography*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

_____. 2003b. Corpus processing. In: STERKENBURG, Piet van (Ed.). *A Practical Guide to Lexicography*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Sketch Engine. Disponível em <https://www.sketchengine.co.uk> Acesso em 20 nov. 2016.

Sketch Engine. *Portuguese Web 2011 (ptTenTen11, Freeling v3)*. Disponível em <https://the.sketchengine.co.uk/bonito/corpus/corp_info?corpname=preloaded/pttenten11_freeling_v3_1> Acesso em 23 nov. 2016.

Sketch Engine. *Portuguese Web 2011 (ptTenTen11, Palavras parsed)*. Disponível em <https://the.sketchengine.co.uk/bonito/corpus/first_form?corpname=preloaded/pttenten11> Acesso em 6 abr. 2016.

SUMMERS, Della. 1993. Longman / Lancaster English Language Corpus - criteria and design. *International Journal of Lexicography*, 6/3: 181-208.

WYNNE, Martin (Ed.). 2005. *Developing Linguistic Corpora: A Guide to Good Practice*. Oxford: Oxbow Books. Disponível em <http://ahds.ac.uk/linguistic-corpora/>. Acesso em 25 jun. 2017.

Recebido em: 16/12/2018
Aprovado em: 10/02/2020