Universidade de São Paulo - USP
Universidade Federal de São Carlos - UFSCar
Universidade Estadual Paulista - UNESP

Um exercício em Linguística de *Corpus* no âmbito do Projeto TermiNet

Jackson Wilke da Cruz Souza
Ariani Di Felippo

NILC-TR-10-08

Agosto, 2010

Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional
NILC - ICMC-USP, Caixa Postal 668, 13560-970 São Carlos, SP, Brasil

Resumo

Neste relatório, descrevemos as atividades realizadas no projeto "*Construção do corpus para o desenvolvimento de uma wordnet terminológica em português do Brasil*", o qual integra o projeto "*Instanciação e aplicação de uma metodologia para o desenvolvimento de wordnets terminológicas em português do Brasil*" (simplesmente, projeto TermiNet). Um dos objetivos do projeto TermiNet é a instanciação ou especificação da metodologia genérica de pesquisa no Processamento Automático das Línguas Naturais, proposta por Dias-da-Silva (2006), para a construção de um tipo específico de recurso linguístico-computacional: as bases lexicais terminológicas no formato *wordnet* ou *terminets* (do inglês, *terminological wordnets*). A estratégia de pesquisa de Dias-da-Silva, denominada aqui de "metodologia trifásica", destaca-se principalmente por equacionar todo empreendimento no PLN em três fases: a linguística, representacional e implementacional. Na fase linguística, em especial, a metodologia instanciada prevê as seguintes tarefas: (i) delimitação do domínio especializado para o qual a *terminet* será construída; no nosso caso, o domínio escolhido foi o da Educação a Distância (EaD); (ii) delimitação das fontes e estratégias de extração do conhecimento necessário a construção de uma *terminet*; (iii) delimitação do conhecimento léxico-conceitual a ser extraído das fontes que caracteriza as bases *wordnets;* e (iv) extração do conhecimento. Quanto à tarefa de delimitação das fontes, optou-se no projeto TermiNet pelas fontes não-estruturadas, os *corpora* textuais. No caso do domínio da EaD, foi necessária a construção de um *corpus.* O relatório ora apresentado relata essa construção do *Corpus.*EaD. Especificamente, relatamos as questões teóricas e metodológicas envolvidas pela tarefa de construção do *corpus,* assim como destacamos os resultados alcançados.

Este trabalho contou com o apoio financeiro
da FAPESP e do CNPq/UFSCar.

Sumário

Resumo ii

1. INTRODUÇÃO 1
2. *CORPUS*: DEFINIÇÃO E CARACTERIZAÇÃO GERAL 3
3. *CORPUS* E TERMINOLOGIA 4
4. ETAPAS DE CONSTRUÇÃO DO *CORPUS* 5
   - 4.1. Projeção do *Corpus* 5
     - 4.1.1. Critérios provenientes da definição do objeto *corpus* 6
     - 4.1.2. Critérios provenientes do tipo de recurso a ser construído a partir do *corpus* 7
     - 4.1.3. Critérios provenientes de decisões de projeto 8
   - 4.2. Compilação do *corpus* 9
     - 4.2.1. Identificação das fontes e coleta dos textos 9
     - 4.2.2. Identificação (e seleção) das páginas e seleção dos textos 10
   - 4.3. Pré-processamento dos textos 10
     - 4.3.1. Conversão 10
     - 4.3.2. Limpeza 11
     - 4.3.3. Nomeação e Anotação Estrutural 11
     - 4.3.4. Armazenamento 13
   - 4.4. Disponibilização 14
5. DESCRIÇÃO QUANTITATIVA DO *CORPUS*.EaD 15
6. RESULTADOS E DISCUSSÕES 16
7. CONSIDERAÇÕES FINAIS 17

REFERÊNCIAS IBLIOGRÁFICAS 17

APÊNDICE 20

1. INTRODUÇÃO

Na área do Processamento Automático das Línguas Naturais (PLN), buscamos desenvolver, em última instância, sistemas computacionais "capazes" de processar (interpretar/ gerar) as línguas naturais, principalmente em meio escrito (DIAS-DA-SILVA, 2006). Dentre eles, citamos os sistemas de: tradução automática, correção ortográfica e gramatical, sumarização automática, etc. (MITKOV, 2004). Quando baseados em conhecimento linguístico, tais sistemas podem apresentar uma arquitetura composta por três "bases de conhecimento estático": a gramatical, a conceitual e a lexical. À base de conhecimento lexical (ou base lexical), em especial, cabe a tarefa de fornecer ao sistema computacional uma coleção de unidades lexicais da língua que se está processando, juntamente com suas propriedades morfológicas, sintáticas, semânticas e pragmático-discursivas, dependendo da especificidade do sistema (MITKOV, 2004).

No caso do processamento semântico do inglês norte-americano, a Wordnet de Princeton (WN.Pr) (FELLBAUM, 1998) é uma base lexical amplamente utilizada, principalmente por sua adequação científica e tecnológica (MORATO et al., 2004). Diante de sua reconhecida potencialidade tecnológica, a WN.Pr tem motivado a construção de bases lexicais no formato *wordnet* para inúmeras línguas. Atualmente, é possível encontrar wordnets para a maioria das línguas europeias, africanas e asiáticas. A wordnet do português do Brasil (PB), a WordNet.Br (WN.Br) (DIAS-DA-SILVA et al., 2008), está em desenvolvimento.

Na WN.Pr, as unidades lexicais (palavras ou expressões) do inglês norte-americano estão divididas em quatro categorias sintáticas: nome, verbo, adjetivo e advérbio. As unidades de cada categoria estão codificadas em *synsets* (do inglês, *synonym sets*), ou seja, em conjuntos de formas sinônimas ou quase-sinônimas (p.ex.: {car; auto; automobile; machine; motorcar}). Cada *synset* é construído de modo a representar um único conceito lexicalizado por suas unidades constituintes. Assim, não é preciso explicitar o valor semântico de cada conjunto de sinônimos por meio de um rótulo conceitual. Os *synsets* estão inter-relacionados pela relação léxico-semântica da antonímia[1] e pelas relações semântico-conceituais de hiponímia, meronímia, acarretamento e causa. A WN.Pr, assim como todas as demais bases no formato *wordnet,* são construídas segundo um paradigma de representação do conhecimento específico, as redes semânticas (do inglês, *semantic network*) (DI-FELIPPO; DIAS-DA-SILVA, 2010).

Nos últimos anos, dadas as aplicações reais para as quais os sistemas de PLN têm sido projetados, é premente que estes sejam "capazes" de processar textos técnicos ou especializados. Assim, encontramos várias *wordnets* terminológicas, as quais podem ser exemplificadas pelas:

a) JurWordnet (SAGRI et al., 2004) e ArchiWordnet (BENTIVOGLI et al., 2004), responsáveis por enriquecer a *wordnet* do italiano com unidades terminológicas do domínio jurídico e da arquitetura, respectivamente;

[1] A antonímia é uma relação entre unidades lexicais, ou seja, formas linguísticas. A relação de antonímia entre *synsets* (ou conceitos) indica, na verdade, uma oposição conceitual e não propriamente uma antonímia.

b) Medical Wordnet (SMITH; FELLBAUM, 2004) e BioWordnet (POPRAT et al., 2008), que ampliam a WN.Pr para os domínios da medicina e da biomedicina, respectivamente.

Embora exista um número razoável de *wordnets* terminológicas, observa-se a carência de uma metodologia suficientemente clara e genérica que facilite e estimule a criação de bases de conhecimento lexical especializado nesse formato. Diz-se isso porque, nesses trabalhos, são comumente utilizados recursos-fonte estruturados (p.ex.: *thesauri*, dicionários, enciclopédias, etc.) para a construção das bases, as quais nem sempre existem ou estão disponíveis, dependendo do domínio que se pretende sistematizar. Além disso, para a análise dessas fontes, os pesquisadores comumente utilizam métodos manuais, nos quais a extração do conhecimento é lenta e sujeita à subjetividade. Diante desse cenário, está sendo desenvolvido o Projeto TermiNet, cujos objetivos são apresentados a seguir.

O projeto TermiNet (FAPESP 2009/06262-1 / CNPq 471871/2009-5) busca instanciar ou especificar a metodologia genérica de pesquisa no PLN, proposta por Dias-da-Silva (2006), para a construção de *wordnets* terminológicas e validar essa metodologia com a construção de uma *terminet* em PB, no caso, do domínio da Educação a Distância (EaD), a WordNet.EaD. Com base na metodologia genérica de pesquisa no PLN e no formato *wordnet*, a instanciação da metodologia para a construção *terminets* ficou assim delimitada (DI-FELIPPO; ALMEIDA, 2010) (Figura 1):

- Domínio linguístico: (i) seleção e delimitação do domínio de conhecimento especializado (p.ex.: medicina, arquitetura, etc.); (ii) delimitação dos recursos-fonte e da estratégia de extração do conhecimento necessário à criação de uma *wordnet* (p.ex.: dicionários, taxonomias, *corpora*, etc.), e (iii) delimitação e extração do conhecimento léxico-conceitual, ou seja, das categorias sintáticas; das unidades lexicais, das relações lexicais de sinonímia e antonímia, das relações semântico-conceituais de hiponímia, meronímia, acarretamento e causa, das glosas e das frases-exemplo;
- Domínio representacional: representação do conhecimento delimitado no domínio linguístico em um formalismo que seja "computacionalmente tratável"; no caso de uma base *wordnet*, tal representação baseia-se nas noções de *forma lexical* (do inglês, *word form*), *synset*, *matriz lexical* e *ponteiros relacionais* (do inglês, *relational pointers*);
- Domínio implementacional: transformação do conhecimento representado no formato *wordnet* para uma base lexical relacional; especificamente, essa etapa engloba as tarefas de (i) seleção de um editor/sistema de gerenciamento de bases relacionais (SGBDR) (do inglês *relational database management system*) e (ii) inserção dos dados no editor e construção da base.

No que diz respeito às tarefas do domínio linguístico, ressalta-se, em especial, a delimitação dos recursos-fontes para a aquisição do conhecimento léxico-conceitual. As fontes a partir dos quais o conhecimento léxico-conceitual necessário à construção de uma *terminet* pode ser adquirido classificam-se em dois grupos: as estruturadas (p.ex.: dicionário, taxonomia, *thesaurus*, etc.) e as não-estruturadas (ou seja, *corpus* textual)

(RIGAU, 1998). Tendo em vista a carência de fontes estruturadas, principalmente no caso de domínios especializados emergentes ou ainda em formação, o *corpus* torna-se a principal fonte de conhecimento.

Como a validação da metodologia instanciação esta sendo feita por meio da construção da WordNet.EaD, foi necessária a construção de um *corpus* desse domínio. Tal construção fora o objetivo do projeto de IC ora aqui descrito.

Dessa forma, neste relatório, procuramos descrever detalhadamente os pressupostos teóricos adotados, as fases metodológicas empregadas no desenvolvimento do nosso projeto e as estratégias adotadas em cada fase. Além disso, pretendemos apresentar os números finais do *corpus* do domínio da EaD construído, o *Corpus*.EaD.

Para tanto, na Seção 2, definimos e caracterizaremos o objeto *corpus* com base em uma breve revisão teórica. Na Seção 3, destacamos a relação entre os *corpora* e a Terminologia, ou seja, salientamos a relevância da utilização dos *corpora* no desenvolvimento de bases lexicais terminológicas. Na Seção 4, descrevemos as etapas de construção do *Corpus.*EaD. Na Seção 5, o *Corpus.*EaD é descrito em função de diferentes critérios. Na Seção 6, destacamos as principais contribuições do subprojeto ora relatado. E, por fim, na Seção 7, algumas considerações finais são feitas.

2. *CORPUS*: DEFINIÇÃO E CARACTERIZAÇÃO GERAL

Segundo Sinclair (2005), o *corpus* é uma coletânea de textos em certo idioma que esteja em formato eletrônico. Especificamente, esses textos devem ser selecionados de acordos com critérios externos, ou seja, critérios que nascem a partir das necessidades da pesquisa na qual o *corpus* será usado e que sejam capazes de representar uma língua ou uma parcela de língua. Segundo essa definição, podemos dizer que nem toda coletânea de textos se caracteriza como *corpus.* Seguindo as palavras de Berber Sardinha (2004), uma coletânea de textos não é um *corpus* quando: (i) o conjunto de textos não apresenta uma organização prévia; no caso, dizemos apenas que é um Arquivo e não um *corpus;* (ii) o conjunto não obedece certos critérios de seleção, coleta, organização e nomeação; quando isso ocorre, dizemos ser uma Biblioteca Eletrônica.

Dentre as várias definições existentes na literatura, destacamos a de Sanchez (1995), que corrobora a de Sinclair (2005):

> "Um conjunto de dados lingüísticos (pertencentes ao uso oral ou escrito da língua, ou a ambos), sistematizados segundo determinados critérios, suficientemente extensos em amplitude e profundidade, de maneira que sejam representativos da totalidade do uso lingüístico ou de algum de seus âmbitos, dispostos de tal modo que possam ser processados por computador, com a finalidade de propiciar resultados vários e úteis para a descrição e análise" (SANCHEZ, 1995, p.8-9)

Com base na definição de Sanchez, vê-se que o *corpus* é sistematizado em função de determinados critérios. Segundo a maioria dos autores que se preocupa com a definição de *corpus* (p.ex.: KENNEDY, 1998; BIBER et al., 1998; RENOUF, 1998; BERBER SARDINHA, 2004; SINCLAIR, 2005), tais critérios são: representatividade,

amostragem, tamanho, autenticidade, diversidade e balanceamento. Além disso, vê-se na definição de Sanchez que corpus é um artefato produzido para a pesquisa. Em outras palavras, isso quer dizer que, embora os textos devem ser autênticos, o *corpus* é um objeto criado com fins específicos de pesquisa (BERBER SARDINHA, 2004). Com base em Sinclair (2005), Berber Sardinha (2004) e Sanchez (1995), é possível estabelecer certas diretrizes que a compilação ou coleta dos textos deve seguir para resultar em um *corpus*: (i) selecionar textos autênticos por meio de critérios regidos pela pesquisa para a qual o *corpus* está sendo construído e (ii) selecionar textos representativos da língua, modalidade da língua ou domínio.

Quando caracteriza como *corpus,* uma coletânea de textos é inegavelmente uma fonte de conhecimento não-estruturado que permite a extração de dados linguísticos reais e em larga escala (ALUÍSIO, ALMEIDA, 2006).

Na próxima Seção, salientamos a relação entre os *corpora* e a área da Terminologia. Na verdade, destacamos como os *corpora* se inserem no contexto de construção de uma base de conhecimento lexical.

3. *CORPUS* E TERMINOLOGIA

No projeto TermiNet, adotamos os pressupostos teóricos da Teoria Comunicativa da Terminologia (TCT) (CABRÉ, 1999) para a construção da WordNet.EaD. Tal escolha teórica implica em acolher determinadas escolhas metodológicas durante todas as etapas de construção de um produto terminológico, seja ele um glossário, dicionário, mapa conceitual, listas de termos ou mesmo uma base de dados. (ALMEIDA, 2006).

Um projeto que se proponha a aplicar a metodologia da TCT deve se embasar nos seguintes pressupostos que, segundo Almeida (2006), são imprescindíveis para um projeto que se vincula a essa vertente teórica:

a) o objeto central da Terminologia são as unidades terminológicas e não os conceitos; isso significa que se deve eleger as unidades lexicais terminológicas como objeto central, reforçando uma perspectiva linguística da Terminologia e não uma abordagem semasiológica;
b) os termos e as palavra não são em princípio diferentes, o que há são signos linguísticos que podem realizar-se no discurso como termo ou palavra, dependendo da situação comunicativa; neste ponto, ressaltamos as palavras de Cabré (2003): *(…) we postulate that a lexical unit is by itself neither terminological nor general but that it is general by default and acquires special or terminological meaning when this is activated by the pragmatic characteristics of the discourse. [...] Any lexical unit would thus have the potential of being a terminological unit.* (CABRÉ, 2003, p.189-190);
c) os níveis lexical, morfológico, sintático e textual podem veicular conhecimento especializado; ou seja, a Terminologia pode ter como objetivo a descrição linguística em todas esses níveis; no caso de uma *terminet,* o conhecimento a ser descrito engloba o nível léxico;
d) os termos devem ser analisados em seu ambiente natural de ocorrência, ou seja, dos discursos especializados;

e) a variação conceitual e denominativa deve ser considerada;
f) do ponto de vista cognitivo, as unidades terminológicas (i) estão subordinadas a um contexto temático, (ii) ocupam um lugar preciso num mapa conceitual e (iii) o seu significado específico é determinado pelo lugar que ocupam nesse mapa (CABRÉ, 2003); no caso de uma *terminet,* o conhecimento a ser descrito também engloba a dimensão conceitual dos termos;

Tendo em vista especificamente o pressuposto teórico descrito em (d), observamos que, na prática, o emprego da TCT pressupõe o uso dos *corpora.* Em outras palavras, os *corpora* são de suma importância para qualquer projeto que visa à construção de um objeto terminológico, como uma *terminet,* por exemplo. O emprego dos *corpora* permite que os termos sejam descritos e analisados em seu contexto usual ou real (CABRÉ, 1999) e, sobretudo, em larga escala. Em outras palavras, podemos dizer que a análise baseada em *corpus* permite que sejam feitas observações precisas sobre o real comportamento linguístico, proporcionando dados ou informações confiáveis sobre os fatos de uma língua. Berber Sardinha (2004) salienta que até fatos novos podem ser descobertos por meio da análise de *corpus,* os quais não seriam perceptíveis pela intuição.

Em suma, segundo os pressupostos gerais da TCT, os termos (isto é, os signos que ocorrem como unidades lexicais terminológicas) e suas propriedades só podem ser identificados e descritos no seu ambiente natural de ocorrência, ou seja, nos discursos especializados. Dessa forma, esses princípios teóricos e metodológicos põem em evidência a importância do uso dos *corpora* em qualquer trabalho terminológico (NASCIMENTO, 2003; AGBAGO, BARBIÈRE, 2005; CABRÉ ET AL., 2005; ALMEIDA, 2006). Quanto à utilização dos *corpora* como fontes de conhecimento especializado, Nascimento (2003) salienta que a construção de uma base lexical terminológica depende diretamente da qualidade do *corpus*.

Na próxima Seção, descrevemos a metodologia empregada para a construção do *Corpus.*EaD.

4. ETAPAS DE CONSTRUÇÃO DO *CORPUS*

Embora existam *corpora* disponíveis para vários domínios, a construção de uma *terminet* para certos domínios pode requerer a construção de um *corpus*. Seguindo os pressupostos da Linguística de Corpus no geral (KENNEDY, 1998; RENOUF, 1998; BIBER et al, 1998; BERBER SARDINHA, 2004; SINCLAIR, 2005) e de trabalhos que focalizaram a construção de *corpora* especializados (COLETI et al. 2008), o *Corpus.*EaD foi construído com base nas seguintes etapas (DI-FELIPPO, SOUZA, 2010): a) projeção do *corpus*, que consiste na definição do tipo de *corpus* necessário à pesquisa; b) compilação; c) pré-processamento (conversão, limpeza, nomeação, anotação e armazenamento); d) disponibilização.

4.1. Projeção do *Corpus*

Essa etapa constitui na delimitação do tipo de *corpus* necessário à pesquisa em questão.
A projeção foi feita com base em três conjuntos de critérios: (i) critérios provenientes da definição do objeto *corpus*, (ii) critérios provenientes do tipo do recurso lexical a ser

construído e (iii) critérios provenientes das decisões do projeto ao qual a construção do *corpus* se vincula (DI-FELIPPO;SOUZA, 2010).

4.1.1. Critérios provenientes da definição do objeto *corpus*

Na literatura geral da Linguística de Corpus, encontramos critérios que definem, como mencionado, a essência do objeto denominado *corpus*. Esses critérios referem-se a:

a) *Representatividade*

A essência do objeto *corpus* está relacionada à questão da representatividade. A questão da representatividade é bastante polêmica no âmbito da Linguística de Corpus. Um dos motivos para tal polêmica é o fato de a representatividade ter sido sempre atacada pelos gerativistas, pois, para eles, um corpus nunca será grande o suficiente para representar a língua porque é apenas um conjunto de exemplos de realizações linguísticas. Sardinha (2004), por exemplo, partindo do pressuposto estabelecido por Halliday (1991, 1992) de que a língua é um sistema probabilístico, defende que a representatividade pode ser assegurada por meio das respostas a três perguntas. São elas: "Representativo do quê?", "Representativo para quem? e "Representativo para quê?" Para responder às duas primeiras perguntas, destacamos os trabalhos de Nascimento (2003) e Aluísio e Almeida (2007). Esses autores demonstram, sob um viés teórico, que um *corpus* especializado de no mínimo 1 milhão de palavras pode ser considerado representativo de uma área especializada do conhecimento. Para responder à terceira pergunta, baseamo-nos nos trabalhos de Almeida et al. (2006), Kasama (2009) e Pino (2009) que, assim como o projeto TermiNet, realizaram a extração (semiautomática) de conhecimento léxico-conceitual especializado a partir de *corpora* para sistematização vocabular. Em tais trabalhos, os *corpora* utilizados tinham, no mínimo, 1 milhão de palavras, o que corrobora empiricamente as palavras de Nascimento (2003) e Aluísio e Almeida (2006).

b) *Autenticidade*

Um *corpus* deve conter somente textos naturais, ou seja, "aqueles que existem na linguagem e que não foram criados com o propósito de figurarem no corpus" (BERBER SARDINHA, 2004). Além de restringir os textos àqueles que não foram produzidos com o propósito de serem alvo de pesquisa linguística, o termo natural restringe os textos àqueles produzidos somente por humanos. Dessa forma, excluímos textos criados em linguagem artificial, como em linguagens de programação. A noção de autenticidade limita o *corpus* a conter somente textos escritos por falantes nativos. Tal consideração implica no descarte de traduções.

c) *Balanceamento*

Entendemos que os componentes de um *corpus* devem estar distribuídos em quantidades semelhantes ou uniformes; no caso, buscamos construir *um* corpus que tenha certo equilíbrio quanto aos seguintes gêneros discursivos: (i) técnico-científico, que recobre os tipos textuais artigo científico, tese, dissertação, etc.; (ii) científico de divulgação, que engloba especificamente artigo científico de divulgação; (iii) instrucional, que engloba tipos textuais como livro-texto e apostila; e (iv) informativo, que engloba tipos textuais

como reportagem e notícia. A recolha de textos desses gêneros busca também satisfazer outro requisito, a diversidade.

d) *Diversidade*

A opção pela heterogeneidade de gênero, isto é, a recolha de textos com diferentes graus de complexidade quanto ao tema, deve-se à tentativa de capturar diferentes "níveis" de informação. A tentativa de capturar diferentes níveis de informação, por sua vez, pauta-se no fato de que textos especializados, como os artigos científicos, não expressam certo conhecimento básico sobre o domínio (por exemplo, definições e exemplificações de termos), pois pressupõem que seus leitores são especialistas que compartilham tal conhecimento básico. Essas informações mais básicas, por sua vez, são comumente explicitadas em textos informativos ou didáticos e de divulgação. Dessa forma, tendo em vista o objetivo de extrair conhecimento léxico-conceitual para a construção de uma terminet, um corpus que apresenta heterogeneidade de gênero pode garantir que os diferentes níveis de conhecimento sejam identificados nos textos.

4.1.2. Critérios provenientes do tipo de recurso a ser construído a partir do *corpus*

A WN.Pr, assim como as *wordnets* das demais línguas, tem sido amplamente utilizada no processamento automático da língua escrita. Mais especificamente, em aplicações como: (i) recuperação de informação (do inglês, *information retrieval*) e (ii) sumarização automática (do inglês, *automatic summarization*). Consequentemente, o *corpus* necessário à construção de uma *terminet* deve englobar apenas textos registrados nesse meio, sejam eles digitais ou impressos. Quanto à modalidade, os textos registrados em meio escrito podem ser, em princípio, tanto de língua falada quanto de língua escrita. No entanto, diante da dificuldade de aquisição de material transcrito, o *corpus* deve ser composto preferencialmente por textos da modalidade escrita.

Se, por um lado, as *terminets* compartilham com as *wordnets* a característica de serem *lingwares* destinados ao processamento de língua natural em meio escrita, por outro lado, as *terminets* e as *wordnets* diferem quanto à cobertura da língua que se pretende processar. As *terminets,* como o próprio nome indica, são recursos lexicais que auxiliam o processamento automático de textos técnicos ou especializados (ou seja, textos de determinados domínios do conhecimento), enquanto que as *wordnets* são recursos para o processamento de textos de língua geral. Consequentemente, uma *terminet* deve sistematizar o conhecimento léxico-conceitual de determinado domínio. Dessa forma, o *corpus* a partir do qual esse conhecimento léxico-conceitual será adquirido deve ser de conteúdo especializado ou terminológico e não de língua geral.

Além disso, as bases *wordnets* são originalmente recursos monolíngues, com exceção das bases resultantes do alinhamento de duas ou mais *wordnets.* As *terminets* também são consideradas, em princípio, bases monolíngues e, portanto, o *corpus* também deve ser monolíngue.

Dessa forma, vê-se que, o tipo de recurso que se quer construir a partir de um *corpus* é responsável por delinear certas características desse *corpus.* Com base da nomenclatura empregada por Giouli e Peperidis (2002) para caracterizar a tipologia de

um *corpus*, pode-se dizer que uma base terminológica do tipo *wordnet* delineia as seguintes características: (i) a variação histórica (sincrônico/ contemporâneo); (ii) o tipo de texto (registrado em meio escrito); (iii) a modalidade da língua (escrita); (iv) a cobertura da língua (especializado); e (v) a quantidade de línguas (monolíngue).

4.1.3. Critérios provenientes de decisões de projeto

Além dos critérios provenientes da própria definição de *corpus* e das características do recurso a ser construído a partir dele, algumas decisões de projeto também interferem na tipologia do *corpus.*

No caso do TermiNet, essas decisões foram: (i) utilizar métodos automáticos para extrair os dados do *corpus* (ou seja, por meio de programas computacionais) e (ii) disponibilizar para o *corpus* para as comunidades da Linguística e do PLN. Consequentemente, o *corpus* precisa ser: (i) anotado morfossintaticamente (ou seja, as ocorrências (palavras) estejam associadas a etiquetas que indicam categorias sintáticas; p.ex.: *casa*_N(ome)) e (ii) disponível na *Web*.

Ao final da projeção, chegamos às seguintes características para um *corpus* que servirá de base para uma *terminet*. Tais características foram reunidas na tipologia apresentada no Quadro 1, a qual foi tomada como guia para a compilação dos textos que compõem o *Corpus*.EaD.

Tamanho (Representatividade e amostragem)	**Médio-grande (ao menos, 1 milhão)**
Balanceamento	Por gênero
Modalidade	Escrito (vs. *corpus* de áudio)
Tipo Textual	Escrito (vs. *corpus* com transcrições)
Meio	Jornais, livros, revistas, manuais e outros
Cobertura da língua	*Corpus* especializado
Gêneros	Técnico-científico, Científico de Divulgação, Informativo e Instrucional
Quantidades de línguas	Monolíngue
Anotação	Anotado (em nível morfossintático)
Comunidade Produtora	Falantes nativos
Mutabilidade	Aberto
Variações Históricas	Sincrônico
Disponibilidade	Disponível na Web

Tabela 1: Tipologia do *Corpus*.EaD

Na próxima Seção, discutimos como foi feita a tarefa de compilação dos textos disponível na *web* para a efetiva construção do *corpus* do domínio da EaD. Destacamos que a *web* é inegavelmente uma fonte inestimável de textos de diferentes gêneros e tipos, os quais podem ser livremente acessados e "adquiridos".

4.2. Compilação do *corpus*

Essa etapa constituiu necessariamente na coleta dos textos que compõem o *corpus*. A compilação foi feita em etapas, as quais estão descritas na sequência.

4.2.1. Identificação das fontes e coleta dos textos

Segundo Aluísio e Almeida (2006), há duas estratégias possíveis para a tarefa de coleta dos textos disponível na *web*, as quais podem ser assim sistematizadas:

1. busca na *web* com máquinas de busca:
   a. uso de uma máquina de busca como o *Google* para pesquisar toda a *web* (podemos utilizar palavras-chave escolhidas para a pesquisa em foco, sobretudo, no caso de pesquisas terminológicas);
   b. uso de ferramentas que pré-processam e/ou pós-processam os resultados das buscas de tais máquinas como fazem o *WebCorp*[2] e *KWiCFinder*[3];
2. coleta de páginas da *web*, organizando-as em um computador local:
   a. construção automática de *corpus* com ajuda de *offline browsers* como o *HTTrack*[4] ou com a ajuda de ferramentas de apoio para a compilação de *corpora* descartáveis (do inglês, *disposable corpora*) como o *Corpógrafo*[5] e o *Toolkit BootCat*[6]; o *HTTrack,* em especial, busca uma URL[7] fornecida pelo pesquisador e salva todo o conteúdo do *site* correspondente (fotos, imagens, textos, links, etc.) em uma máquina local;
   b. coleta do *corpus* pela seleção de páginas de forma manual ou semi-automática de acordo com um projeto específico de *corpus*.

Mesmo havendo inúmeras ferramentas computacionais que auxiliam a coleta em massa de textos na *web* como as citadas, a estratégia mais simples de seleção das fontes e de coleta dos textos, caracterizada pelo acesso às páginas desejadas e *download* dos arquivos no computador, tem-se mostrado mais eficaz, haja vista que é essa estratégia tem sido na construção de grandes *corpora*, como o *British National Corpus* (BNC)[8].

Diante disso, optamos pela estratégia manual de compilação. Para tanto, o motor de busca que utilizamos foi o *Google* que, a partir das palavras-chave "educação a distância" e "ead", retornou várias páginas vinculadas ou indexadas a essas palavras. Contudo, nem todas as páginas foram consideradas fontes para a compilação dos textos.

---

[2] http://www.webcorp.org.uk/
[3] http://miniappolis.com/KWiCFinder/KWiCFinderHome.html
[4] http://www.httrack.com/page/2/
[5] http://poloclup.linguateca.pt/corpografo/
[6] http://sslmit.unibo.it/~baroni/bootcat.html
[7] *Uniform Resource Locator*; indica o endereço único de um recurso em uma rede (p.ex.: a Internet).
[8] http://www.natcorp.ox.ac.uk/

4.2.2. Identificação (e seleção) das páginas e seleção dos textos

Para a escolha de uma página da *web* como fonte[9] para compilação dos textos, esta deveria cumprir um dos dois requisitos: (i) vínculo à instituições de ensino na modalidade EaD, principalmente públicas e reconhecidamente vinculadas a projetos estaduais e/ou nacionais de EaD e/ou (iii) cujos textos tivessem características previstas na tipologia do *corpus*. As páginas que passaram pelos critérios de seleção estavam prontas para serem utilizadas como fontes de aquisição dos textos. Em cada página, a seleção dos textos relevantes para a construção do *corpus* foi feita por meio das mesmas palavras-chave utilizadas para a identificação das páginas ou *sites.*

Após a compilação, o *corpus* precisou ser preparado para que pudesse receber um tratamento ou processamento computacional nas etapas futuras do projeto TermiNet.

4.3. Pré-processamento dos textos

A preparação ou pré-processamento englobou os processos de: (i) conversão manual e/ou automática dos textos nos formatos doc, pdf e html para o formato txt, (ii) limpeza manual dos dados corrompidos pela conversão; (iii) nomeação padronizada dos arquivos, anotação estrutural dos textos e geração de cabeçalho e armazenamento do *corpus*. Os processos descritos em (iii) são comumente realizados por uma ferramenta computacional denominada "editor de cabeçalho".

4.3.1. Conversão

Nessa tarefa, é preciso converter os textos e/ou documentos originais para um formato legível pela máquina. Essa é uma parte essencial da construção de um *corpus* que será processado por qualquer ferramenta de PLN. As conversões dos textos partiram dos formatos html, pdf, doc e docx, os quais caracterizaram os textos compilados da *web*.

Para converter os textos, testamos alguns conversores automáticos livremente disponíveis na *web*. O primeiro a ser testado foi o *PDF Text Reader 1.1*. Apesar de apresentar várias vantagens, dentre elas a de converter qualquer tipo de arquivo, pdf, doc, docx, etc., observamos que o conversor gerava muito dado corrompido e/ou inseria dados cuja limpeza seria demorada. O segundo extrator testado foi o *Very PDF2Word v3.0*. Comparado com o *PDF Text Reader 1.1,* o *Very PDF2Word v3.0* apresenta vantagens. Na configuração do programa, há uma opção para remover, já no processo de conversão, alguns dados que não são processados pela máquina, como tabelas, fórmulas matemáticas, imagens, etc.. Essa funcionalidade facilita o processo de limpeza dos textos, tarefa descrita da próxima subseção. Contudo, a versão que obtivemos convertia somente as 5 primeiras páginas de um arquivo, o que inviabilizava a usa utilização em nosso projeto. Por fim, testamos o *Adobe Reader 9.0*[10]. Esse programa possui uma opção que permite salvar uma cópia do arquivo pdf diretamente no formato txt. Além disso, apresenta as seguintes vantagens: (a) exclusão de figuras e imagens do corpo de texto durante o processo de conversão e (b) transformação de tabelas em listas de palavras, facilitando a visualização no momento de limpeza dos textos.

---

[9] As fontes utilizadas no projeto TermiNet estão descritas no Apêndice 1.

[10] Disponível em: http://www.baixaki.com.br/site/dwnld907.htm. Acessado em 16 ago 2010.

4.3.2. Limpeza

Na sequência, realizamos a limpeza manual dos arquivos convertidos. Essa tarefa prepara os textos, agora em formato txt, para serem processados pela máquina. No caso, a limpeza garante que o arquivo contenha apenas os dados que podem ser processados pelas ferramentas de PLN, como etiquetadores, *parsers*, etc.

Especificamente, a limpeza consistiu na exclusão de: (i) dados corrompidos pela conversão, como equações e fórmulas matemáticas, figuras e quadros, que não são suportados pelo formato txt; (ii) dados referentes à paginação, cabeçalho, rodapé, autoria, filiação e referência bibliográfica; e (iii) quebras de linha inexistentes no original, resultantes da conversão para txt. As legendas das tabelas, figuras e quadros, por sua vez, foram mantidas nos arquivos em formato txt, pois podem conter candidatos a termo.

Vale ressaltar que as versões originais de todos os textos também foram armazenadas, caso seja necessário recuperar dados que foram corrompidos e/ou excluídos na limpeza. Todos esses dados foram retirados manualmente, contudo, como citado acima, alguns desses dados puderam já ser retirados durante o processo de conversão pelo programa *Adobe Reader 9.0*, facilitando essa tarefa.

4.3.3. Nomeação e Anotação Estrutural

Uma vez limpos, os arquivos txt referentes aos textos que compõem o *corpus* foram submetidos aos processos de (i) nomeação padronizada, que tem o objetivo de facilitar a recuperação posterior de cada texto, e (ii) anotação estrutural de dados externos, como informações de autoria, tipologia textual e de gênero, tipo de fonte, etc. Tais tarefas quais são comumente feitas por meio de uma ferramenta computacional denominada "editor de cabeçalho".

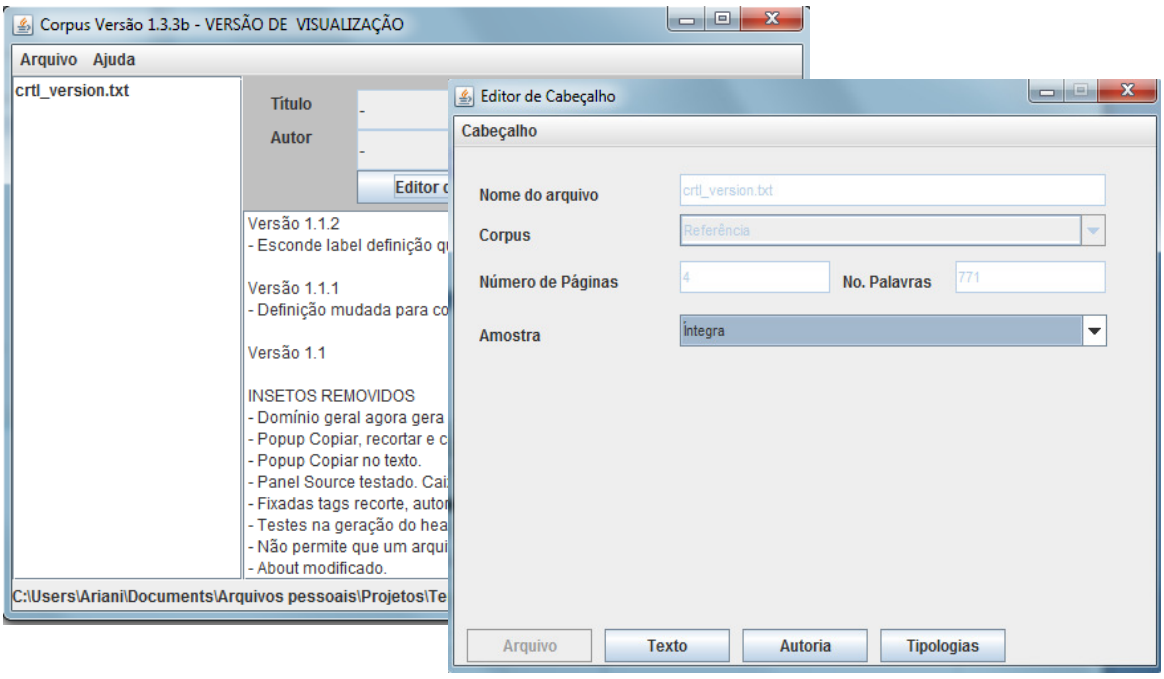

Figura 1: Interface da versão *offline* do Editor de Cabeçalho.

Inicialmente, pensamos em usar a versão *offline* do Editor de Cabeçalho 1.0[11] desenvolvido pelo NILC (Figura 1). Ao testar a ferramenta, no entanto, identificamos que a mesma apresenta certas funcionalidades que não eram de interesse direto do projeto, por exemplo, a geração de cabeçalhos sempre acoplados aos textos, ou seja, a ferramenta não gera os cabeçalhos em arquivos separados, o que não seria viável para o processamento a que o *corpus* será submetido no projeto TermiNet.

Assim, os processos de nomeação dos arquivos e de anotação estrutural foram feitos por meio do editor de cabeçalho (*Header Editor*) disponível no Portal de Córpus[12] do Projeto PLN-Br[13], que apresenta a funcionalidade de gerar os cabeçalhos em arquivos separados (em inglês, *stand off annotation*). O *Header Editor,* em especial, permite que os dados externos sobre cada texto sejam especificados. Com base nessa especificação, a ferramenta nomeia o arquivo de forma padronizada e gera um cabeçalho corresponde no formato XCES (do inglês, *Corpus Encoding Standard for XML*), o qual sistematiza os dados por meio de um conjunto de etiquetas. Na Figura 2, apresentamos a interface gráfica do referido editor.

Figura 2: Interface do *Header editor do* Portal de Corpus.

Os cabeçalhos, como o ilustrado na Figura 3, têm o objetivo de propiciar a geração de *subcorpora.* Quando um *corpus,* cujos textos possuem cabeçalhos que sistematizam dados estruturais externos, é processado por um editor que reconhece as etiquetas do cabeçalho, pode-se gerar *subcorpora* em função das informações contidas no cabeçalho, p.ex.: *subcorpora* de todos os textos publicados por um autor específico ou de um gênero específico. Ademais, no caso do *Header Editor,* parte das informações

[11] Disponível em: http://www.nilc.icmc.usp.br/lacioweb/downloads.htm. Acessado em: 16 ago. 2010.
[12] http://www.nilc.icmc.usp.br:8180/portal/
[13] http://www.nilc.icmc.usp.br/plnbr/

sistematizadas no cabeçalho é utilizada na elaboração do padrão de nomeação. No caso, a nomeação dos arquivos seguiu o padrão: gênero textual, tipo textual e número de registro do arquivo, p.ex.: Instrucional_Livro-texto_73.txt.

```
<?xml version="1.0" encoding="UTF-8"?>
<cesHeader xmlns="http://www.xces.org/schema/2003"
xmlns:xlink="http://www.w3.org/1999/xlink"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.xces.org/schema/2003" version="1.0.4">
      <fileDesc>
            <titleStmt>
                  <title>A Educação a Distância possibilitando a
formação do professor com base no ciclo da prática pedagógica</title>
            </titleStmt>
            <editionStmt version="01" />
            <extent>
                  <wordCount>8085</wordCount>
                  <byteCount units="bytes">108744.0</byteCount>
                  <extNote>14</extNote>
            </extent>
            <sourceDesc>
                  <biblStruct>
                        <monogr>
                              <title>A Educação a Distância
possibilitando a formação do professor com base no ciclo da prática
pedagógica</title>
                              <author>Maria E. B. Brito Prado</author>
                              <edition></edition>
                              <imprint>
```

Figura 3: Parte de um cabeçalho gerado pelo *Header Editor.*

Após o processo de especificação dos dados externos, os textos submetidos ao editor *Header Editor* foram armazenados na base de dados do Portal de Córpus e estão disponíveis, aos membros do projeto, para consulta e *download*, com ou sem cabeçalhos.

4.3.4. Armazenamento

Além do armazenamento no Portal de Corpus, a versão sem cabeçalho do *Corpus.*EaD também foi armazenada em uma máquina local, que funciona como o servidor do projeto TermiNet. Os textos foram armazenados em uma estrutura de pastas do *Windows Explorer* que tem como critério de organização o nível de especialização dos gêneros/tipos textuais (Figura 4).

Com base na Figura 4, vê-se especificamente que os textos do *Corpus.*EaD foram armazenados em dois *subcorpora*. O *subcorpus* +Técnico engloba apenas os textos do gênero técnico-científico, os quais podem ser do tipo tese, dissertação, projeto de pesquisa ou artigo científico. Tais textos são normalmente escritos para a comunidade de especialistas. O *subcorpus* –Técnico engloba os textos dos gêneros científico de divulgação, instrucional e informativo, os quais podem ser do tipo artigo de divulgação, livro-texto, apostila, reportagem ou notícia. Esses tipos textuais caracterizam textos que são comumente escritos para principiantes na área-objeto ou para leigos.

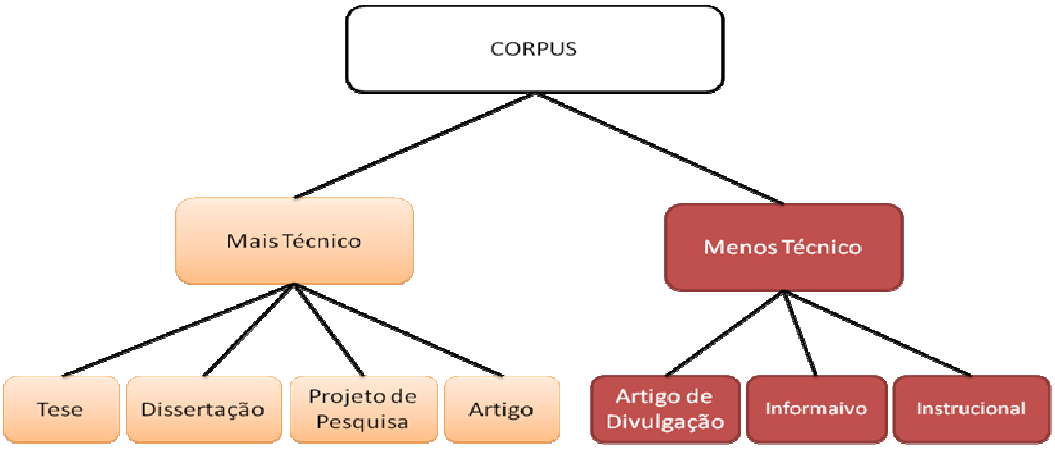

Figura 4: Organização do *Corpus.*EaD.

A organização e o armazenamento dos textos segundo o grau de especificidade têm duas justificativas: (i) busca pelo balanceamento dos gêneros e (ii) processamento individual dos *subcorpora* +Técnico e –Ténico nas etapas futuras do projeto TermiNet.

Quanto à motivação (i), ressalta-se que o balanceamento por gênero previsto na tipologia (Tabela 1) buscava garantir o equilíbrio entre os textos com diferentes graus ou níveis de especificidade, pois o conhecimento a ser extraído para a construção de uma *terminet* é explícito de forma diferente em textos mais e menos especializados. Para satisfazer ao balanceamento, os textos do gênero técnico-científico foram agrupados no *subcorpus* +Técnico e os dos gêneros científico de divulgação, instrucional e informativo foram agrupados no *subcorpus* –Ténico. Mesmo assim, o *Corpus.*EaD não está efetivamente balanceado quanto aos "supergêneros" representados pelos *subcorpora* (Gráficos 1 e 2). No caso, a falta de equilíbrio reflete uma característica do próprio domínio: maior volume de publicações do gênero técnico-científico que dos demais gêneros. Em outras palavras, essa discrepância entre o número de ocorrências ("palavras") por gênero reflete, na verdade, o estágio atual das produções dos textos de áreas em consolidação, como a EaD, e não as falhas na seleção dos textos. Conclusão semelhante fizeram Coleti et al. (2008) na compilação de um *corpus* da área de Nanociência e Nanotecnologia.

Quanto à motivação (ii), ressalta-se que se objetiva analisar diferenças e/ou semelhanças no número final de termos, sinônimos e hipônimos extraídos de cada *subcorpora*.

4.4. Disponibilização

Seguindo os pressupostos da Linguistica de Corpus, a disponibilização ampla requer o pedido e concessão de permissão de uso. Como isso é uma tarefa muito custosa optamos por dois tipos de disponibilização do *corpus.*EaD. Nosso intuito é que, ao longo do projeto TermiNet, possamos criar um site, onde somente usuários cadastrados, com fins de pesquisa, solicite o envio de senha e possa utilizar total ou parcialmente o *Corpus.*EaD. Enquanto esse site não fica pronto, mais uma vez o Portal *Corpus* se faz importante: todo o nosso *corpus*, convertido, limpo, nomeado e anotado estruturalmente,

está hospedado na base do portal. Porém, a disponibilidade do *Corpus*.EaD se faz somente para os pesquisadores do projeto TermiNet, ou, se para fins de pesquisa, outro pesquisador desvinculado ao projeto solicite permissão de acesso ao *corpus.* Caso contrário, a visualização e disponibilização do *corpus* ficam restrita aos pesquisadores do projeto referido.

Na próxima seção está disponível, quantitativamente, toda a descrição do *Corpus*.EaD, desde seus números de ocorrências, até as quantidades de textos recolhidos separados por gêneros que consideramos ao longo da pesquisa.

5. DESCRIÇÃO QUANTITATIVA DO *CORPUS*.EaD

Ao final, o *Corpus*.EaD (versão 1) apresenta o total de 1.350.683 ocorrências, sendo 534.147 do *subcorpus* –Ténico e 816.536 do *subcorpus* +Ténico[14] (Gráfico 1).

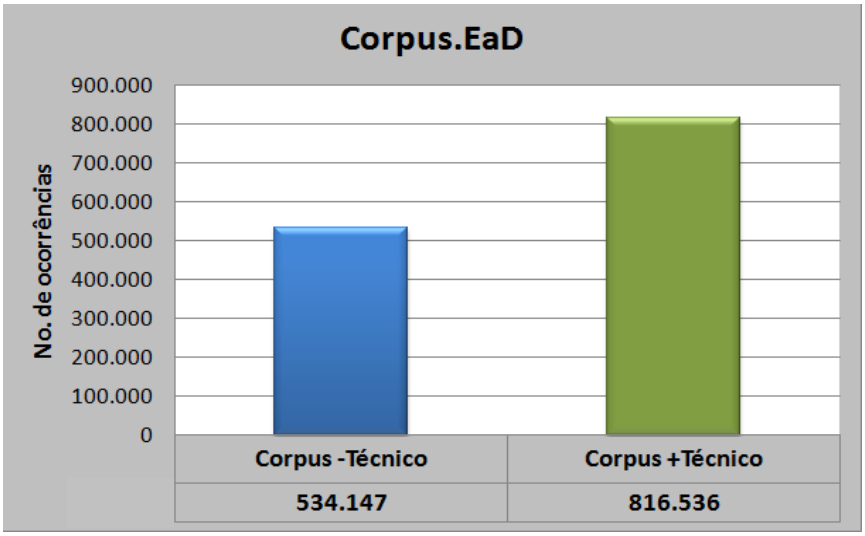

Gráfico 1: Número de ocorrências em função dos *subcorpora*

Na Tabela 1, apresentamos a distribuição dos textos em função dos gêneros/tipos textuais.

Subcorpora	**Gêneros Textuais**	**Tipos Textuais**	**Quantidade**	**Total**
-Técnico	Científico de divulgação	Artigos de Divulgação	138	307
	Instrucional	Livro-Texto	9	
		Apostila	2	
	Informativo	Notícias/Reportagens	158	
+Técnico	Técnico-científico	Tese	3	40
		Dissertação	14	
		Projetos de Pesquisa	1	
		Artigos Científicos	22	

Tabela 2: *Corpus.*EaD: número de textos por gênero.

[14]Toda a contabilização foi feita pelo programa Word Smith Tools v2. 0. Disponível em: http://software.informer.com/getfree-wordsmith-tools-5.0-reviews/. Acessado em 16 ago 2010.

Em termos de porcentagem, a distribuição dos textos pode ser vista na Figura 2.

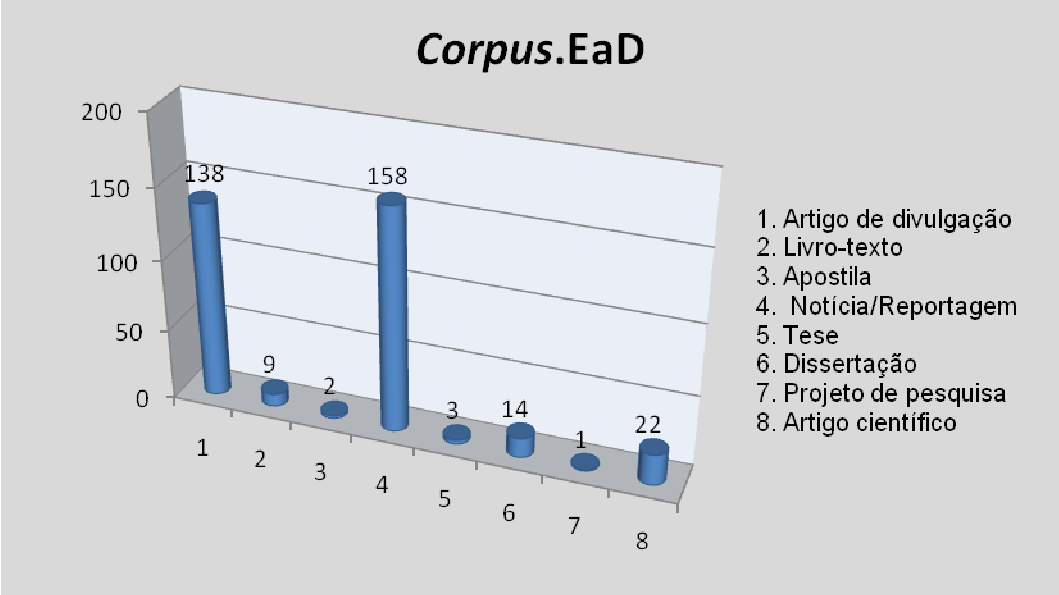

Figura 2: *Corpus.*EaD: porcentagem de textos por gênero.

Na próxima seção, apresentamos as principais contribuições do projeto de IC ora relato para as áreas do PLN e Terminologia.

6. RESULTADOS E DISCUSSÕES

Diante de toda a pesquisa, podemos sistematizar os resultados do nosso projeto em:

a) Aquisição de um arcabouço teórico-metodológico para a construção de *corpora* com vistas à construção de *wordnets* terminológicas: a metodologia utilizada na construção do *Corpus.*EaD poderá ser aplicada a outros projetos que buscam construir *wordnet* terminológicas.

b) Especificação das principais características ou requisitos que um *corpus* precisa satisfazer para subsidiar a construção de *wordnets* terminológicas ou *terminets*, ou seja, especificação de sua tipologia.

c) Construção do *Corpus.*EaD, o primeiro do domínio da EaD em PB, o qual poderá servir de fonte para outros projetos terminológicos, computacionais ou não.

d) Disponibilização do *Corpus.*EaD na *web* como fonte para pesquisas futuras: a disponibilidade atual do *Corpus.*EaD através do Portal de Córpus é restrita aos membros do projeto TermiNet. No entanto, como os *corpora* especializados são recursos extremamente úteis e de construção cara, pretende-se disponibilizá-lo para as comunidades linguística e do PLN via *web*. Tal disponibilização será feita por meio do futuro *site* do projeto TermiNet. Especificamente, o *Corpus.*EaD estará disponível para *download* mediante o cadastramento *online* do pesquisador/solicitante e subsequente envio de senha por parte do pesquisador responsável pelo TermiNet.

7. CONSIDERAÇÕES FINAIS

Tendo em vista os objetivos iniciais do projeto, que foram projetar e construir um *corpus* do domínio da EaD em PB, enfatizamos que todos eles foram alcançados. Aliás, do ponto de vista teórico, ressaltamos que o trabalho de IC ora relatado produziu como principal resultado não a tipologia do *corpus,* mas sim a discussão sobre os critérios que culminaram nessa tipologia. De um modo geral, acreditamos que essa discussão pode auxiliar trabalhos futuros na tarefa de projeção de *corpora*, etapa fundamental na construção desse tipo de recurso. Sob o ponto de vista prático, destacamos o próprio *Corpus*.EaD, que é o único que se tem conhecimento em PB dessa área-objeto.

AGRADECIMENTOS

Agradecemos à Coordenadoria de Iniciação Científica e Tecnológica da Pró-Reitoria de Pesquisa da UFSCar pela bolsa concedida no âmbito do Programa Institucional de Bolsas de Iniciação Científica – PIBIC/CNPq/UFSCar. Agradecemos também aos pesquisadores dos grupos de pesquisa NILC (Núcleo Interinstitucional de Linguística Computacional – USP, UFSCar, UNESP) e GETerm (Grupo de Estudos e Pesquisa em Terminologia - UFSCar) pela participação ativa no projeto.

REFERÊNCIAS IBLIOGRÁFICAS

AGBAGO, A., BARRIÈRE, C. *Corpus* construction for Terminology. In: *CORPUS* LINGUISTICS CONFERENCE, 2005. **Proceedings**... Birmingham, 2005. p. 14-17.

ALMEIDA, G.M.B. A Teoria Comunicativa da Terminologia e a sua prática. **Alfa**, v. 50, p. 81-97, 2006.

ALMEIDA, G. M. B et al. Recolha e sistematização de corpora para elaboração do primeiro dicionário-piloto em Nanociência e Nanotecnologia em Língua Portuguesa. In: JORNADA DA REDE PANLATINA DE TERMINOLOGIA (REALITER´06), 2006, Rio de Janeiro. **Anais...**Paris (França), 2006.

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 lingüística. **Calidoscópio**, v. 4 (3), p. 155-177, set/dez 2006.

BENTIVOGLI, L.; BOCCO, A.; PIANTA, E. ArchiWordnet: integrating Wordnet with domain-specific knowledge. In: INTERNATIONAL GLOBAL WORDNET CONFERENCE, 2, 2004. **Proceedings**... Masaryk University, Brno, 2004. P. 39-47. Disponível em: <www.fi.muni.cz/gwc2004/proc/101.pdf>. Acesso em: 10 maio 2008.

BERBER SARDINHA, T. **Linguística de Corpus**. Manole. Barueri, SP. 2004.

BIBER, D.; CONRAD, S.; REPPEN, R. **Corpus linguistics***:* Investigating language structure and use. Cambridge: Cambridge University Press, 1998.

CABRÉ, M. T. **La terminología: representación y comunicación: elementos para una teoría de base comunicativa y otros artículos.** Barcelona: Institut Universitari de Linguística Aplicada, 1999.

___. Theories of terminology: their description, prescription and explanation. **Terminology**, v.9(2), p.163-200, 2003.

___. Application-driven terminology engineering. **Terminology**, v.11(2), p. 1-19, 2005.

COLETI, J. S.; et al. A compilação de corpus em língua portuguesa na área de nanociência/nanotecnologia: problemas e soluções. In: TAGNIN, S. E. O.; VALE, O. A. (Org.). **Avanços da Lingüística de Corpus no Brasil**. 1 ed. São Paulo: Humanitas, 2008, p. 167-191.

DIAS-DA-SILVA, B.C. O estudo linguístico-computacional da linguagem. **Letras de Hoje,** Porto Alegre, v. 41 (2), p. 103-138, 2006.

___.; DI FELIPPO, A.; NUNES, M.G.V. The automatic mapping of Princeton Wordnet lexical-conceptual relations onto the Brazilian Portuguese Wordnet database. In: International Conference on Language Resources and Evaluation, 6, 2008. **Proceedings**... Marrakech, Morocco, 2008.

DI-FELIPPO, A.; ALMEIDA, G. M. B. Uma metodologia para o desenvolvimento de *wordnets* terminológicas em português do Brasil. **TradTerm***,* n.16, 2010. ISSN 0104-639X *In press*

DI-FELIPPO, A.; DIAS-DA-SILVA, B. C. As abordagens teóricas e os formalismos para o tratamento computacional do significado lexical. **Revista Brasileira de Linguística Aplicada**, Belo Horizonte, v.10, n.01, p. 43-69, 2010. ISSN 1676-0786

DI-FELIPPO, A.; SOUZA, J. W. C. O projeto do *corpus* para a construção de uma *wordnet* terminológica. In: Shepherd, T., Berber Sardinha, T. e Veirano Pinto, M. (Orgs). ENCONTRO DE LINGUÍSTICA DE *CORPUS*, 8, 2009. **Anais**... Rio de Janeiro (RJ), 2010. *In press*.

FELLBAUM, C (Ed.). **Wordnet**: an electronic lexical database. Ca, MA: MIT Press, 1998.

GIOULI, V.; PIPERIDIS, S. **Corpora and HLT: current trends in corpus processing and annotation**. Bulgaria: Institute for Language and Speech Processing, 2002. Disponível em: < http://www.larflast.bas.bg/balric/index/index_eng.htm>. Acesso em: 01 junho 2008.

HALLIDAY, M. A. K. Corpus studies and probabilistic grammar. In: Aijmer, K.; Alternberg, B. (Eds). **English corpus linguistic**. London: Longman, 1991, p. 30-43.

KASAMA, D. Y. **Estruturação do conhecimento e relações semânticas: uma ontologia para o domínio da Nanociência e Nanotecnologia**. São José do Rio Preto, 2009. 178p. Dissertação (Mestrado em Linguística) – Faculdade de Biociências, Letras e Ciências Exatas, Universidade Estadual Paulista, São José do Rio Preto, 2009.

KENNEDY, G. **An introduction to corpus linguistics**. London: Longman, 1998.

LEECH, G. **English in advertising: a linguistic study of advertising in Great Britain.** London: Longman, 1966.

MITKOV, R. (Ed.). **The Oxford handbook of computational linguistics**. Oxford, New York: Oxford University Express, 2004.

NASCIMENTO, M. F. B. O papel dos corpora especializados na criação de bases terminológicas. In: Castro, I.; Duarte, I. (Orgs.). **Razões e emoções, miscelânea de estudos em homenagem a Maria Helena Mateu***s*. Lisboa: Imprensa Nacional-Casa da Moeda, v. II, p. 167-179, 2003.

PINO, D. H. P. **Terminologia do biodiesel: aspectos semânticos e morfológicos**. São Carlos, 2009. 98p. Dissertação (Mestrado em Linguística) – Departamento de Letras, Universidade Federal de São Carlos, São Carlos, 2009.

POPRAT, M. et al. Building a BioWordnet using Wordnet data structures and Wordnet's software infrastructure – a failure story. In: ACL WORKSHOP ON SOFTWARE

ENGINEERING, TESTING, AND QUALITY ASSURANCE FOR NATURAL LANGUAGE, 2008. **Processing**… Ohio, 2008. P.31-39.

RENOUF, A. (Ed.) **Explorations in Corpus Linguistics**. Amsterdam: Rodopi, 1998.

RIGAU, G. **Automatic acquisition of lexical knowledge from MRDs**. Tesis doctoral, Departament de Llenguatges i Sistemes Informàtics, UPC, Barcelona, 1998.

SAGRI et al. Jur-Wordnet. In: INTERNATIONAL GLOBAL WORDNET CONFERENCE, 2, 2004. **Proceedings**... Masaryk University, Brno, 2004. P. 305-310.

SINCLAIR, J. Corpus and text: basic principles. In: Wynne, M. (Ed.). **Developing linguistic corpora: a guide to good practice**. Oxford: Oxbow Books, 2005. p.1-16. Disponível em: <http://ahds.ac.uk/linguistic-corpora/>. Acesso em: 30 out. 2006.

SMITH, B.; FELLBAUM, C. *Medical Wordnet*: a new methodology for the construction and validation of information resources for consumer health. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS,20,2004. **Proceedings**… Geneva, 2004.

SUÁREZ, M.; CABRÉ, M.T. La variación denominativa en los textos de especialidad: indicios linguísticos para su recuperación automática. In: Simposio Iberoamericano de Terminología, 8, 2002. **Actas**… Cartagena de Indias, 2002. p.1-12.

APÊNDICE

Fontes utilizadas na compilação do *Corpus*.EaD.

Gênero textual	Tipo textual	Fontes
Técnico-científico	Artigo Tese Dissertação Projeto de pesquisa	Site da Associação Brasileira de Educação a Distância (http://www2.abed.org.br/) Site Vivência Pedagógica (http://www.vivenciapedagogica.com.br/textos_educacao_a_distancia.html) Banco de teses (Capes) (http://servicos.capes.gov.br/capesdw/) Biblioteca digital Domínio Público (http://www.dominiopublico.gov.br/) Biblioteca Virtual da UNICAMP (http://libdigi.unicamp.br/) Portal Scielo (http://search.scielo.org/?q=ensino%20a%20dist%E2ncia&where=ORG) Instituto Benjamin Constant (http://www.ibc.gov.br/Nucleus/index.php?catid=147&blogid=1&itemid=10170) Revista Digital de Tecnologia Educacional e Educação a Distância http://www.pucsp.br/tead/n1a/artigos.htm
Científico de divulgação	Artigo de divulgação	Site Edutecnet (http://www.edutecnet.com/) Conect@a - Revista on-line de Educação a Distância (http://www.revistaconecta.com/) Revista Pesquisa FAPESP (on-line) http://www.revistapesquisa.fapesp.br/?search=educa%E7%E3o+a+dist%E2ncia
Instrucional	Livro-texto Apostila	Núcleo de Informática Aplicada à Educação (Nied/UNICAMP) (http://www.nied.unicamp.br/oea/pub/livro3/) Livros on-line (http://www.nied.unicamp.br/oea/pub.html) Faculdade de Educação/UnB (http://www.fe.unb.br/acontece/noticias/lancamento-de-livro) Universidade Federal da Bahia (Educação Online) (http://www.moodle.ufba.br/mod/resource/view.php?inpopup=true&id=48331) Portal do Ministério da Educação (http://portal.mec.gov.br/seesp/arquivos/pdf/aee_ead.pdf) Associação Brasileira de Pesquisadores em Cibercultura (Abciber) (http://www.abciber.org/publicacoes/livro1/)
Informativo	Reportagem Notícia	Site da Associação Brasileira de Educação a Distância (http://www2.abed.org.br/) Site Edutecnet (http://www.edutecnet.com/) Associação Brasileira dos Estudantes de Educação a Distância http://www.estudantesead.org.br//mod/forum/view.php?id= Anuário Brasileiro de Ensino a Distância (http://www.abraead.com.br/) Anuário Brasileiro Estatístico de Educação Aberta e a Distância (http://www.abraead.com.br/anuario/anuario_2008.pdf) Site da Folha de São Paulo (http://www.folha.uol.com.br/) Site do G1(http://g1.globo.com/) Site da Revista IstoÉ (http://www.istoe.com.br/revista/edicoes-anteriores/) Site do Jornal Hoje (http://g1.globo.com/jornal-hoje/) Site do Jornal Nacional (http://g1.globo.com/jornal-nacional/) Site da E-Learning Brasil (http://www.elearningbrasil.com.br/) Portal E-Learnig (http://www.portalelearning.com.br/) Site da Revista Exame (http://portalexame.abril.com.br/pme/) Site da Universidade Aberta do Brasil (http://uab.pti.org.br/) Site da Revista Veja (http://veja.abril.com.br/)

Tabela 4: Fontes consultadas para a construção do *Corpus*.EaD