O *Corpus*.EaD no Projeto TermiNet: Estratégias de Construção

Ariani Di Felippo
Jackson W. da Cruz Souza

Departamento de Letras – Universidade Federal de São Carlos (UFSCar)
Núcleo Interinstitucional de Linguística Computacional (NILC)
Grupo de Estudos e Pesquisas em Terminologia (GETerm)

Financiadores

Desenvolvedores

Introdução e Objetivos

- A sistematização do **conhecimento especializado** no **formato *wordnet*** é feita com base em **recursos estruturados**.
- No **projeto TermiNet**, propôs-se uma metodologia de construção de *wordntes* terminológicas com base em ***corpus***.
- A validação dessa metodologia está sendo feita por meio da construção da **WordNet.EaD** português do Brasil.
- Para tanto, construiu-se o ***Corpus*.EaD** com base nas seguintes etapas: (a) projeção do *corpus*; (b) compilação dos textos; (c) pré-processamento (conversão, limpeza, nomeação e anotação) dos textos e (d) disponibilização do *corpus.*
- Neste trabalho, objetiva-se apresentar as **estratégias** utilizadas nas etapas (a)-(d) e as principais características do *corpus.*

Projeção do *corpus*

Tamanho	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
Quantidade de línguas	Monolíngue
Anotação	Anotado (em nível morfossintático)
Comunidade produtora	Falantes nativos
Mutabilidade	Aberto
Variações históricas	Sincrônico (*corpus* contemporâneo)
Disponibilidade	Disponível na *Web*

Quadro 1: A tipologia do *corpus* para a construção de uma *terminet.*

Compilação dos textos

- Estratégia de compilação dos textos
  - Busca em massa na *web*
    - motor de busca “genérico” (p.ex.: *Google*);
    - motor de busca dedicado à compilação de *corpus* (p.ex.: *WebCorp*);
    - *offline browsers* (p.ex.: HTTrack)
    - ferramentas específicas de compilação de textos (p.ex.: Corpógrafo e BootCat).
  - Seleção de páginas de forma “manual” de acordo com um projeto específico de *corpus*
- Critérios de seleção das páginas
  - Páginas de instituições públicas ou de conteúdo confiável.
- Especificação dos termos de busca
  - Delimitação do domínio da EaD
    - educação a distância/ead + gestão/ docência/ discência/ tecnologia

Pré-processamento dos textos (1)

- As estratégias de realização das etapas do pré-processamento:

1. Conversão semiautomática dos textos para o formato *txt*
   - PDF (Adobe Acrobat 9 Pro) > TXT
     (a) Quantidade reduzida de dados corrompidos;
     (b) Exclusão de figuras e imagens do texto durante a conversão;
     (c) Transformação de tabelas em listas de palavras.

2. Limpeza “manual” dos textos pós-conversão, ou seja, exclusão de:
   a) Dados corrompidos (p.ex.: equações e fórmulas);
   b) Dados referentes à paginação, cabeçalho, rodapé, autoria, filiação e referência bibliográfica

Pré-processamento dos textos (2)

3. Nomeação padronizada dos arquivos, anotação estrutural dos textos e geração de cabeçalho;
   - Portal de Córpus > Header Editor (http://www.nilc.icmc.usp.br:8180/portal/)
   - ❖ Anotação morfossintática
     - Etiquetador PALAVRAS
     - Formatos XML e SCES (PLN-Br)

Figura 1:*Header Editor* do Portal de Córpus.

4. Armazenamento do *corpus*
   - Portal de Córpus *(web)*
   - Máquina local

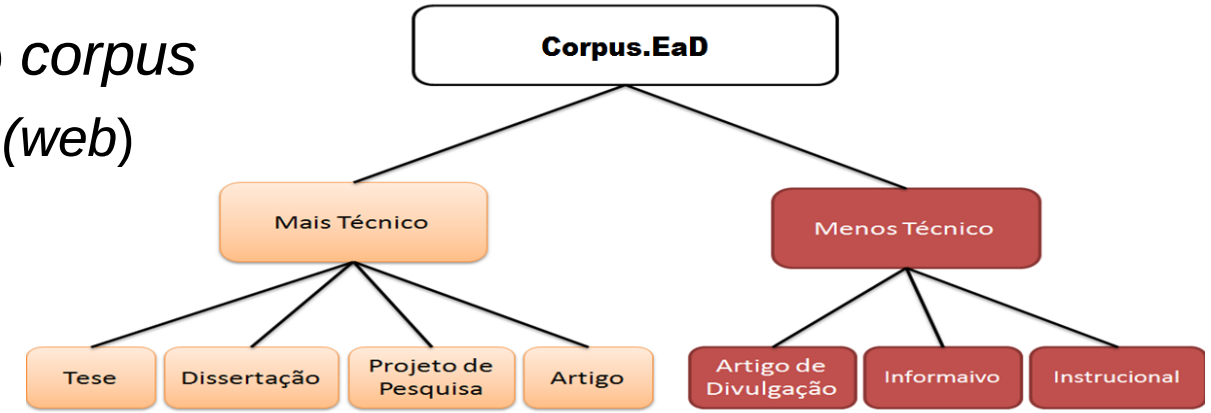

Figura 2: Organização local do *Corpus.EaD.*

Disponibilização do *corpus*

- Portal de Córpus
  - Disponibilização restrita aos membros do projeto TermiNet
- *Website* do projeto TermiNet
  - Em construção
  - Via cadastro do pesquisador e envio de solicitação de senha para *download.*

Figura 3: A disponibilização do *Corpus.EaD.* no Portal de Córpus.

Estatísticas do *Corpus*.EaD

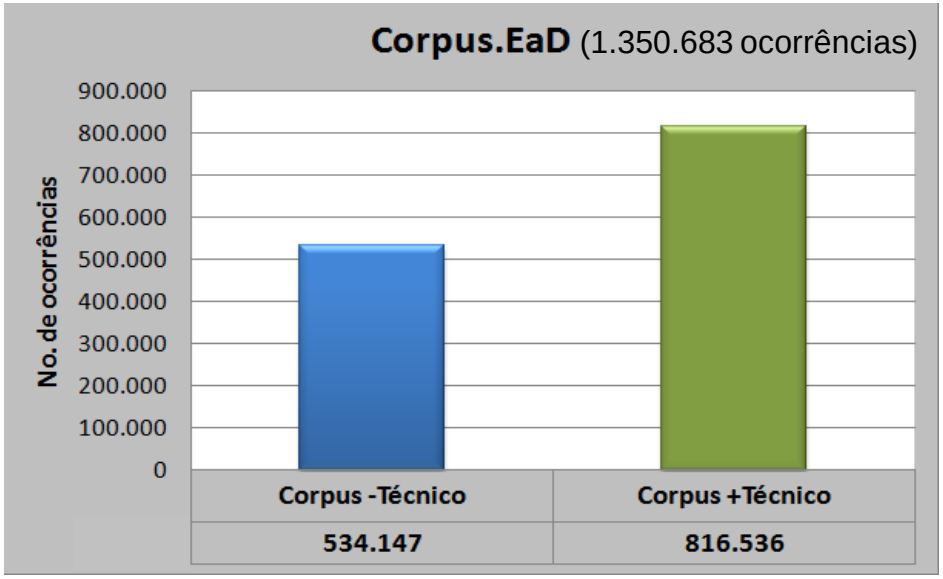

Gráfico 1: Número de ocorrências por *subcorpus.*

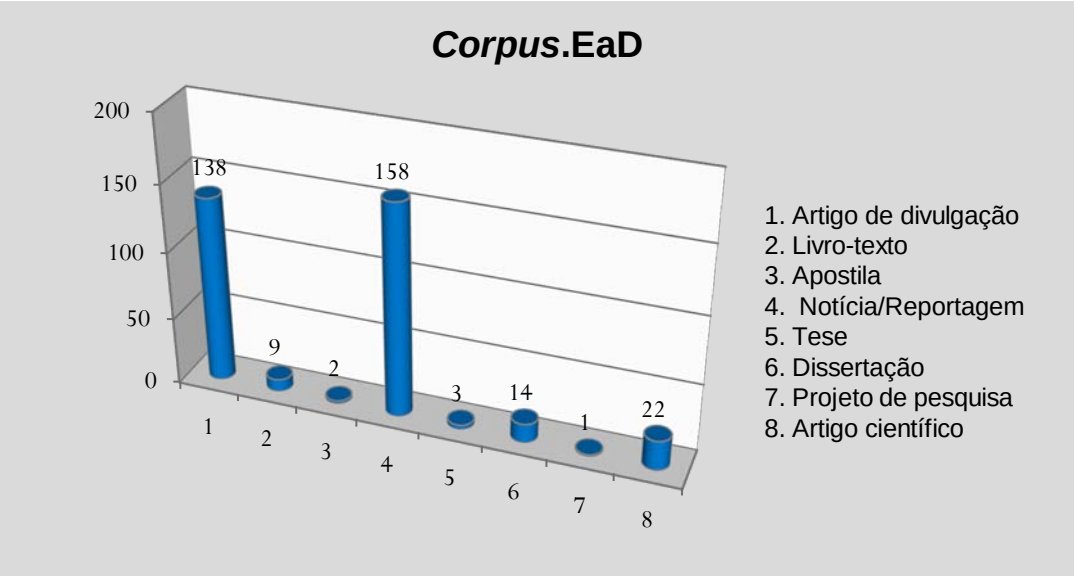

Gráfico 2: Porcentagem de textos por gênero.

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 1: *Corpus.*EaD: número de textos por gênero

Considerações finais

- As estratégias de construção do *corpus* adotadas no projeto TermiNet se mostraram viáveis;
- Apesar das estratégias de compilação de textos em massa da *web,* a seleção manual das fontes e dos textos se mostrou mais eficaz no cenário do TermiNet;
- O *Corpus.*EaD é o primeiro do referido domínio em PB;
- Do *Corpus.*EaD, estão sendo extraídos automaticamente os candidatos a termo que formarão os *synsets* da WordNet.EaD.