



# 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*

<b>Tamanho</b>	Médio-grande (ao menos, 1 milhão )
<b>Balanceamento</b>	Por gênero
<b>Modalidade</b>	Escrito (vs <i>corpus</i> de áudio)
<b>Tipo textual</b>	Escrito (vs <i>corpus</i> com transcrições)
<b>Meio</b>	Jornais, livros, revistas, manuais e outros
<b>Cobertura da língua</b>	<i>Corpus</i> especializado
<b>Gêneros</b>	Técnico-científico, científico de divulgação, informativo e instrucional
<b>Quantidade de línguas</b>	Monolíngue
<b>Anotação</b>	Anotado (em nível morfossintático)
<b>Comunidade produtora</b>	Falantes nativos
<b>Mutabilidade</b>	Aberto
<b>Variações históricas</b>	Sincrônico ( <i>corpus</i> contemporâneo)
<b>Disponibilidade</b>	Disponível na <i>Web</i>

**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)

- 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.

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

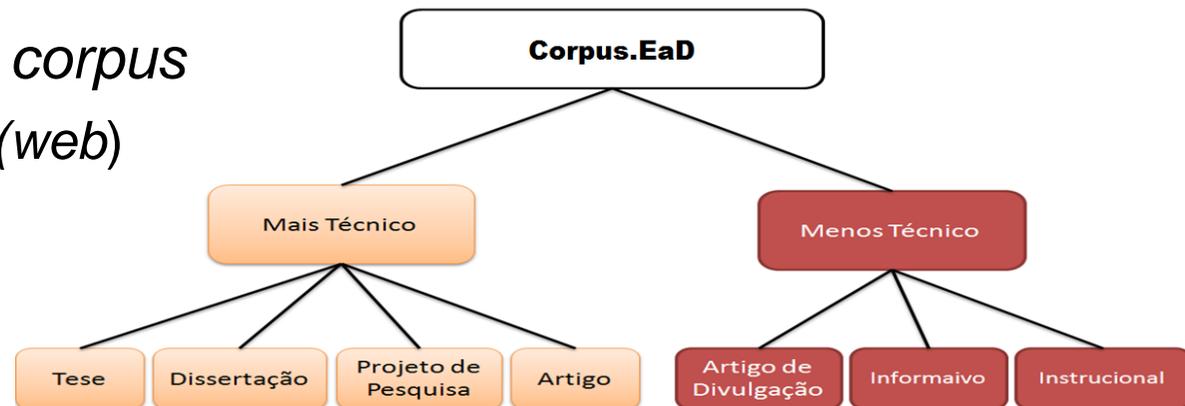


Figura 2: Organização local do *Corpus.EaD*.

# 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*.

Portal de Córpus PROJETO PLN-BR

USP  
PUC-RS  
UNISINOS  
PUC-RJ  
Mackenzie  
UFScar  
UNESP-Araraquara

VERSÃO 1.0  
Monday, October 4th of 2010. [RSS-Feed](#) [contact](#)

menu

- Home
- Corpus Selection
- Header Editor and Corpus Uploader
- Project Description
- Corpus Available and Text Classification Schema
- Manuals
- Downloads
- Related Publications
- Team
- Support
- Collaborators
- FAQ
- Contact us

user: mister\_jackbauer  
logged since: 04/10/10 14:58:59  
last access: 22/09/10 08:41:19  
searches: 50

Log out

**Select one corpus**

**1. PLN-BR GOLD - Córpus público**

PLN-BR GOLD é o *corpus* gold standard do Projeto PLN-BR, formado por textos do jornal Folha de São Paulo que podem ser acessados integralmente na Web sem necessidade de senha de acesso. Ele é uma amostra aleatória estratificada e proporcional à distribuição do *corpus* global do projeto PLN-BR (chamado de PLN-BR FULL) com relação aos textos dos cadernos do jornal. Ele é formado por 1% dos textos do *corpus* PLN-BR FULL, o que equivale a 1.024 textos, e possui somente notícias e reportagens para as quais a Folha de São Paulo possui direitos de republicação. Este *corpus* está contido no *corpus* PLB-BR CATEG, também criado no escopo do projeto PLN-BR. O *corpus* PLN-BR FULL, por sua vez, é formado por 103,080 mil textos do jornal Folha de São Paulo, compondo um ano construído a partir do ano de 1994 (toma um mês aleatório até o ano de 2005). A classificação em notícias e reportagens foi feita de forma automática usando-se um classificador de tipos de textos treinado com os 40 tipos de textos do Projeto Lácio-Web (<http://www.nilc.icmc.usp.br/lacioweb/>) no *corpus* montado para o projeto de doutorado de Rachel Aires que foi defendido no ICIMC-USP em 2005 sob orientação da Profa. Sandra Aluísio (mais informação sobre o classificador em <http://www.nilc.icmc.usp.br/nilc/projects/linguarudo.html>).

[Search](#)  Full download(6,27MB)  Download just the text files (1,43MB)

**4. Corpus EAD - Córpus privado**

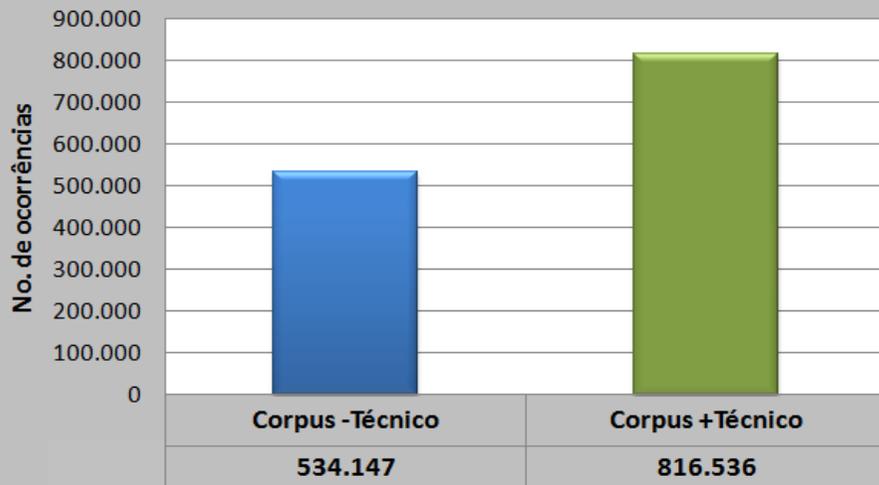
O *Corpus.EaD* é um *corpus* especializado do domínio da Educação a Distância (EaD) em português do Brasil (PB). Esse *corpus* foi construído no âmbito do projeto Terminet, cujos objetivos são: (a) especificar uma metodologia semiautomática para a construção de *wordnets* terminológicas a partir de *corpus* e (b) validar tal metodologia com a construção da WordNet.EaD. O referido *corpus* possui 1.350.683 ocorrências e 347 textos compilados manualmente de sites "confiáveis" relacionados à EaD. No *Corpus.EaD*, os textos estão organizados em dois *subcorpora*. O *subcorpus* –Técnico possui 534.147 ocorrências distribuídas em 307 textos dos gêneros científico de divulgação, instrucional e informativo. O *subcorpus* +Técnico possui 816.536 ocorrências distribuídas em 40 textos dos gêneros tese, dissertação, projetos de pesquisa e artigos científicos.

[Search](#)  Full download(5 9672 MB)  Download just the text files (0.1989 MB)

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

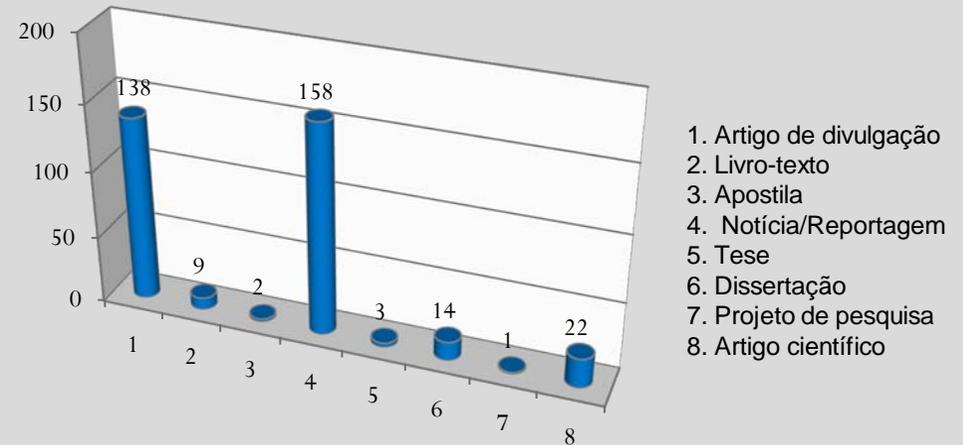
# Estatísticas do *Corpus.EaD*

**Corpus.EaD** (1.350.683 ocorrências)



**Gráfico 1:** Número de ocorrências por *subcorpus*.

**Corpus.EaD**



**Gráfico 2:** Porcentagem de textos por gênero.

<i>Subcorpora</i>	Gêneros Textuais	Tipos Textuais	Quantidade	Total
-Técnico	Científico de divulgação	Artigos de Divulgação	138	307
		Instrucional	Livro-Texto	
	Informativo	Apostila	2	
+Técnico		Técnico-científico	Notícias/Reportagens	158
	Tese		3	
	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.