Uma coleção de artigos científicos de Português compondo um *Corpus* no domínio educacional

LUÍS HENRIQUE G. DE AGUIAR

Mestrando no Programa de Pós-graduação em Educação da Universidade Federal dos Vales do Jequitinhonha e Mucuri (UFVJM). E-mail: luishenriqueaguiar@gmail.

VALDIR JR. CORDEIRO ROCHA

Mestrando no Programa de Pós-graduação em Educação da Universidade Federal dos Vales do Jequitinhonha e Mucuri (UFVJM). E-mail: vjcordeiror@gmail.com

MARCUS VINÍCIUS C. GUELPELI

Professor Doutor da Universidade Federal dos Vales do Jequitinhonha e Mucuri (UFVJM). E-mail: marcus.guelpeli@ufvjm.edu.br

Resumo: Este artigo descreve a construção de um *corpus* de textos, o qual é formado por artigos científicos, no domínio educacional, trazendo as estatísticas que o compõem. Com este trabalho, pretende-se obter um *corpus* que torne possíveis diversas pesquisas na área de Processamento de Linguagem Natural e especificamente na área de Sumarização Automática, para possibilitar análise da performance de sumarizadores na língua portuguesa.

Palavras-chave: Corpus. Linguística Computacional. PLN. Sumarização.

Abstract: This article describes the construction of a Portuguese texts corpus in the educational domain. The corpus consists of scientific articles. The work also brings the statistics that make up this corpus. This work aims to obtain a corpus which make possible several research in natural language processing area and specifically in Automatic Summarization area to enable analysis of summarizers performance in Portuguese.

Keywords: Corpus. Computational Linguistics. PLN; Summarization.

Introdução

O termo *corpus* é usado para fazer referência a uma coleção cujos textos são escritos, armazenados eletronicamente e processados por computador com propósitos de pesquisa linguística. Ao construir um *corpus* de textos, procura-se fazer uma seleção de dados representativa, isto é, que constitua um corpo de evidências linguísticas que possa suportar generalizações e contra as quais se possam testar hipóteses (MARTINS, CASELI e NUNES, 2001).

Segundo Zavaglia e Ferraresi (2006), a Linguística de *Corpus* (LC) exerce ampla influência na pesquisa linguística, principalmente na Europa. No Brasil, esse tipo de pesquisa ainda está em estágio inicial, mas não deixa de ocorrer, por exemplo, em centros especializados em Processamento de Linguagem Natural, Lexicografia e Linguística Computacional.

Este trabalho relata a construção de um *corpus*, que adota a língua portuguesa, de domínio educacional, formado por 10 subcategorias. Nas pesquisas, utilizaram-se repositórios acadêmicos e adotou-se um critério segundo o qual os arquivos deveriam possuir resumos e palavras-chave formados por seus autores, já que se pretende com este *corpus* realizar futuras pesquisas na área de Sumarização Automática (SA).

Este trabalho está dividido em cinco seções. A seção dois define e classifica o termo *corpus*; na seção três, apresenta-se a metodologia aplicada na construção do *corpus* e a quarta seção expõe suas estatísticas.

Corpus

Segundo Rebechi e Andreetto (2015), a LC é uma disciplina que se ocupa da investigação de unidades convencionais da língua em um texto ou conjunto de textos em formato eletrônico, ou seja, daquilo que é mais provável de ocorrer na língua, não necessariamente daquilo que é apenas possível. Uma das vantagens da observação dos dados por meio de corpora é que ela revela evidências empíricas e sistemáticas da linguagem, as quais poderiam passar despercebidas sem o auxílio de ferramentas computacionais específicas para esse tipo de investigação. Além disso, uma pesquisa puramente intuitiva pode levar o pesquisador a "encontrar" dados que simplesmente "comprovem" suposições ou hipóteses previamente estabelecidas, sem revelar dados novos que possam levar a outras descobertas.

Segundo Oliveira e Guelpeli (2014), a palavra *corpus* (plural: corpora) tem origem no latim corpo, conjunto de textos, que em LC determina uma coleção de textos selecionados e organizados. Segundo Sardinha (2000), a Linguística de *Corpus* ocupa-se da coleta e exploração de corpora, ou conjunto de dados coletados de forma criteriosa com a finalidade de serem utilizados para pesquisa de uma língua ou variedade linguística.

Em 1961 foi construído o primeiro *corpus* linguístico eletrônico denominado *Brown* (KUCERA e FRANCIS, 1961), formado por 1 milhão de palavras, impulsionando o desenvolvimento da LC. Para Aluísio e Almeida (2006), um *corpus* computadorizado observa um conjunto de considerações que influencia a validade e confiabilidade da pesquisa baseada em *corpus*, sendo a sua criação um processo repetitivo. Tal processo começa com a seleção dos textos, baseada em algum critério significativo para a pesquisa (critério externo), continua com as investigações empíricas da língua ou variedade linguística sob análise (critério interno) e, por fim, tem-se a revisão de todo o projeto.

Para Humblé (2001), um *corpus* pode ser compreendido como uma quantidade grande de textos armazenados no computador e que são acessados com programas próprios de pesquisa. Esses dados podem ser variados, como jornais, bulas de remédios ou artigos científicos, sendo

obtidos de diversas fontes, dependendo do contexto e necessidade, sendo os mais comuns os textos digitalizados e retirados da internet.

Metodologia

Esta seção descreve a metodologia de construção do que foi proposto, adotando a descrita por Aluísio e Almeida (2006), que divide em três estágios a compilação de um *corpus* próprio. Primeiramente é a fase de projeto, que inclui a seleção dos textos, logo após realiza-se a captura, manipulação, nomeação dos arquivos e, por fim, efetua a anotação. A manipulação dos textos, que abrange a limpeza e formatação do *corpus* para o processamento computacional, bem como a organização estrutural deste, segue a metodologia proposta por Guelpeli (2012).

A compilação do *corpus* teve duração de quatro meses (dezembro de 2015 a março de 2016) e para tanto foi utilizada a tabela de áreas de conhecimento da grande área da Educação, para a definição de dez subáreas que compõem as categorias do *corpus*. A referida tabela encontra-se disponível no site da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) [1]. As categorias escolhidas foram: Educação Especial; Educação Permanente; Educação Pré-escolar; Ensino-aprendizagem; Filosofia da Educação; História da Educação; Política Educacional; Psicologia Educacional; Sociologia da Educação e Tecnologia Educacional.

Na fase de projeto, denominada por Aluísio e Almeida (2006) critério externo, ocorreu a seleção dos textos, observando os critérios de (1) gratuidade; (2) possibilidade de reprodução dos arquivos originais; (3) classificação das bases para o domínio e subáreas escolhidas para a pesquisa; (4) resumo do texto original, denominado sumário de referência, elaborado pelo autor. Foram escolhidos somente artigos científicos, por possuírem resumo e palavras-chave, que poderão ser utilizados posteriormente para testes com sumarizadores automáticos de texto. Os artigos científicos foram retirados, em sua maioria, do repositório *Scientific Electronic Libary online* (SCIELO)[2], nas categorias já mencionadas. As categorias em que não houve preenchimento total pelos artigos do repositório da SCIELO – Educação permanente; Ensino aprendizagem; Filosofia da Educação; Política Educacional e Sociologia da Educação – foram preenchidas por

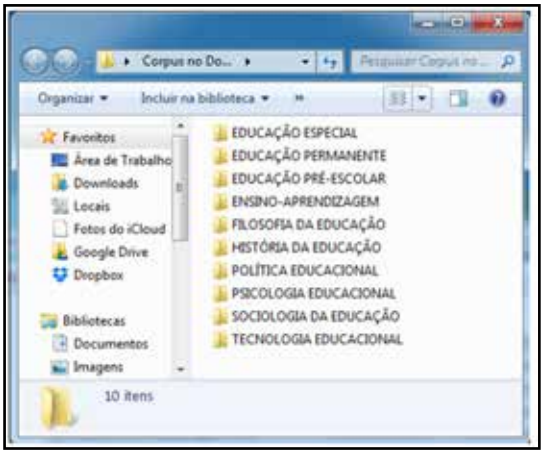
1Disponível em: http://www.capes.gov.br/avaliacao/instrumentos-de-apoio/tabela-de-areas-do-conhecimento-avaliacao

2 Disponível em: http://www.scielo.br/

artigos encontrados no repositório Buscador Coruja[3], além disso, o domínio Educação Pré-escolar também foi preenchido com artigos encontrados segundo pesquisa com termos relevantes feitas no Google.

Na segunda fase, conforme realizado por Guelpeli (2012), ocorreram a limpeza e a formatação do *corpus* para o processamento computacional. Foi retirado tudo que não fazia parte do texto – gráficos, tabelas, figuras e números de páginas. Os arquivos fontes foram convertidos do formato em "PDF" para o "TXT", que é compatível para o processamento. Para cada uma das 10 categorias que compõem o domínio, criou-se uma subpasta, conforme exemplifica a Figura 1.

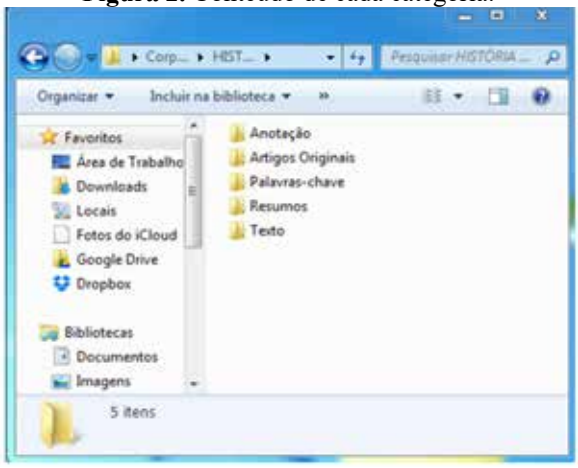
Figura 1. Categorias que compõem o domínio.



Dentro de cada categoria, o *corpus* foi divido em cinco pastas, conforme mostra a figura 2. Na primeira pasta, "Anotação", contém as estatísticas e referências externas do texto; na segunda, "Artigos Originais", armazenam-se os artigos originais no formato PDF. Na pasta "palavras-chave" são armazenadas aquelas escolhidas por cada autor do artigo, essas palavras podem ser utilizadas no processo de sumarização com a finalidade de melhorar a qualidade do sumário. A pasta "Resumos" traz os resumos manuais de cada artigo, estes são importantes para realizar a avaliação dos sumários automáticos, e, por fim, a pasta "Textos" armazena os corpos dos artigos, que serão submetidos aos sumarizados.

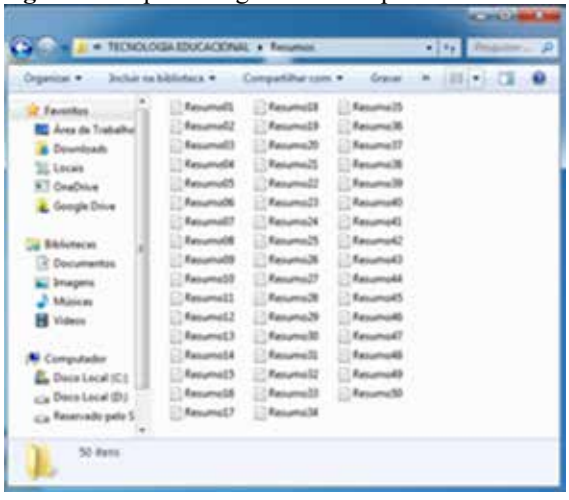
3 Disponível em: https://buscadorcoruja.com/

Figura 2. Conteúdo de cada categoria.



Os arquivos dentro das pastas “Palavras-chave”, “Resumos” e “Texto” foram nomeados segundo um padrão no qual o nome do arquivo é formado pelo nome da pasta que ele está inserido e um número de 01 a 50, dessa maneira os arquivos referentes a um mesmo artigo têm o mesmo número. Assim, o artigo armazenado na pasta “Artigos Originais” com o número 02 terá suas respectivas informações nas outras pastas com o nome “chave02”, “resumo02” e “texto02”. Esse padrão pode ser visualizado na figura 3.

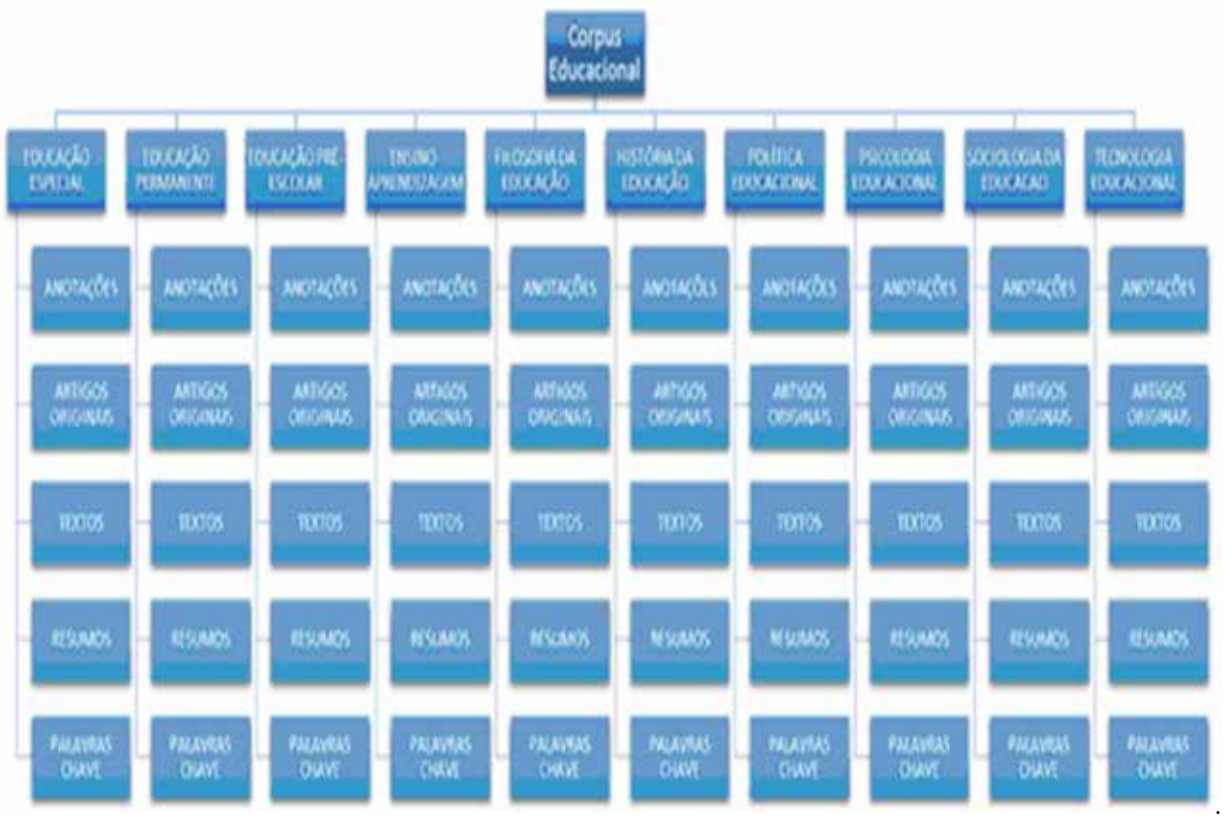
Figura 3. Arquivos organizados na pasta “Resumos”.



Na figura 4 é apresentado o diagrama do *Corpus* em Educação. O *corpus* possui 500 arquivos de texto, sendo 50 textos para cada categoria.

Concluindo, na última fase tem-se a anotação, que se compreende como a retirada dos métodos que, segundo Aloísio e Almeida (2006), são dados estruturados sobre dados, isto é, dados bibliográficos comuns, de catalogação, tais como tamanho do arquivo, tipo da autoria, a tipologia textual e informação sobre a distribuição do *corpus*, neste caso específico, a retirada dos dados estatísticos do *corpus* gerado. As referências externas deste *corpus* se encontram nas pastas "Anotação".

Figura 4. Diagrama do *Corpus* em Educação



Estatística do *corpus*

Esta seção mostra as estatísticas do *corpus* em educação. Segundo Peixoto e Brito (2015), a LC e a existência de corpora de grande tamanho seriam inconcebíveis sem que houvesse o auxílio de ferramentas computacionais, uma vez que a área trabalha com grandes quantidades de informação textual, impossíveis de se processar manualmente com rapidez, mesmo que por grandes equipes. Portanto, são usados programas de análise lexical para efetuar operações de processamento da linguagem, tais como contagem de palavras, geração de listas de frequência, geração de listas de palavras-chave e exibição de linhas de concordância.

Neste trabalho, as informações estatísticas do *corpus* foram coletadas através do software *FineCount* 2.6 *free*[4]. O *corpus* é formado por 2.999.646 palavras no total, cuja distribuição se dá nos 500 artigos selecionados. A tabela 1 sintetiza as estatísticas dos textos separadas nas 10 categorias que compõem o *corpus*, mostra também o número de palavras por categorias e o número médio de palavras por texto de cada uma destas. De acordo com a tabela, a média de palavras por categoria é 299964,6 e a média por texto é de 5999,29.

Tabela 1. Estatísticas dos textos fonte do *corpus*.

Arquivos	Caracteres	Caracteres e espaços	Palavras	Palavras e numerais	Sentenças	Média de palavras por texto
EDUCAÇÃO ESPECIAL	1389371	1649313	274182	281052	10267	5483,64
EDUCAÇÃO PERMANENTE	1391696	1652722	281035	285382	14274	5620,7
EDUCAÇÃO PRÉ-ESCOLAR	1442290	1722467	281091	286480	23248	5621,82
ENSINO-APRENDIZAGEM	1374495	1634176	275404	279809	13299	5508,08
FILOSOFIA DA EDUCAÇÃO	1514556	1810378	313779	317365	18554	6275,58
HISTÓRIA DA EDUCAÇÃO	1918883	2283568	392515	400849	14646	7850,3
POLÍTICA EDUCACIONAL	1846727	2188932	371540	379529	12607	7430,8
PSICOLOGIA EDUCACIONAL	1365382	1616528	266666	271889	13157	5333,32
SOCIOLOGIA DA EDUCAÇÃO	1643482	1952429	329689	334182	17144	6593,78
TECNOLOGIA EDUCACIONAL	1074809	1275323	213745	218345	7688	4274,9
Total	**14961691**	**17785836**	**2999646**	**3054882**	**144884**	**59992,92**
Desvio Padrão	**249063,72**	**296407,39**	**53007,45**	**54042,24**	**4361,13**	**1060,14**
Média Geral	**1496169,1**	**1778583,6**	**299964,6**	**305488,2**	**14488,4**	**5999,29**

4 Disponível em: http://www.tilti.com/software-for-translators/finecount/

Na Tabela 2, são mostradas as estatísticas dos sumários manuais dos 10 domínios que formam o *corpus*, há um total de 168993 palavras e uma média geral de 3379,86 por texto.

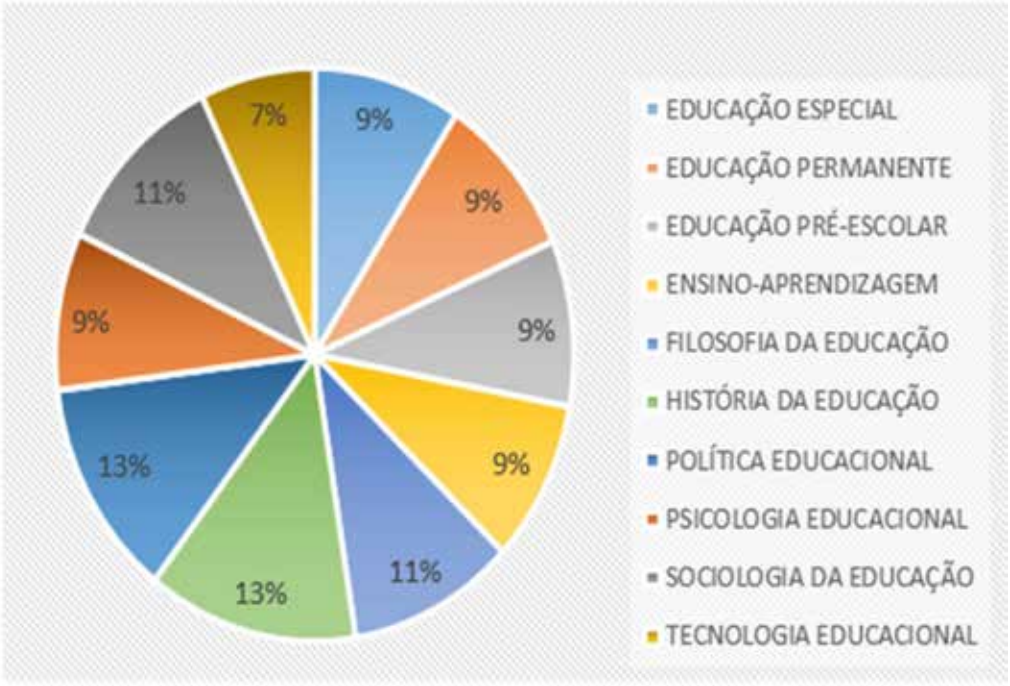
Tabela 2. Estatísticas dos Sumários Manuais

Arquivos	Caracteres	Caracteres e espaços	Palavras	Palavras e numerais	Sentenças	Média de palavras por texto
EDUCAÇÃO ESPECIAL	116123	135793	22633	22899	723	452,66
EDUCAÇÃO PERMANENTE	91635	107240	18425	18654	628	368,5
EDUCAÇÃO PRÉ-ESCOLAR	81817	96157	16054	16279	760	321,08
ENSINO-APRENDIZAGEM	88281	103381	17413	17593	553	348,26
FILOSOFIA DA EDUCAÇÃO	77857	92091	16240	16304	595	324,8
HISTÓRIA DA EDUCAÇÃO	71999	84727	14804	14984	420	296,08
POLÍTICA EDUCACIONAL	76407	89677	15506	15694	429	310,12
PSICOLOGIA EDUCACIONAL	81145	94957	15594	15826	693	311,88
SOCIOLOGIA DA EDUCAÇÃO	77136	90700	15320	15368	675	306,4
TECNOLOGIA EDUCACIONAL	88193	103060	17004	17341	603	340,08
Total	**850593**	**997783**	**168993**	**170942**	**6079**	**3379,86**
Desvio Padrão	**12535,78**	**14464,25**	**2289,02**	**2326,87**	**115,01**	**45,78**
Média Geral	**81481**	**95557**	**16147**	**16291,5**	**615,5**	**322,94**

A Figura 5 mostra a disposição da porcentagem do total de palavras bem como os números dos textos fonte das 10 categorias que constituem o *corpus*. A variação do tamanho dos textos fonte é pequena, conforme mostrado no gráfico. Esse balanceamento, de acordo com

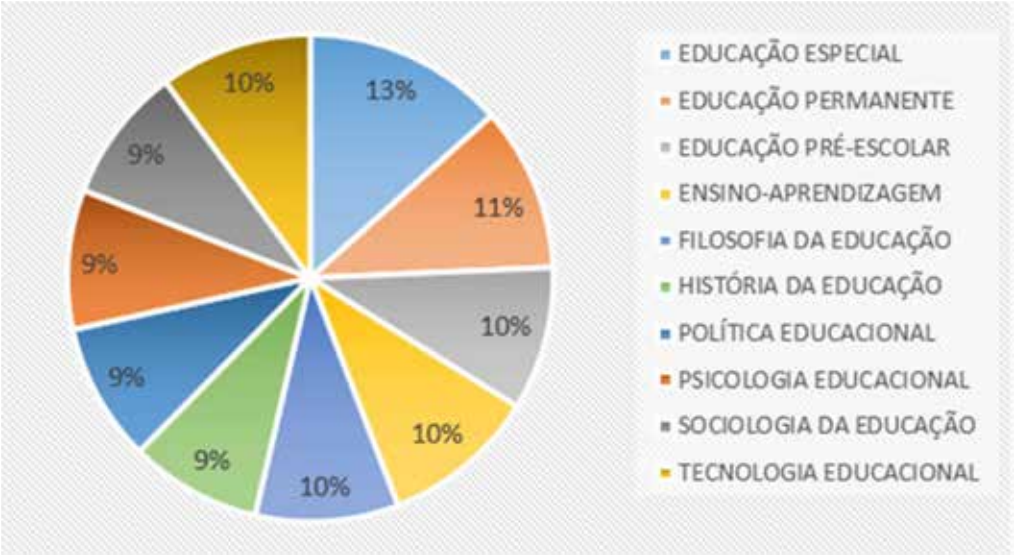
Aloísio e Almeida (2006), é um requisito que impacta na validade e confiabilidade do *corpus* computadorizado. A categoria Tecnologia Educacional possui os textos com a menor quantidade de palavras e numerais, por sua vez, as categorias Política Educacional e História da Educação possuem os maiores textos.

Figura 5. Divisão da porcentagem de palavras e números por categoria dos textos fontes.



A Figura 6 traz as porcentagens desta distribuição para os sumários manuais. Mostra um balanceamento entre a quantidade de palavras e números das 10 categorias do *corpus*.

Figura 6. Divisão da porcentagem de palavras e números por categoria dos sumários Manuais



Conclusão

Este artigo descreve os procedimentos para a construção e as estatísticas de um *corpus* que foi construído a partir de artigos científicos na língua portuguesa e de domínio educacional. Espera-se poder contribuir com a comunidade científica ao disponibilizar este *corpus* para simulação.

O *corpus* construído neste trabalho faz parte dos estudos desenvolvidos no grupo de pesquisa Mineração de Textos e Processamento de Linguagem Natural e Aprendizado de Máquina (MTPLNAM), que também já produziu outros corpora, conforme os trabalhos de Oliveira e Guelpeli (2014), Fernandes e Guelpeli (2014) e Guelpeli e Fernandes (2016).

O grupo MTPLNAM tem como objetivo pesquisar, gerar conhecimentos e desenvolver aplicações sobre mineração de texto (MT), processamento de linguagem natural (PLN) e aprendizagem de máquina (AM), sua página pode ser acessada em http://www.mtplnam.com.br,

onde estão outros corpora produzidos pelo grupo e disponibilizados para a comunidade científica para testes e trabalhos.

A necessidade deste *corpus* é viabilizar os estudos na área de sumarização automática e possibilitar que sumarizadores possam ser avaliados por ferramentas especializadas, como o ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) (LIN e HOVY, 2003), além de dar suporte a sumarizadores que utilizam palavras-chave para personalizar seus sumários, como é feito no sumarizador PragmaSUM (ROCHA, 2014), que também é um projeto do grupo MTPLNAM.

Assim, este *corpus* tem a finalidade de contribuir com a continuidade dos estudos da eficiência de sumarizadores, com possibilidade de utilização das palavras-chave no processo de sumarização dos textos, além de viabilizar estudos da influência destas palavras nos conteúdos dos textos em questão.

REFERÊNCIAS

ALUÍSIO, S.M.; ALMEIDA, G.M.B. **O que é e como se constrói um corpus? Lições aprendidas na compilação de vários corpora para pesquisa linguística.** Calidoscópio, (UNISINOS). vol. 4, n. 3, p. 155-177, set/dez 2006.

FERNANDES, H M; GUELPELI, M. V. C. **Creación de corpus en lengua española para su utilización en testes acerca de Sumarización Automática**. In: 6th International Conference on Corpus Linguistics-CILC 2014, 2014, Las Palmas de Gran Canaria. 6th International Conference on Corpus Linguistics-CILC 2014, 2014.

GUELPELI, M. V. C.; FERNANDES, H. M. **Input a Word, Analyze the World: Selected Approaches to Corpus Linguistics...**ISBN-13: 978-1-4438-8513-3 e ISBN-10: 1-4438-8513-4 1ª. ed. United Kingdom: Cambridge Scholars Publishing, 2016. v. I. 521p

GUELPELI, M.V.C. (2012) **"Cassiopeia: Um Modelo de Agrupamento de Textos Baseado em Sumarização**". http://nlx.di.fc.ul.pt/~guelpeli/Arquivos/Tese.pdf.

HUMBLÉ, Philippe. **Dictionaries and Language Leaners.** Frankfurt am Main, 2001.

KUCERA, H and FRANCIS W. N. **Brown University Standard Corpus of Present-Day American English (or just Brown Corpus) as a general corpus (text collection) in the field of corpus linguistics.** It contains 500 samples of English-language text, totalling roughly one million words, compiled from works published in the United States in 1961.

LIN, C-Y. and HOVY, E.H. **Automatic Evaluation of Summaries Using N-gram Cooccurrence Statistics.** In the Proceedings of Language Technology Conference – HLT. Edmonton, Canadá, 2003.

MARTINS, M S. CASELI, H M. NUNES, M G V N. **A construção de um corpus de textos paralelos inglês-português.** Série de Relatórios do Núcleo Interinstitucional de Linguística Computacional NILC - ICMC-USP, Caixa Postal 668, 13560-970 São Carlos, SP, Brasil, Set. 2001.

OLIVEIRA, R.R.; GUELPELI, M.V.C. **Building a Corpus in Italian Written Language.** In:- 6th International Conference on Corpus Linguistics (CILC2014). Las Palmas de Gran Canaria, Espanha, 2014. No prelo.

PEIXOTO, L. M.; BRITO, L. F. A. **Procedimentos para compilação de um corpus composto por legendas e construção de uma ferramenta de corpus on-line: o Corpus of English Language Videos**. DOMÍNIOS DE LINGU@GEM. <http://www.seer.ufu.br/index.php/dominiosdelinguagem> - v. 9, n. 3 (jul/set. 2015) - ISSN 1980-5799.

REBECHI, Rozane R.; ANDREETTO, Marlene D. **As retraduções de Trauerund Melancholie para o português: o léxico freudiano sob o olhar da Linguística de Corpus.** Pandaemoniumger., São Paulo, v. 18, n. 26, p. 126-157, Dec. 2015. Available from <http://dx.doi.org/10.1590/1982-88371826126157>. access on 30 May 2016.

ROCHA, V. J. C., **PragmaSUM: Um Sumarizador Automático De Textos Baseado Em Perfil De Usuário.** Trabalho de Conclusão de Curso de Sistemas de Informação da Universidade Federal dos Vales do Jequitinhonha e Mucuri. Diamantina- MG, Brasil, 2014.

SARDINHA, T. B. (2000) **"Linguística de Corpus: Histórico e Problemática".** DELTA, São Paulo, v. 16, n. 2, p 323-367, 2000a <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-44502000000200005>

ZAVAGLIA, C. FERRARESI, M. L. **Construção de um corpus paralelo e alinhado português italiano-português para o domínio literário.** Estudos Linguísticos XXXV, p. 502-511, 2006.