Pontifícia Universidade Católica do Rio Grande do Sul
Faculdade de Informática
Programa de Pós-Graduação em Ciência da Computação

Integração de ferramentas para compilação e exploração de *corpora*

Filipe Pereira da Silveira

Dissertação apresentada como requisito parcial à obtenção do grau de mestre em Ciência da Computação.

Orientadora: Profa. Dra. Vera Lúcia Strube de Lima

Porto Alegre, agosto de 2008.

Dados Internacionais de Catalogação na Publicação (CIP)

S587i Silveira, Filipe Pereira da.
Integração de ferramentas para compilação e exploração de corpora / Filipe Pereira da Silveira. – Porto Alegre, 2008.
99 f.

Diss. (Mestrado) – Fac. de Informática, PUCRS.
Orientador: Profa. Dra. Vera Lúcia Strube de Lima

1. Informática. 2. Linguística Computacional. 3. Linguística de Corpus. 4. Processamento de Textos (Computação). I. Lima, Vera Lúcia Strube de. II. Título.

CDD 006.35

Ficha Catalográfica elaborada pelo
Setor de Tratamento da Informação da BC-PUCRS

Pontifícia Universidade Católica do Rio Grande do Sul
FACULDADE DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

TERMO DE APRESENTAÇÃO DE DISSERTAÇÃO DE MESTRADO

Dissertação intitulada ***"Integração de Ferramentas para Compilação e Exploração de Corpora"***, apresentada por Filipi Pereira da Silveira, como parte dos requisitos para obtenção do grau de Mestre em Ciência da Computação, Inteligência Computacional, aprovada em 25/08/08 pela Comissão Examinadora:

Profa. Dra. Vera Lúcia Strube de Lima – PPGCC/PUCRS
Orientadora

Profa. Dra. Milene Selbach Silveira – PPGCC/PUCRS

Profa. Dra. Sandra Maria Aluísio – ICMC/USP

Prof. Dr. Carlos Augusto Prolo – FACIN/PUCRS

Homologada em 14/07/09, conforme Ata No. 12 pela Comissão Coordenadora.

Prof. Dr. Fernando Gehm Moraes
Coordenador.

PUC
Campus Central
Av. Ipiranga, 6681 - P32 - sala 507 - CEP: 90619-900
Fone: (51) 3320-3611 - Fax (51) 3320-3621
E-mail: ppgcc@inf.pucrs.br
www.pucrs.br/facin/pos

Dedico este trabalho aos meus pais,
que sempre apoiaram e incentivaram meus estudos.

AGRADECIMENTOS

Agradeço à minha família, pelo carinho, pela compreensão e por ter sempre apoiado minhas escolhas, e ter investido desde cedo na minha formação. Agradeço ao meu pai, exemplo de vontade e dedicação, pelos conselhos e agradeço principalmente à minha mãe pelo apoio incondicional nos momentos mais difíceis.

Agradeço à minha professora e orientadora, Vera Lúcia Strube de Lima, pela experiência e conhecimentos transmitidos. Obrigado especialmente pela paciência, pela compreensão, pela disponibilidade, e pelas palavras de incentivo durante toda minha caminhada no mestrado.

Agradeço à minha namorada pelo carinho e pela compreensão nos momentos finais da escrita.

Agradeço também aos meus amigos e colegas da PUCRS e do TRT, em especial ao meu amigo e colega Flávio Knob pelo apoio e incentivo. Agradeço também aos amigos que fiz no NILC, em especial ao colega Arnaldo Cândido e a professora Sandra Aluísio pelo apoio e acolhida em São Carlos.

Agradeço à Josiane Brandolt, Lilian Teixeira e Susana Azeredo pelas contribuições e pelo precioso tempo dedicado aos experimentos.

Agradeço à CAPES por ter financiado meus estudos e permitido o desenvolvimento deste trabalho.

RESUMO

Este trabalho realiza um estudo da tipologia e disponibilidade de *corpora*. São discutidas questões referentes ao projeto de um corpus no que se refere a sua compilação. São apresentadas funcionalidades para exploração de *corpora* e analisadas ferramentas e recursos disponíveis para trabalhar com *corpus*. A seleção de ferramentas adequadas para compilação e exploração de *corpora* de textos em língua natural representa hoje um desafio aos pesquisadores da área. Muitas das ferramentas disponíveis dependem de licenças e plataformas específicas para serem executadas, limitam o uso de vários formatos de documento ou criam padrões próprios de codificação de corpus e de anotações, dificultando a criação, a interoperabilidade e o compartilhamento de recursos lingüísticos entre grupos de pesquisas. Nesse sentido é apresentada e descrita uma ferramenta para a lingüística de *corpus* que construímos e oferecemos à comunidade de pesquisadores em língua portuguesa – a ferramenta ENTRELINHAS. Esta ferramenta facilita a compilação e agrega funcionalidades essenciais para exploração de *corpora*. A ferramenta adere a um formato de codificação compatível com o Portal de Córpus do NILC/USP contribuindo com o intercâmbio de recursos para o processamento da língua portuguesa. Uma análise quanto ao uso dessa ferramenta também é apresentada.

ABSTRACT

In this work we present a brief study on the taxonomy and availability of text corpora in order to introduce questions concerning corpus design and corpus compiling. We present corpus exploring functionalities and we bring comments on available tools and resources to work with corpora. Selecting the suitable tools for corpora compiling and analysis is still a challenge to researchers in the field. Many of the available tools are commercially distributed, depend on specific platforms, restrict file format usage or create their own standards for corpus codification and annotation, what makes it more difficult to interoperate and to share linguistic resources among research groups. In this context we present and detail ENTRELINHAS, the corpus linguistics tool we built and we make available to Portuguese language researchers in this field. ENTRELINHAS eases corpus compiling and makes basic resources for Portuguese language corpora exploring available. The tool adheres to an encoding standard that keeps it compatible with NILC/USP's Portal de Córpus. A discussion on the report of the use of ENTRELINHAS is also presented.

ÍNDICE

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AC/DC - Acesso a corpora/Disponibilização de corpora
ANC – American National Corpus
API – Application Programming Interface
BNC - British National Corpus
CLIR – Cross-Language Information Retrieval
CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico
CNRS - Centre National de la Recherche Scientifique
CRPC - Corpus de Referência do Português Contemporâneo
DELA – Dictionnaires Électroniques du LADL
ELRA - European Language Resources Association
DTD – Document Type Definition
ICAME - International Computer Archive of Modern and Medieval English
ICU – International Components for Unicode
IGM - Institut Gaspard-Monge
J2EE – Java 2 Platform Enterprise Edition
JSP – Java Server Pages
JSTL – Java Standard Tag Library
JVM – Java Virtual Machine
LDC - Linguistic Data Consortium
LGPL - Lesser General Public License
LORIA - Laboratoire Lorrain de Recherche en Informatique et ses Applications
NILC - Núcleo Interinstitucional de Lingüística Computacional
NSP – Ngram Statistics Package
OTA - The Oxford Text Archive
POS - *Part-of-speech*
SEU - Survey of English Usage
SGML - Standard Generalized Markup Language
TEI – Text Encoding Initiative
URL – Uniform Resource Locator
UTF-8 – 8-bit Unicode Transformation Format
UTF-16 – 16-bit Unicode Transformation Format
XML – Extensible Markup Language

Capítulo 1

Introdução

A linguagem[1], elemento fundamental da comunicação, está constantemente evoluindo e se modificando. A cada dia, novas palavras são incorporadas ao vocabulário, enquanto outras caem em desuso ou ganham novos significados. O ser humano é capaz de compreender e se adaptar rapidamente a essas evoluções e a toda a variedade lingüística que nos rodeia. Porém, esse dinamismo da língua torna-se um enorme desafio ao processamento da linguagem natural, passo fundamental para aproximar a comunicação entre humanos e computadores.

A lingüística, ciência que estuda a linguagem humana, é uma ciência empírica, na qual todo o conhecimento é resultado de nossas observações e experiências sobre o uso da linguagem. A lingüística de *corpus* estuda a linguagem através de amostras de "textos reais". *Corpus* (plural *corpora*) é uma grande coleção de textos com milhares de palavras escritas por humanos (RUSSELL; NORVIG, 2003) e é também a base do processamento estatístico da linguagem e da lingüística de *corpus*.

A abordagem empírica nos estudos sobre a linguagem, através da lingüística de *corpus*, era muito comum entre os anos de 1920 e 1960. Porém, por volta de 1960, a abordagem racionalista passou a dominar grande parte das pesquisas, especialmente devido aos trabalhos de Noam Chomsky (1957). Nos últimos 20 anos, a abordagem empírica vem experimentando uma espécie de renascimento, impulsionada principalmente pela crescente disponibilidade de

[1] Apesar de, na língua portuguesa, o termo *linguagem* ter um significado distinto de *língua*, neste trabalho, os dois termos serão utilizados com o mesmo significado, como empregado na língua inglesa.

corpora eletrônicos. Este recente crescimento no número de *corpora* eletrônicos fez com que as pesquisas em processamento estatístico da linguagem e lingüística de *corpus* voltassem a se intensificar e contribuir com a descrição da linguagem (KENNEDY, 1998).

A lingüística de *corpus* serve a diversas aplicações como, por exemplo, o estudo de como ensinar e aprender uma língua. A análise de um *corpus* contribui para o processamento computacional da língua natural fornecendo evidências que melhoram a descrição da estrutura e do uso das línguas (KENNEDY, 1998).

Santos (1998) agrupa os pesquisadores da área de lingüística de *corpus* basicamente em dois grupos: compiladores de *corpora* e usuários de *corpora*. Os compiladores de *corpora* preocupam-se especialmente com questões tais como criar, estruturar e anotar *corpora*. Já os usuários de *corpora*, preocupam-se em extrair informações a partir dos *corpora*. Além destes, emerge nos grupos de pesquisa e vem ganhando espaço os desenvolvedores de ferramentas para *corpora*.

1.1. Motivações

A lingüística de *corpus* pode contribuir significativamente para a descrição da linguagem natural e, apesar do crescimento no número de recursos e ferramentas disponíveis nessa área, ainda existem muitos problemas em aberto. A seleção de ferramentas adequadas às necessidades de cada projeto representa hoje um desafio aos pesquisadores da área.

Muitas ferramentas para compilação e exploração de *corpora* utilizam padrões de codificação de texto e anotações incompatíveis entre si, dificultando a interoperabilidade e o compartilhamento de recursos lingüísticos entre grupos de pesquisas. Outras ferramentas, algumas comerciais, necessitam plataformas específicas para serem executadas e não suportam adequadamente esquemas de codificação de caracteres para a língua portuguesa.

1.2. Objetivos

O objetivo geral deste trabalho é fazer um estudo amplo sobre *corpora* e ferramentas para *corpora*, propondo e desenvolvendo uma ferramenta que possa ser distribuída sem custos, voltada a usuários da língua portuguesa.

Tem-se como objetivos específicos:

- contribuir com o projeto PLN-BR[2] disponibilizando uma ferramenta para compilação e exploração de *corpora* aderente ao formato de codificação XCES, adequado para a língua portuguesa e compatível com o Portal de Córpus[3];
- facilitar a compilação de *corpora* permitindo que formatos populares de arquivos sejam utilizados;
- oferecer na ferramenta funcionalidades básicas de exploração de *corpora*. A solução deve ser independente de plataforma e integrar ferramentas já existentes para a implementação de suas funcionalidades.

1.3. Organização desta dissertação

No Capítulo 2 apresentaremos o modo como alguns autores definem o que é um *corpus* e introduzimos as noções de tamanho e representatividade de um corpus. Apresentaremos também o modo como *corpora* costumam ser classificados. Descrevemos alguns dos maiores *corpora* em língua inglesa e língua portuguesa hoje existentes e abordaremos aspectos sobre uso da *web* como um *corpus*.

No Capítulo 3 abordaremos as principais etapas envolvidas na compilação de um *corpus*. Discutiremos questões e critérios relacionados ao projeto de um *corpus* e a coleta de documentos. Descreveremos tarefas relacionadas à preparação, à segmentação e à anotação dos textos. Apresentaremos também projetos e iniciativas em vista de estabelecer um padrão de codificação para corpora e anotações.

No Capítulo 4 apresentaremos funcionalidades básicas de exploração de *corpora* e algumas aplicações baseadas em *corpus*.

No Capítulo 5 apresentaremos algumas ferramentas para compilação e exploração de *corpus*, abordando seus principais recursos e funcionalidades.

No Capítulo 6 apresentaremos características e funcionalidades da ferramenta Entrelinhas, que desenvolvemos no âmbito da compilação e da exploração de *corpora* em língua portuguesa.

---

[2] PLN-BR (Recursos e Ferramentas para a Recuperação de Informação em Bases Textuais em Português do Brasil) financiado pelo CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), projeto #550388/2005-2.

[3] http://incubadora.fapesp.br/projects/ portal-corpus

No Capítulo 7 apresentaremos experiências realizadas com a Ferramenta e mostraremos o relato de um usuário especializado quanto às suas funcionalidades, em decorrência de suas percepções no uso da Entrelinhas.

No Capítulo 8 apresentaremos as considerações finais, retomando os assuntos abordados nos capítulos anteriores e lições aprendidas, e indicaremos trabalhos futuros sobre a ferramenta Entrelinhas e sobre o tema estudado.

Capítulo 2

Corpora

O termo *corpus* vem do latim e significa corpo, conjunto. *Corpora* (plural de *corpus*) lingüísticos são enormes coleções de textos, criteriosamente selecionados, apresentando exemplos escritos ou falados em uma língua. Na literatura, há várias definições sobre o que exatamente constitui um *corpus*.

De acordo com Manning e Schütze (1999):

> *"Um corpo de textos é chamado corpus – corpus é simplesmente o termo em latim para 'corpo' e, quando você tem muitas coleções de textos, você tem corpora."*

Kennedy (1998) escreve:

> *"Na Lingüística, um corpus é um corpo de texto escrito ou de falas transcritas que pode servir de base para análise e descrição lingüística."*

Sardinha (2004) analisou a definição de corpus apresentada por vários autores. Por mencionar vários aspectos importantes – tais como a origem e a formatação dos dados, o propósito, a composição, a representatividade a extensão do corpus – o autor selecionou a definição *corpus* dada por Sanchez, Cantos e Cumbre (apud SARDINHA, 2004) como a mais completa:

> *"Um conjunto de dados linguísticos (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 linguístico 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."

Atualmente, a maioria dos *corpora* está armazenada em formato eletrônico, porém nem sempre *corpora* foram armazenados eletronicamente para serem lidos por computadores. *Corpora* eletrônicos passaram a ser usados mais intensamente a partir dos anos 60. Teoricamente poderíamos considerar um *corpus* qualquer coleção de textos, em formato eletrônico ou não (KENNEDY, 1998; MCENERY; WILSON, 1996).

Ainda que não sejam exatamente textos, existem também *corpora* de falas gravadas, que não serão estudados neste trabalho. Neste documento, consideraremos *corpus* como uma coleção de textos computacionalmente armazenada e organizada para o estudo de fenômenos lingüísticos ou a criação de ferramentas computacionais, como etiquetadores e corretores ortográficos.

Corpora são geralmente extensos e, quanto maior um *corpus*, mais representativo ele é. Geralmente, o tamanho de um *corpus* é medido pela quantidade de palavras que ele contém. *Corpora* devem conter o maior número possível de estruturas existentes na linguagem pertencentes ao domínio de interesse. A representatividade de um *corpus* permite que dele sejam extraídas observações confiáveis sobre o uso da língua (GASPERIN; STRUBE DE LIMA, 2001). Diz-se que um *corpus* é representativo quando ele é uma amostra representativa da população de interesse, ou seja, os resultados obtidos na amostra também valem para a população pesquisada (MEGERDOOMIAN, 2003).

2.1. Tipologia de *corpora*

Corpora podem ser compilados para diferentes finalidades, a partir de diferentes tipos e quantidades de texto, em diferentes línguas. Estes *corpora* podem diferenciar-se significativamente uns dos outros em diversos aspectos como, por exemplo: origem dos textos, gênero dos textos, áreas de domínio dos textos, idioma dos textos e disposição interna. Estes diferentes tipos de *corpora* buscam atender a diferentes propósitos. O objetivo desta seção é apresentar os tipos mais comuns de *corpora* citados na literatura.

2.1.1. *Corpora* escritos e falados

Corpora podem ser do tipo escrito ou falado. A maior parte dos *corpora* são do tipo escrito, pois são mais fáceis de obter e processar. Um *corpus* escrito é formado simplesmente pela produção escrita da língua em estudo, como: textos de livros, revistas, jornais, artigos, páginas na Internet, etc. Enquanto que *corpora* falados, também chamados de *corpora* de falas, são compostos por falas transcritas, que podem ser originadas, por exemplo, de diálogos, monólogos ou conversas telefônicas.

Em um *corpus* falado, a transcrição das falas implica o tratamento de aspectos fonéticos e prosódicos bastante complexos. Esse tratamento da transcrição torna a compilação de um *corpus* falado muito mais lenta e cara (KENNEDY, 1998). Transcrições de um *corpus* falado, normalmente, possuem muitas contrações, vários tipos de representações fonéticas, variações de pronúncia, interjeições e muitos fragmentos de frases e de palavras. O tratamento de todas estas ocorrências ainda representa um enorme desafio (MANNING; SCHÜTZE, 1999).

Existem *corpora* formados de textos escritos e também de falas transcritas. Um exemplo é o Survey of English Usage (SEU) Corpus[4]. O SEU Corpus, considerado um dos mais importantes *corpora* pré-eletrônicos e mais tarde convertido para o formato eletrônico, possui 100 textos de origem escrita e 100 textos de origem falada em inglês britânico.

2.1.2. *Corpora* balanceados e especializados

Corpora podem ser classificados quanto ao gênero dos textos em dois tipos principais: balanceados ou especializados. Um *corpus* balanceado normalmente é compilado quando se deseja construir um *corpus* para uso geral, sem um objetivo específico de pesquisa. *Corpora* balanceados são formados de textos de diferentes gêneros e domínios, estes incluídos em iguais quantidades por gênero, ou em quantidades proporcionais à relevância que cada gênero de texto tem na língua. É bastante comum encontrar-se *corpora* balanceados também chamados de *corpora* genéricos, *corpora* equilibrados ou *core corpora* (KENNEDY, 1998).

Corpora especializados**,** também chamados de *corpora* oportunistas ou especiais, são compilados para objetivos específicos de pesquisa. Um *corpus*

[4] http://www.ucl.ac.uk/english-usage/

especializado pode objetivar o estudo, por exemplo, do desenvolvimento da linguagem falada ou escrita por crianças, de uma língua falada ou escrita através de falantes não nativos dessa língua ou, ainda, do uso de uma língua numa área específica do conhecimento, como medicina ou informática. A maior parte dos *corpora* especializados foi compilada tendo como objetivo o estudo de aspectos sociais do uso da língua e variações regionais da linguagem. Portanto, *corpora* específicos de um dialeto ou de uma região também são classificados como *corpora* especializados (KENNEDY, 1998).

Algumas vezes a noção de *corpus* balanceado tem sido usada de maneira relativa. Um *corpus* especializado pode ser balanceado, dentro do escopo de um dado domínio, se forem incluídos vários gêneros de texto dentro do domínio em estudo. Por exemplo, se o objetivo é analisar fenômenos lingüísticos em notícias econômicas, então o *corpus* pode ser balanceado pela inclusão de notícias de vários jornais, revistas ou outras fontes, com o objetivo de capturar diferentes estilos, vocabulário ou padrões (MEGERDOOMIAN, 2003).

2.1.3. *Corpora* estáticos e dinâmicos

Um *corpus* estático é um *corpus* compilado e planejado para ser uma amostra finita da linguagem, e nenhum aumento ou diminuição do *corpus* ocorre dinamicamente. *Corpora* dinâmicos, também chamados de *corpora* orgânicos, são *corpora* que podem crescer ou diminuir dinamicamente, opondo-se aos *corpora* estáticos.

Um *corpus* monitor é um tipo de *corpus* dinâmico cujo conteúdo é constantemente substituído. Através das substituições se inserem novos textos e, em quantidade equivalente, se removem os textos mais antigos. O objetivo das substituições é compor um *corpus* capaz de refletir as mudanças da língua. *Corpora* monitores são normalmente usados para fins lexicográficos pois, por serem constantemente reciclados, representam o estado atual da língua (HERNÁNDEZ, 2002; MEGERDOOMIAN, 2003).

2.1.4. *Corpora* de estudo, de referência e de treinamento

Corpora podem ser classificados quanto a sua finalidade. Denomina-se *corpus* de estudo o *corpus* principal de uma pesquisa, que contém os textos que representarão a língua que se pretende descrever.

Muitas pesquisas contam também com *corpora* de referência e *corpora* de treinamento. *Corpora* de referência são utilizados para a comparação de

diferenças em relação a outros *corpora*, podendo, por exemplo, guiar a compilação de *corpora* de estudos. *Corpora* de treinamento, por sua vez, são utilizados para o teste de ferramentas de PLN em desenvolvimento (SARDINHA, 2004).

2.1.5. *Corpora* monolíngües, multilíngües, paralelos e alinhados

Corpora também podem ser classificados como monolíngües ou multilíngües. *Corpora* monolíngües contêm textos em uma única língua, enquanto *corpora* multilíngües contêm textos em diferentes línguas. *Corpora* multilíngües que contêm o mesmo texto em mais de uma língua são chamados *corpora* paralelos ou comparáveis. Entretanto, nem todo *corpus* multilíngüe é um *corpus* paralelo. O *corpus* jurídico Aarthus, por exemplo, contém textos em inglês, francês e dinamarquês, porém não tem qualquer compromisso com a equivalência e a tradução entre seus textos (MCENERY; WILSON, 1996; SARDINHA, 2004).

Corpora paralelos são formatados para que seus textos, em diferentes línguas, possam ser facilmente comparados. Este tipo de *corpus* é utilizado em sistemas mais recentes de tradução automática. Porém, para tornar um *corpus* paralelo realmente valioso, é preciso identificar quais sentenças e palavras são traduções umas das outras. *Corpora* que trazem estas anotações são chamados de *corpora* alinhados.

Corpora bilíngües têm sido muito usados em projetos de tradução automática baseados em métodos estatísticos de desambiguação contextual do significado. O mais conhecido *corpus* bilíngüe de textos paralelos e alinhados é o Canadian Hansards, formado pela transcrição dos debates do parlamento canadense. O *corpus* consiste de textos em francês e inglês, equivalentes de tradução um do outro. O Canadian Hansards[5] está disponível no Linguistic Data Consortium (LDC) (MEGERDOOMIAN, 2003).

2.1.6. *Corpora* sincrônicos, diacrônicos, históricos ou contemporâneos

Corpora podem ser classificados como sincrônicos ou diacrônicos (KENNEDY, 1998). *Corpora* sincrônicos são formados por textos que refletem o estado de uma língua em um momento específico no tempo. Geralmente *corpora* sincrônicos são formados por textos contemporâneos à época em que foram

[5] http://www.parl.gc.ca/common/chamber.asp

criados. Sardinha (2004) classifica *corpora* contemporâneos como um tipo específico de *corpus*. A maioria dos *corpora* são sincrônicos: um exemplo é o Brown Corpus[6] (consulte a Seção 2.2.1), que contém textos em inglês americano publicados em 1961.

Corpora diacrônicos são formados por textos escritos em diferentes períodos de tempo, ou seja, são formados por textos históricos. Alguns autores, como Sardinha (2004), classificam *corpora* históricos, com textos escritos no passado, como um tipo específico de *corpus*. *Corpora* diacrônicos ajudam os lingüistas a estudar e entender as mudanças de uma língua. O primeiro *corpus* eletrônico diacrônico é o Helsinki Corpus of English Texts[7], que contém textos que vão do século XIII ao século XVIII, como mostrado na Tabela 1.

Tabela 1 – Distribuição das palavras da parte diacrônica do Helsinki Corpus of English Texts (KYTÖ, 1996)

Período	Sub-períodos	Palavras	%
Old English	I -850	2 190	0.5
	II 850-950	92 050	22.3
	III 950-1050	251 630	60.9
	IV 1050-1150	67 380	16.3
	Total	**413 250**	**100.0**
Middle English	I 1150-1250	113 010	18.6
	II 1250-1350	97 480	16.0
	III 1350-1420	184 230	30.3
	IV 1420-1500	213 850	35.1
	Total	**608 570**	**100.0**
Early Modern English (British)	I 1500-1570	190 160	34.5
	II 1570-1640	189 800	34.5
	III 1640-1710	171 040	31.0
	Total	**551 000**	**100.0**

Um dos principais *corpus* diacrônicos em língua portuguesa é o Corpus Anotado do Português Histórico Tycho-Brahe[8]. O Tycho-Brahe contém textos em prosa, escritos em português europeu entre os séculos XVI e XIX, anotados morfológica e sintaticamente.

2.2. *Corpora* disponíveis

Há pouco mais de 20 anos eram pouquíssimos os *corpora* em formato eletrônico disponíveis, a maioria sem fins lucrativos. Porém, o rápido desenvolvimento da informática que ocorreu nas últimas décadas permitiu que mais e mais pesquisadores trabalhassem com *corpora*. Nos anos 90 muitos projetos de compilação de *corpora* surgiram em todo o mundo. Muitos dos

[6] http://icame.uib.no/brown/bcm.html
[7] http://khnt.hit.uib.no/icame/manuals/HC/INDEX.HTM
[8] http://www.tycho.iel.unicamp.br/

maiores projetos foram concebidos para fins comerciais, principalmente por editores de dicionário (KENNEDY, 1998).

Em Manning e Schütze (1999) é apresentada uma lista com as maiores organizações que distribuem *corpora*. A lista com os endereços atualizados das organizações pode ser vista na Tabela 2. A maioria delas cobra pelos *corpora* e o custo depende da finalidade do uso. Obviamente, licenças para uso comercial tendem a ser mais caras que licenças educacionais ou privadas.

Tabela 2 – Distribuidores de *corpora* eletrônicos e seus endereços na Internet (MANNING; SCHÜTZE, 1999) (atualizado)

Organização	Endereço na Internet
Linguistic Data Consortium (LDC)	http://www.ldc.upenn.edu/
European Language Resources Association (ELRA)	http://www.elra.info/
International Computer Archive of Modern and Medieval English (ICAME)	http://icame.uib.no/
The Oxford Text Archive (OTA)	http://ota.ahds.ac.uk/
Child Language Data Exchange System	http://childes.psy.cmu.edu/

2.2.1. *Corpora* em língua inglesa

Diversos *corpora* eletrônicos estão disponíveis para análise em várias línguas, a maior parte deles em língua inglesa. Alguns dos maiores e mais importantes *corpora* lingüísticos existentes em língua inglesa são:

- American National Corpus[9] (ANC): O objetivo do projeto American National Corpus (IDE; MACLEOD, 2001; IDE; SUDERMAN, 2004, 2006) é criar um grande *corpus* de textos de vários gêneros e falas transcritas da língua inglesa americana produzidos a partir de 1990. O projeto pretende que o ANC chegue a 100 milhões de palavras. A edição mais recente do *corpus* ANC tem 22 milhões de palavras e é disponibilizada pelo LDC.
- British National Corpus[10] (BNC): O BNC (ASTON, 1996; BURNAGE; DUNLOP, 1992) foi completado em 1994 e é um *corpus* com aproximadamente 100 milhões de palavras. É composto tanto por textos em língua escrita como falada. Os textos que formam BNC provêm de uma enorme variedade de fontes e buscam representar o inglês britânico utilizado no final do século XX. A parte escrita do BNC representa cerca de 90% do *corpus* e é composta por amostras de

[9] http://americannationalcorpus.org/
[10] http://www.natcorp.ox.ac.uk/

jornais, revistas, livros acadêmicos, romances, cartas e trabalhos escolares, entre outros. A parte falada do BNC representa cerca de 10% do *corpus* e é composta por uma grande quantidade de conversações informais registradas por voluntários selecionados, de diferentes idades, regiões e classes sociais. As conversas estão demograficamente balanceadas e foram coletadas em diferentes contextos como, por exemplo: reuniões governamentais, programas de rádio e chamadas telefônicas. O *corpus* foi codificado de acordo com as regras do Text Encoding Initiative[11] (TEI) usando Standard Generalized Markup Language (SGML).

- Brown Corpus: Criado em 1964 na Brown University[12], o Standard Sample of Present-Day American English, ou simplesmente Brown Corpus (FRANCIS; KUČERA, 1971), como é mais conhecido, é historicamente importante por ter sido o primeiro *corpus* eletrônico a ser criado. O Brown Corpus é genérico e foi criado para ser uma amostra representativa do inglês americano usado em 1961. Apesar de hoje ser considerado pequeno e pouco atualizado, o Brown Corpus ainda é muito utilizado, pois serviu como modelo para muitos outros *corpora*. A estrutura do *corpus*, que conta com 500 amostras de texto, pode ser vista na Tabela 3. O Brown Corpus é disponibilizado pelo LCD e pelo ICAME.

Tabela 3 – Estrutura do Brown Corpus
(FRANCIS; KUČERA, 1971, traduzido do inglês pelo autor)

I. Não-ficção (374 textos)			
A. Imprensa: reportagens			
	Diário	Semanal	Total
Política	10	4	14
Esporte	5	2	7
Sociedade	3	0	3
Últimas notícias	7	2	9
Finanças	3	1	4
Cultural	5	2	7
Total			44
B. Imprensa: editorial			
	Diário	Semanal	Total
Institucional	7	3	10
Pessoal	7	3	10
Relatórios industriais			2
Catálogos universitários			1
Construção civil			1
Total			30
J. Didático			
Ciências naturais			12
Medicina			5
Matemática			4
Ciências sociais			14
Ciência política, direito e educação			15
Humanas			18
Engenharia e tecnologia			12
Total			80
II. Ficção (126 textos)			

[11] http://www.tei-c.org/
[12] http://www.brown.edu/

Cartas para o editor	5	2	7
Total			27

C. Imprensa: críticas e resenhas (teatro, livros, música, dança)

Diário	Semanal	Total
14	3	17

D. Religião	
Livros	7
Periódicos	6
Discursos	4
Total	17
E. Lazer e talentos	
Livros	2
Periódicos	34
Total	36
F. Tradições	
Livros	23
Periódicos	25
Total	48
G. Literatura, biografias, memórias, etc.	
Livros	38
Periódicos	37
Total	75
H. Miscelânea	
Documentos governamentais	24
Relatórios institucionais	2
K. Geral	
Romances	20
Contos	9
Total	29
L. Suspense e policial	
Romances	20
Contos	4
Total	24
M. Ficção científica	
Romances	3
Contos	3
Total	6
N. Aventura e faroeste	
Romances	15
Contos	14
Total	29
P. Romances e histórias de amor	
Romances	14
Contos	15
Total	29
R. Humor	
Romances	3
Ensaios, etc.	6
Total	9
Total	**500**

- Penn Treebank[13]: O Penn Treebank é um grande *corpus* com cerca de 4,5 milhões de palavras. Todas as palavras do *corpus* foram anotadas com suas classes gramaticais e rótulos marcando a análise sintática. Este *corpus* foi coletado do jornal The Wall Street Journal[14]. Apesar de ser bastante usado, o Penn Treebank não está disponível gratuitamente (MEGERDOOMIAN, 2003).

2.2.2. *Corpora* em língua portuguesa

A principal fonte de *corpora* em língua portuguesa é o projeto AC/DC[15] (Acesso a corpora/Disponibilização de corpora) da Linguateca[16]. Alguns dos *corpora* existentes em língua portuguesa são:

[13] http://www.cis.upenn.edu/~treebank/
[14] http://online.wsj.com/
[15] http://acdc.linguateca.pt/

- *Corpus* NILC/São Carlos: O *corpus* NILC/São Carlos (PINHEIRO; ALUÍSIO, 2003) do Núcleo Interinstitucional de Lingüística Computacional[17] (NILC) contém aproximadamente 35 milhões de palavras em português brasileiro contemporâneo. O *corpus* é composto por diversos tipos de textos, como: didáticos, jornalísticos, legais, literários, entre outros. A Figura 1 mostra um trecho de uma notícia de turismo contida no *corpus*. O Corpus NILC/São Carlos está disponível gratuitamente para fins de pesquisa na página da Linguateca.

```
<ext id=165601 cad="Turismo" sec="soc"
sem="94a">
<s> <t> Manhattan abriga em 11 mil km de
ruas algumas das melhores diversões do
planeta </t> </s>
<s> <caixa> Central Park, oásis urbano de
340 hectares, dezenas de arranha-céus,
centros culturais, museus indispensáveis,
restaurantes e bairros boêmios como o
Greenwich Village, garantem lazer 24 horas
para o visitante </caixa> </s>
<s> <a> Da enviada especial a Nova York </a>
</s>
<p>
<s> São 11 mil quilômetros de ruas em
Manhattan e, portanto, há passeios por toda
a parte . </s>
<s> Dos tradicionais locais públicos, o
mais gostoso de visitar é o Central Park  .
</s>
</p>
<p>
<s> Para os nova-iorquinos, o parque que se
estende da rua 59 até a 110, entre a Quinta
Avenida e a Central Park West parece ser um
oásis dentro da cidade . </s>
</p>
</ext>
```

Figura 1 – Trecho do *Corpus* NILC/São Carlos
Fonte: (LINGUATECA, 2006)

- CETEMPúblico: O Corpus de Extractos de Textos Electrónicos MCT/Público (CETEMPúblico) é um *corpus* com aproximadamente 180 milhões de palavras da área jornalística em português europeu. O *corpus* foi criado pelo projeto Processamento Computacional do Português (projeto que deu origem à Linguateca) após a assinatura de um protocolo entre o Ministério da Ciência e da Tecnologia (MCT) português e o jornal Público em Abril de 2000. O Público, fundado em 1990, é o primeiro jornal diário português de grande circulação a disponibilizar uma edição eletrônica na rede. A Figura 2 mostra um trecho de uma notícia esportiva contida no *corpus*. O CETEMPúblico está disponível gratuitamente para fins de pesquisa na página da Linguateca através do projeto AC/DC (ROCHA; SANTOS, 2000; SANTOS; ROCHA, 2001).

---

[16] A Linguateca (http://www.linguateca.pt/) é um centro de recursos para o processamento computacional da língua portuguesa criado pelo projeto Processamento Computacional do Português, com financiamento do Ministério da Ciência e Tecnologia de Portugal.

[17] O Núcleo Interinstitucional de Lingüística Computacional (http://www.nilc.icmc.usp.br) é um núcleo de pesquisa criado em 1993 com o objetivo de fortalecer a pesquisa e o desenvolvimento de projetos na área de Lingüística Computacional e Processamento da Linguagem Natural. O grupo reúne pesquisadores da Universidade de São Paulo (USP) em São Carlos, Universidade Federal de São Carlos (UFSCar) e Universidade Estadual Paulista (UNESP) de Araraquara.

```
<ext n=1360002 sec=des sem=92a>
<p>
<s>
O Farense aceitou o natural domínio dos primeiros 20'.
</s>
<s>
Depois, controlou o jogo, na sequência do seu dinâmico contra-ataque, onde Pitico, pela direita, e Djukic, pelo outro lado, criavam perigo para as redes de Zivanovic.
</s>
</p>
<p>
<s>
Só de bola parada e num livre de Baía (18'), é que os madeirenses se aproximaram da baliza de Lemajic.
</s>
<s>
Mas, no minuto seguinte, o protagonista seria o Farense, ao ver o árbitro Carlos Valente anular um golo, por irregularidade, num remate de Ricardo, a cruzamento de Pitico.
</s>
<s>
O União, após várias tentativas de aproximação da área do adversário, acabou por perder uma grande oportunidade de golo (30') quando o jugoslavo Lepi falha o cabeceamento à entrada da baliza, na sequência da marcação de um canto.
</s>
</p>
</ext>
```

Figura 2 – Trecho do CETEMPúblico (LINGUATECA, 2006)

- CETENFolha: O Corpus de Extratos de Textos Eletrônicos NILC/Folha de São Paulo (CETENFolha) é um *corpus* de cerca de 24 milhões de palavras em português brasileiro. O *corpus* foi criado no âmbito do projeto Processamento Computacional do Português (mesmo projeto que criou o CETEMPúblico) com base nos textos do jornal Folha de São Paulo que fazem parte do *corpus* NILC/São Carlos, compilado pelo NILC. O CETENFolha está disponível gratuitamente para fins de pesquisa na página da Linguateca através do projeto AC/DC.

- CRPC: O Corpus de Referência do Português Contemporâneo (CRPC, 2008; NASCIMENTO, 2000) (CRPC) teve sua criação iniciada em 1988 no Centro de Lingüística da Universidade de Lisboa[18]. O *corpus* contém atualmente 201 milhões de palavras. O CRPC é constituído por amostras de diversos tipos de texto escrito (literário, jornalístico, técnico, científico, didático, econômico, jurídico, parlamentar, etc.) e de falas transcritas (elocuções informais e formais). São amostras que guardam variedades nacionais e regionais do português de textos que vão desde a segunda metade do séc. XIX até 2002, sendo, na sua maior parte, posteriores a 1970. Estão incluídas, no *corpus*, amostras do português europeu, português do Brasil, português dos cinco países africanos de língua oficial portuguesa (Angola, Cabo Verde, Guiné-Bissau, Moçambique, São Tomé e Príncipe), português de Macau[19], português do Timor-Leste e do português de

[18] http://www.clul.ul.pt/

[19] Antiga colônia portuguesa que hoje é administrada pela República Popular da China.

Goa[20]. A distribuição entre as regiões se dá conforme a Figura 3. Note que o *corpus* é formado quase que totalmente pelo português europeu.

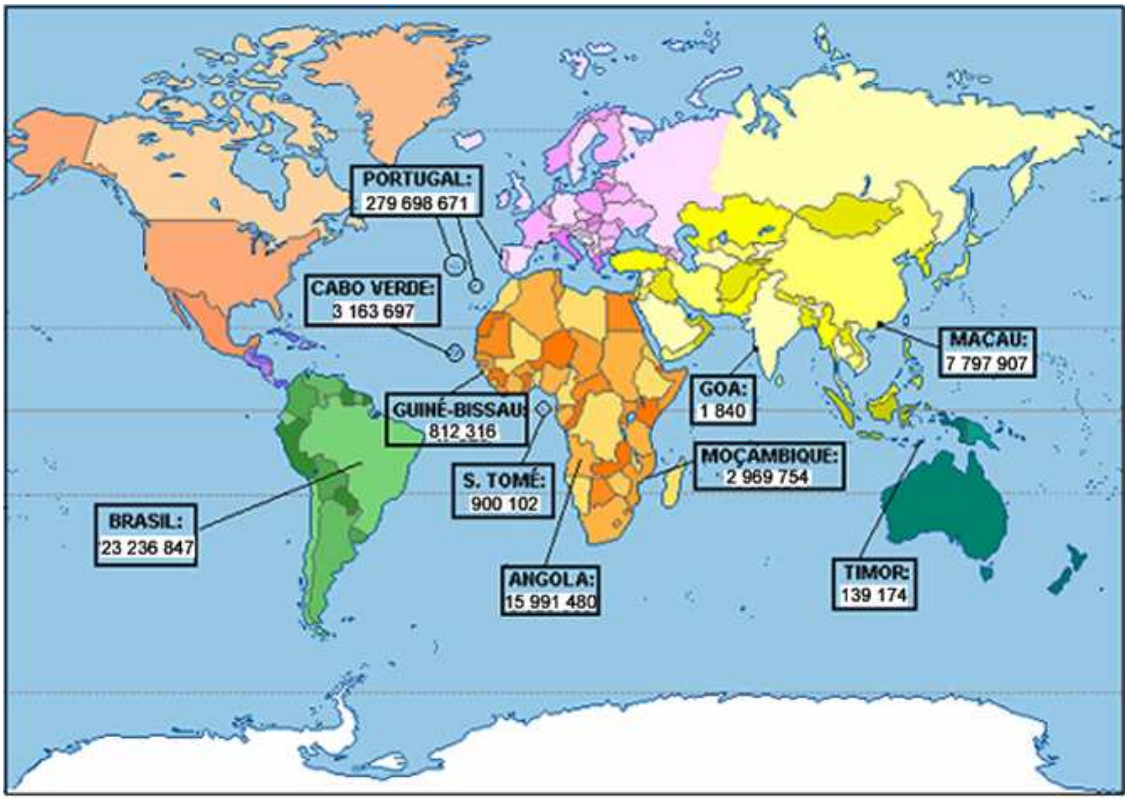


Figura 3 – Número de palavras de cada região do CRPC

(CRPC, 2008)

2.2.3. *Corpora* do projeto Lácio-Web

O projeto Lácio-Web[21] (ALUISIO et al., 2003) tem o objetivo de disponibilizar, principalmente para lingüistas e cientistas da computação, *corpora* e ferramentas lingüístico-computacionais. O projeto disponibiliza quatro *corpora* corretamente compilados de português brasileiro escrito contemporâneo. Os *corpora* estão catalogados e codificados de forma que possam ser facilmente intercambiados, analisados e utilizados.

Os quatro *corpora* disponibilizados no Lácio-Web são:

[20] Antiga colônia portuguesa que hoje é administrada pela Índia.

[21] O Lácio-Web (http://www.nilc.icmc.usp.br/lacioweb/) é um projeto financiado pelo CNPq resultante da parceria entre o NILC (Núcleo Interinstitucional de Lingüística Computacional), o IME/USP (Instituto de Matemática e Estatística da Universidade de São Paulo) e a FFLCH/USP (Faculdade de Filosofia, Letras e Ciências Humanas da USP).

- Lácio-Ref: é o *corpus* de referência do projeto Lácio-Web, composto de textos provenientes de diversos jornais, revistas, teses, dissertações, livros e informativos. As únicas anotações presentes neste *corpus* referem-se à existência de elementos gráficos, e cabeçalhos que contém informações bibliográficas e de catalogação. A grande maioria dos textos está integralmente disponibilizada.
- Par-C: *corpus* paralelo inglês-português que possui textos publicados durante um ano na revista Pesquisa FAPESP[22]. A revista divulga projetos científicos e tecnológicos financiados pela instituição.
- Comp-C: *corpus* de textos jurídicos em inglês de conteúdo comparável a textos jurídicos em português, armazenados no Lácio-Ref.
- Mac-Morpho: *corpus* fechado, formado por artigos publicados no jornal Folha de São Paulo em 1994. O Mac-Morpho contém mais de 1 milhão de palavras e foi anotado pelo etiquetador Palavras (BICK, 2000

2.2.4. Uso da *web* como um *corpus*

Mecanismos de busca na Internet servem como interfaces ou portas de entrada para a exploração da *web* como um poderoso recurso lingüístico (KEHOE; RENOUF, 2002). A *web* pode ser usada, por exemplo, para a extração do contexto de uso das palavras. A simples busca de uma palavra em um dos serviços mais populares de busca na Internet, como Google[23] ou Yahoo[24], é capaz de retornar milhares de trechos de textos (*snippets*) em que o termo pesquisado aparece destacado no contexto, com algumas palavras que o antecedem e que o sucedem.

Diversas publicações demonstram o potencial de uso da *web* para tarefas tais como tradução automática (GREFENSTETTE, 1999), extração de relações semânticas (CHKLOVSKI; PANTEL, 2004), identificação de colocações (SERETAN; NERIMA, 2004) ou desambiguação de palavras (SANTAMARÍA;

[22] http://www.revistapesquisa.fapesp.br/
[23] http://www.google.com/
[24] http://www.yahoo.com/

GONZALO, 2003). Mais adiante neste trabalho, na Seção 5.6, é apresentada a WebCorp[25], um ferramenta com a finalidade de usar a *web* como um *corpus*.

No entanto, a maior parte das ferramentas de busca na *web* existentes está preparada para a recuperação da informação, e não para a extração de dados lingüísticos. Rehbein (2008) destaca algumas desvantagens do uso da web como *corpus*: metadados não confiáveis, ausência de anotações, pouco controle sobre as ferramentas de busca e dificuldade para replicar os resultados devido as modificações que ocorrem nas páginas. A mesma autora apresenta algumas estratégias utilizadas para contornar o problema, tais como: a formatação dos dados retornados pelas ferramentas de busca e a criação do *corpus* a partir de documentos obtidos na *web*.

2.3. Considerações sobre este capítulo

Neste capítulo, apresentamos o modo como alguns autores definem o que é um *corpus* e introduzimos as noções de tamanho e representatividade de um *corpus*. Apresentamos também o modo como *corpora* costumam ser classificados quanto ao propósito para o qual foram criados, o gênero, as áreas de domínio, os idiomas, e a época em que foram produzidos os textos neles contidos. Em seguida apresentamos alguns dos maiores *corpora* em língua inglesa e língua portuguesa hoje existentes e abordamos aspectos sobre uso da *web* como um *corpus*.

No próximo capítulo apresentaremos as principais questões e atividades envolvidas no processo de compilação de um *corpus*, tais como: o projeto, a coleta, a preparação, a anotação e a codificação dos textos.

[25] http://www.webcorp.org.uk/

Capítulo 3

Compilação de *corpora*

Compilar – ou criar – um *corpus* é projetar e codificar uma coleção de documentos coletados dentro de determinados padrões ou exigências, para a realização de estudos lingüísticos ou computacionais de aprendizagem de máquina. Apesar do número de *corpora* disponíveis ter crescido significativamente nos últimos anos, alguns pesquisadores necessitam que novos *corpora* sejam criados. Isso ocorre quando os *corpora* existentes não atendem aos requisitos da pesquisa, possuem um custo de acesso muito elevado, ou, simplesmente, porque uma coleção específica de documentos faz parte da descrição do problema.

Vários aspectos envolvem a compilação de um *corpus*. Características como tamanho, balanceamento, representatividade, direitos autorais, conversão de formatos, limpeza de textos, meta-dados, inserção de anotações e padrões de codificação devem ser discutidas, estudadas e planejadas previamente, de acordo com o propósito do *corpus*. Nesta seção apresentaremos as principais tarefas envolvidas na compilação de um *corpus*: o projeto, a coleta dos documentos, a preparação dos textos e a codificação do *corpus*.

3.1. Projeto de um *corpus*

As características que um *corpus* deve ter não costumam ser objeto de consenso entre os pesquisadores. O tamanho, o balanceamento e a representatividade são características fortemente relacionadas aos objetivos da pesquisa na qual o *corpus* está inserido: alguns tipos de *corpora* tendem a ser mais adequados para tipos específicos de análise e alguns *corpora* simplesmente não são adequados para certos tipos de pesquisas. Mesmo quando concordam sobre

os requisitos que determinado *corpus* deve ter, muitos pesquisadores discordam quanto à forma prática de consegui-los (KENNEDY, 1998; SANTOS, 1998).

O tamanho de um *corpus*, normalmente medido pela quantidade de palavras, depende da quantidade de textos e do tamanho dos mesmos. Em geral, quanto maior um *corpus*, melhor. No entanto, o tamanho de um *corpus* normalmente é limitado pelo tempo e pelo volume de recursos disponíveis para o projeto, que por sua vez influenciam na quantidade de textos que podem ser obtidos e suas respectivas permissões de uso.

Outra característica importante refere-se ao fato de o *corpus* ser uma amostra balanceada da população em geral, ou ser um *corpus* especializado. Um *corpus* balanceado, como já foi discutido na Seção 2.1.2, é uma coleção que tenta cobrir o máximo possível de domínios, gêneros, tipos e estilos textuais. O projeto de um *corpus* deve especificar a quantidade e o tipo dos textos que serão incluídos. O *corpus* será composto de textos escritos ou de falas transcritas? Textos formais ou informais? Textos didáticos, jornalísticos, publicitários ou literários? Textos jurídicos, religiosos ou esportivos? Textos em prosa ou poesia? Narrativas ou dissertações? Textos históricos ou contemporâneos? Textos produzidos por adultos, crianças ou imigrantes? Em que proporção cada tipo deve ser incluído? Textos de diferentes gêneros e domínios devem ser incluídos em iguais quantidades ou em quantidades proporcionais à relevância que cada gênero ou domínio de texto tem na língua? O que verdadeiramente constitui um *corpus* balanceado ainda é uma questão em aberto (ARNOLD; BUCKLEY, 2006; HERNÁNDEZ, 2002; MEGERDOOMIAN, 2003).

É preciso especificar também se o *corpus* será ou não composto de textos completos. Um *corpus* pode ser formado por textos completos ou apenas por amostras dos textos originais. As amostras de textos são partes dos textos originais completos. Essas amostras podem ser de tamanho específico para o *corpus* inteiro. Para o estudo do estilo e do discurso, por exemplo, um *corpus* formado por amostras de textos de duas mil palavras, extraídas de vários textos, não é capaz de capturar confiavelmente características da estrutura interna de textos completos, onde se espera que as características lingüísticas das seções introdutórias e finais sejam diferentes. Estudos como estes requerem *corpora* de textos completos (KENNEDY, 1998).

Um *corpus* deve ser representativo da língua de uma população; a população não abrange, necessariamente, a linguagem como um todo. Os textos

podem ser amostrados a partir de sub-populações de acordo com a região, o gênero ou grupos específicos. Um *corpus* pode ser considerado representativo quando as descobertas feitas a partir dele puderem ser generalizadas para a língua como um todo (EVANS, 2008; KENNEDY, 1998).

3.2. Coleta dos textos

A coleta de textos (ARNOLD; BUCKLEY, 2006; STEFANOWITSCH, 2008) visa obter documentos textuais em formato eletrônico que atendam os requisitos estabelecidos para a composição do *corpus*. Em muitos casos, os textos de interesse podem não estar disponíveis em formato eletrônico, o que pode ocorrer com livros, documentos antigos, impressos ou manuscritos. A coleta destes textos implica digitação, digitalização de documentos impressos ou a transcrição de áudios, tarefas que podem levar muitos dias, pois necessitam ser realizadas de forma quase totalmente manual.

Mesmo documentos impressos, quando digitalizados e processados por software de reconhecimento óptico de caracteres[26], necessitam ser minuciosamente analisados e corrigidos por um humano. Este trabalho manual é imprescindível pois, mesmo com o alto nível de precisão alcançado por este tipo de software, os textos gerados são suscetíveis a erros e podem apresentar diversos problemas, tais como trechos não reconhecidos ou incorretos (EVANS, 2008; STEFANOWITSCH, 2008).

Normalmente, os documentos que atendem aos requisitos da pesquisa podem já estar em formato eletrônico, facilitando significativamente a coleta dos textos. Estes documentos podem estar acessíveis pela Internet, armazenados em discos rígidos, CD-ROMs, DVDs, gravados em um banco de dados, ou diretamente em um sistema de arquivos. Existem ferramentas especilizadas na captura de documentos da Internet, tais como HTTrack[27], WebZIP[28] ou WebCloner[29]. Essas ferramentas, conhecidas como navegadores *offline* ou *web crawlers*, podem acelerar o processo de coleta de textos (EVANS, 2008).

A maioria dos documentos, mesmo quando disponíveis na Internet, está protegida por direitos autorais. Convém sempre verificar se os textos coletados

[26] Reconhecedores óticos de caracteres, também conhecidos como OCR (*optical character recognition*), convertem documentos em formato de imagem para o formato texto.
[27] http://www.httrack.com/
[28] http://www.spidersoft.com/webzip/
[29] http://www.productsfoundry.com/webcloner/

podem ser utilizados da forma pretendida, caso contrário é essencial obter dos autores autorização explícita para a utilização dos mesmos. Permissões para uso de trechos dos textos apenas ou para fins não-comerciais, como pesquisas acadêmicas e uso pessoal, são habitualmente mais fáceis de serem obtidas (ARNOLD; BUCKLEY, 2006; EVANS, 2008; STEFANOWITSCH, 2008).

Durante a coleta dos documentos é recomendado registrar o máximo possível de informações referentes à autoria e à origem dos mesmos, tais como: autores, título, edição, editora, ano e local de publicação, endereço de *download*, idioma, esquema de codificação, entre outras. Mais tarde, essas informações poderão ser incorporadas aos textos, podendo vir a acelerar a obtenção das permissões de uso dos textos e a decodificação dos documentos (EVANS, 2008).

3.3. Preparação dos textos

Existe uma grande variedade de formatos de arquivo publicados na Internet; no entanto, a maioria ferramentas de análise de *corpora* aceita apenas alguns poucos formatos de documento como entrada. Mesmo que estejam em formato eletrônico, muitos dos documentos coletados não estarão prontos para ser lidos pelas ferramentas de PLN disponíveis (EVANS, 2008).

Preparar manualmente um *corpus* com algumas centenas de documentos, copiando e colando textos da área de transferência do computador, sem auxílio de um *software* conversor, pode levar meses. O trabalho de preparação dos textos, através da conversão desses arquivos para um formato adequado, depende muito da forma como foram produzidos e da disponibilidade de *software* conversores.

Alguns editores de texto mais sofisticados podem gerar arquivos em formatos bastante complexos. Muitos conversores podem não conseguir converter adequadamente alguns elementos presentes nos documentos originais ou, simplesmente, esses elementos não podem ser representados no formato aceito pelas ferramentas de análise de *corpora*. Em geral, ferramentas de exploração de *corpora* suportam como entrada apenas documentos em texto puro (txt) ou algum formato específico de codificação de *corpora*. Mesmo quando os textos de um *corpus* são codificados em formato diferente de txt, é comum que cópias em texto puro sirvam como formato intermediário na conversão do arquivo e sejam mantidas internamente na estrutura do *corpus*, com o objetivo de otimizar o processamento (EVANS, 2008; STEFANOWITSCH, 2008).

Formatos populares de documento como *HyperText Markup Language*, *Microsoft Word*, *Rich Text Format*, *Portable Document Format* ou *OpenDocument Format* são extremamente ricos e podem conter elementos não-textuais que podem ser perdidos durante conversão. É importante verificar com cuidado se *layout* de páginas, imagens, fórmulas, tabelas, legendas, estilos (negrito, itálico, sublinhado, etc.), números de linha, cabeçalhos e rodapés foram convertidos de modo satisfatório. Não há um modo definido de converter esses elementos, que podem ser descartados ou, quando interessantes para o projeto, convertidos em anotações (EVANS, 2008; MEGERDOOMIAN, 2003; STEFANOWITSCH, 2008).

Caracteres especiais, acentos, hifenizações, quebras de linha e de parágrafo, também requerem atenção. Problemas na representação de letras de alfabetos diferentes do inglês normalmente estão relacionados ao esquema de codificação de caracteres (*charset*) utilizado.

A versão do documento em texto puro, mesmo quando acompanhada de anotações, é apenas uma representação do documento original. A conservação dos documentos originais permite que falhas ou informações perdidas durante a preparação dos textos sejam recuperadas ou corrigidas manualmente, mesmo quando detectadas mais tarde.

3.4. Segmentação e anotação dos textos

Anotações (EVANS, 2008; HALTEREN, 1999; MEGERDOOMIAN, 2003; STEFANOWITSCH, 2008), também conhecidas como marcações ou etiquetas, são informações explicitamente adicionadas a um *corpus*. Apesar de um *corpus* desprovido de anotações representar um recurso bastante útil, um *corpus* anotado representa para a lingüística de *corpus* um recurso ainda mais útil e valioso. Anotações agregam valor a um *corpus* através da expansão das possibilidades de exploração, permitindo que buscas mais refinadas sejam realizadas. É possível acrescentar a um *corpus* anotações lingüísticas, ou referentes à autoria, à origem e à estrutura dos textos originais.

Informações referentes a autoria e origem dos textos, geralmente são obtidas durante a coleta dos textos. Acrescentar anotações com informações como título, ano de publicação ou até mesmo a origem, o sexo e a idade dos autores pode permitir o estudo de fenômenos sociais. Esse tipo de anotação, por estar vinculada a textos inteiros, normalmente é registrada em seções na forma

de cabeçalhos ou *headers*, independentemente da forma como o *corpus* está codificado.

Os textos também podem receber anotações lingüísticas de vários tipos. A segmentação do texto (MEGERDOOMIAN, 2003) é pré-requisito para os demais tipos de anotação. A segmentação, ou *tokenização*, delimita elementos lingüísticos contínuos, determinando, por exemplo, o início e o final de palavras, sentenças e parágrafos. As menores unidades de um texto são também conhecidas como *tokens*.

Em algumas línguas as palavras não são separadas por espaços em branco; no entanto, na língua portuguesa e em diversas outras línguas, a simples existência de um espaço em branco indica o fim de uma palavra e o início de outra. Há casos especiais, como "Porto Alegre" ou "guarda-chuva", em que pode ser interessante considerar dois segmentos como um único segmento. Há segmentadores que podem fazer essa segmentação de forma automática, como por exemplo, o QToken[30].

Há vários tipos e níveis de anotação lingüística, que podem enriquecer o texto com informações morfológicas, sintáticas e semânticas. Os tipos mais comuns de anotação lingüística em *corpus* são a etiquetagem das classes gramaticais das palavras, também conhecido como *Part-Of-Speech tagging* (*POS tagging*), e a lematização, onde cada palavra recebe uma etiqueta com seu lema.

O processo de anotar, ou etiquetar, um texto pode ser feito de modo manual ou automático, com auxílio de um *software* etiquetador. Atualmente, *software* como o Palavras (BICK, 2000), que fazem a etiquetagem de classes gramaticais, são capazes de marcar com um alto nível de precisão enormes quantidades de texto. A escolha do *software* etiquetador implica a escolha de um conjunto de etiquetas específico (HALTEREN, 1999).

Muitos outros tipos de anotação podem ser adicionados. Em algumas situações, pode ser interessante anotar o texto com informações sobre a estrutura e a formatação utilizada do documento original, indicando, por exemplo, a presença de cabeçalhos, rodapés, quebras de página, trechos em negrito, itálico, sublinhado, etc.

---

[30] Segmentador de texto gratuito, implementado em Java por Oliver Mason. Disponível em: http://www.english.bham.ac.uk/staff/omason/software/qtoken.html

3.5. Codificação dos textos e das anotações

O crescimento na disponibilidade de recursos lingüísticos nas últimas décadas, fez com que diversos formatos de codificação de textos, anotações surgissem. A maioria dos projetos de *corpus*, com objetivo de atender requisitos de ferramentas de anotação e exploração de *corpus* específicas, veio a criar seus próprios formatos para codificação de textos e anotações. A diversidade de formatos aumentou a importância da busca por padrões que facilitassem o compartilhamento, a combinação e o intercâmbio desses recursos. Entre os principais projetos e iniciativas em vista de estabelercer um padrão de codificação para textos e anotações, podemos destacar: MuchMore[31], Tiger-XML[32], Text Encoding Initiative[33] (TEI), Corpus Encoding Standard (CES), Corpus Encoding Standard for XML (XCES) e padrão ISO TC37/SC4[34].

Nas próximas seções apresentaremos o padrão da ISO TC37/SC4, o formato de codificação Corpus Encoding Standard for XML (XCES) e o formato de codificação do projeto MuchMore. A escolha por estes padrões ou formatos se deve à observação do crescente destaque com que eles vêm sendo adotados em sistemas e plataformas.

3.5.1. Padrão ISO TC37/SC4

O Technical Comitee 37 (TC37, *Terminology and Other Languages Resources*), da International Organization for Standardization (ISO), criou um sub-comitê (SC4) para preparar padrões internacionais e recomendações para a modelagem de dados, anotação, intercâmbio de dados e avaliação de recursos lingüísticos. Dentro do ISO TC37/SC4 um grupo de trabalho (WG1-1) foi criado para estabelecer um padrão internacional e prover um *framework* (IDE; ROMARY; DE LA CLERGERIE, 2004) para criação, anotação e manipulação de recursos lingüísticos que sirva de referência para diferentes esquemas de anotação e *software* de processamento. O *framework* deve permitir e estimular o intercâmbio e o reuso de anotações lingüísticas e, ao mesmo tempo, prover codificação e anotação flexíveis. O grupo, formado por especialistas de diversas universidades,

---

[31] O projeto MUCHMORE (http://muchmore.dfki.de) é coordenado pelo centro de pesquisas alemão Deutsche Forschungszentrum für Künstliche Intelligenz (http://www.dfki.de), por Paul Buitellar e Thierry Declerck. O projeto desenvolve tecnologias para a construção de um sistema de recuperação de informações interlíngua, *cross-language information retrieval*, para o domínio médico.

[32] http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/doc/html/TigerXML.html

[33] http://www.tei-c.org/index.xml

[34] http://www.tc37sc4.org/

reuniu-se em novembro de 2002 e identificou princípios e requisitos fundamentais para o desenvolvimento do *framework*.

O *framework* deve permitir a representação de qualquer variedade de informação lingüística, seja ela geral ou específica, e não deve impor ou restringir nenhuma teoria lingüística. As estruturas de representação devem possuir uma semântica definida e, descritores e categorias de informação devem ser compartilhados de maneira centralizada e consistente, evitando que mecanismos distintos sejam usados para descrever o mesmo tipo de informação.

Os dados devem ser representados através de uma nomenclatura padrão, de forma legível para humanos. Além disso, os dados devem também poder ser inseridos a qualquer momento, de forma incremental e, informações específicas devem poder ser facilmente extraídas ou separadas. Dessa forma, deve ser possível isolar camadas específicas da anotação de outras camadas, permitindo que novas informações de anotação sejam adicionadas e somente apontem para os dados originais (*stand-off annotation*) ao invés destas estarem em meio aos dados. Para isso, o modelo de dados deve separar claramente estrutura e conteúdo, porém mantendo-os mapeáveis entre si. O documento anotado deve poder ser facilmente manipulado pelo usuário, e seu mapeamento deve ser documentado em um XML Schema[35] (ou equivalente) associado ao modelo de dados.

O *framework* deve poder ser estendido e ser independente de mídia, capaz de lidar com imagens e vídeos. O mapeamento entre documento anotado e formato final deve poder ser feito através folhas de estilo (*schema-derived stylesheet*). O formato final deve ser permitir a serialização[36] e a desserialização de dados.

3.5.2. XCES – Corpus Encoding Standard for XML

O XCES[37] (IDE; BONHOMME; ROMARY, 2000), Corpus Encoding Standard for XML[38], é um conjunto bastante aceito de padrões de codificação

[35] http://www.w3.org/XML/Schema
[36] A serialização permite que dados, quando gravados ou transmitidos através de uma rede, sejam lidos e carregados corretamente.
[37] O XCES (http://www.xml-ces.org/) foi criado pelo Departamento de Ciência da Computação do Vassar College (http://www.vassar.edu) e a equipe francesa Langue et Dialogue do Laboratoire Lorrain de Recherche en Informatique et ses Applications (http://www.loria.fr/) e do Centre National de la Recherche Scientifique (http://www.cnrs.fr/)
[38] http://www.w3.org/XML/

para aplicações do processamento da língua natural baseado em *corpus*. O XCES estabelece um nível de codificação mínimo para que um *corpus* seja considerado padronizado, e essa padronização refere-se à representação descritiva do *corpus*, como a marcação de informações estruturais e tipográficas. Devemos observar que o XCES ainda está em desenvolvimento.

O objetivo inicial do XCES foi prover um *framework* de acesso e representação para o American National Corpus. O *framework* beneficiou a comunidade de engenheiros da linguagem com a especificação de um formato que permite grande interoperabilidade entre diferentes tipos de anotações e anotações diferentes de um mesmo fenômeno, servindo de interface entre diferentes tipos de anotações lingüísticas. No XCES as anotações devem poder ser facilmente definidas e validadas, não impondo ao desenvolvedor valores ou elementos específicos.

O XCES é a versão XML do CES, Corpus Encoding Standard. O CES integra o EAGLES Guidelines[39]. O EAGLES Guidelines provê um conjunto de padrões de codificação de aplicações para processamento da língua natural baseado em *corpus*. O CES é uma aplicação Standard Generalized Markup Language[40] (SGML) compatível com as especificações do TEI Guidelines for Electronic Text Encoding and Interchange, da Text Encoding Initiative. O CES foi projetado para servir de maneira adequada às pesquisas em engenharia da linguagem e aplicações.

XML, eXtensible Markup Language, é uma linguagem de marcação derivada do SGML. XML tornou-se um padrão amplamente aceito para representação e intercâmbio de dados na rede mundial de computadores. A evolução do CES para o XCES permitiu o uso de diversos mecanismos disponíveis para o XML como, por exemplo: XSTL[41] (eXtensible Stylesheet

[39] O EAGLES Guidelines foi desenvolvido pelo grupo de pesquisa europeu Expert Advisory Group on Language Engineering Standards (http://www.ilc.cnr.it/EAGLES96/home.html).
[40] http://www.w3.org/MarkUp/SGML
[41] http://www.w3.org/TR/xslt

Language Transformations), XML Schemas, XSL[42] (eXtensible Stylesheet Language), XPointer[43] e XLink[44] (XML Linking Language).

O XCES especifica uma arquitetura de dados para *corpora* e provê DTDs (Document Type Definitions) para codificar a estrutura básica dos documentos e suas anotações lingüísticas. Há DTDs XML e XML Schemas do XCES para a codificação de dados segmentados, dados com anotações gramaticais, dados alinhados, entre outros. Na Tabela 4 estão listados e brevemente descritos os oito XML Schemas para a codificação de documentos e anotações lingüísticas providos pelo XCES.

Tabela 4 – *XML Schemas* providos pelo *XCES*

XML Schemas	Descrição
`xcesDoc.xsd`	Codificação para XCES Documents (*level 1*).
`xcesAna.xsd`	Codificação para anotações.
`xcesAlign.xsd`	Codificação para dados alinhados.
`xcesWord.xsd`	Estende xcesDoc para prover *tags* em nível de palavras para anotações *stand-off*.
`xcesSpoken.xsd`	Estende xcesDoc para a codificação de falas transcritas.
`xcesHeader.xsd`	Cabeçalho usado por todos XCES Documents.
`xcesGlobal.xsd`	Definições de comuns de elementos e atributos.
`xcesLink.xsd`	Usado para importar o *Xlink namespace*.

3.5.3. MuchMore

O projeto MuchMore (BUITELAAR et al., 2003; VINTAR et al., 2006), Multilingual Concept Hierarchies for Medical Information Organization and Retrieval, provê um *framework* que permite que as tecnologias correlatas existentes possam ser integradas e aperfeiçoadas, ao mesmo tempo que novas tecnologias possam ser desenvolvidas. Dentre as várias contribuições do projeto destaca-se a descrição de um formato de anotação lingüística e semântica baseado em XML.

O formato de anotação do MuchMore é capaz de integrar múltiplos níveis de análise lingüística. Os níveis de informação, que incluem anotações morfológicas, sintáticas e semânticas, podem ser organizados separadamente e

[42] http://www.w3.org/TR/xsl/
[43] XPointer (http://www.w3.org/TR/xptr/) é um sistema para o endereçamento de componentes de um arquivo XML.
[44] http://www.w3.org/TR/xlink/
[46] http://www.bncweb.info/

referenciados entre si através de identificadores. O formato de anotação, desenvolvido especialmente para atender às necessidades do projeto, especifica uma DTD, Document Type Definition, que define a estrutura de um documento XML. Na Figura 4 é apresentado um exemplo de anotação lingüística XML no MuchMore sobre uma passagem de texto.

No MuchMore, o texto é representado pelo elemento <*text*> que, por sua vez, é composto de um ou mais elementos <*token*>, que identificam as palavras do texto e carregam consigo informações morfossintáticas, além da forma canônica de cada palavra.

```
Balint syndrom is a combination of symptoms including simultanagnosia, a disorder of
spatial and object-based attention, disturbed spatial perception and representation, and
optic ataxia resulting from bilateral parieto-occipital lesions.

<text>
        <token id="w1" pos="NN">Balint</token>
        <token id="w2" pos="NN">syndrom</token>
        <token id="w3" pos="VBZ" lemma="be">is</token>
        <token id="w4" pos="DT" lemma="a">a</token>
        <token id="w5" pos="NN" lemma="combination">combination</token>
        ...
        <token id="w20" pos="JJ" lemma="spatial">spatial</token>
        <token id="w21" pos="NN" lemma="perception">perception</token>
        <token id="w22" pos="CC" lemma="and">and</token>
        <token id="w23" pos="NN" lemma="representation">representation</token>
        ...
        </text>
<chunks>
        <chunk id="c1" from="w1" to="w2" type="NP"/>
        <chunk id="c7" from="w20" to="w23" type="NP"/>
</chunks>
```

Figura 4 – Exemplo de anotações lingüísticas no MUCHMORE

(BUITELAAR et al., 2003)

No formato utilizado pelo MuchMore, o texto, com suas informações morfossintáticas e sintagmáticas, são armazenados em um único arquivo. Desta forma, é necessário repetir o mesmo texto em vários arquivos, de maneira redundante, quando se deseja distribuir diferentes informações lingüísticas em diferentes arquivos. Essa redundância pode representar um significativo desperdício de espaço para armazenamento. Há outros aspectos desfavoráveis, tais como a escassez de informações morfossintáticas, a ausência de informações de gênero e número das palavras, além da ausência de informações semânticas (SOUZA et al., 2006).

3.6. Considerações sobre este capítulo

Neste capítulo apresentamos as principais tarefas envolvidas nas etapas de compilação de um *corpus*: o projeto, a coleta dos documentos, a preparação

dos textos e a codificação do *corpus*. Apresentamos também o padrão ISO, o formato de codificação Corpus Encoding Standard for XML (XCES) e o formato de codificação do projeto MuchMore. No próximo capítulo apresentaremos as principais questões e atividades relacionadas à exploração de *corpora* e suas aplicações.

Capítulo 4

Exploração e uso de *corpora*

Até recentemente, na lingüística computacional as abordagens tradicionais concentravam-se no desenvolvimento de mecanismos formais e na melhora das fontes de conhecimento para análise da linguagem. Porém tem se confirmado uma tendência de empregar a abordagem quantitativa com foco nas técnicas estatísticas de aquisição de conhecimento, usando *corpora* (MEGERDOOMIAN, 2003).

O processamento estatístico da linguagem viabiliza a análise e a compreensão da linguagem natural através de uma abordagem baseada em *corpus*, com o uso de estatística e aprendizado de máquina. Geralmente essa abordagem conduz a modelos probabilísticos da linguagem, que podem ser aprendidos a partir de dados, de maneira mais simples que gramáticas formais e, assim, colocados em evidência com a crescente disponibilização de *corpora* (RUSSELL; NORVIG, 2003).

Modelos probabilísticos apresentam vantagens, pois podem ser treinados a partir de dados: o aprendizado do modelo ocorre, em última análise, a partir da contagem de ocorrências. A tolerância a erros e a capacidade de lidar com qualquer palavra, tornam os modelos probabilísticos bastante robustos. Além disso, esses modelos são capazes de representar o fato de que as pessoas nem sempre concordam sobre todos os aspectos no uso da língua e, nos casos de ambigüidade, a estatística pode ser usada para interpretar os dados da maneira mais provável (RUSSELL; NORVIG, 2003).

Segundo Manning e Schütze (1999), os três principais requisitos para o processamento estatístico da linguagem natural são computadores, *corpora* e

software. Computadores, porque *corpora* são geralmente grandes e demandam recursos computacionais compatíveis com grandes volumes de texto (diferentemente da época em que os primeiros *corpora* foram criados, atualmente os computadores não representam um custo significativo). *Corpora*, porque são base do processamento estatístico da linguagem natural (felizmente, hoje, muitas organizações distribuem *corpora* gratuitamente). E *software*, como ferramentas de busca, de marcação e outras, para analisar os dados de um *corpus*.

4.1. Exploração de *corpora*

Através da exploração de listas de palavras, concordâncias ou colocações feitas sobre um *corpus* pode-se obter informações que contribuem, por exemplo, para a melhora no ensino e no aprendizado de uma língua. Nesta seção apresentaremos três funcionalidades, muito comuns em ferramentas de exploração de *corpora*: contadores de ocorrências (às vezes chamados de listas de palavras ou contadores de freqüência), concordanciadores e buscadores de colocações.

4.1.1. Contadores de ocorrências

Contadores de ocorrências realizam contagens e calculam a freqüência de ocorrências de itens lexicais ou palavras em um *corpus*. O cálculo das freqüências permite a identificação de palavras-chave e *stopwords*. A contagem de palavras pode contribuir também para a identificação de documentos similares, sendo usada em abordagens voltadas à recuperação da informação, classificação automática de documentos e outras.

As contagens mais comuns, que podem ser aplicadas a *corpora* inteiros ou textos específicos, são:

- o número total de palavras;
- o número de palavras distintas;
- o número de ocorrências de cada palavra e sua freqüência em relação ao total de palavras;
- o número de textos em que a palavra ocorre e sua freqüência em relação ao total de textos.

Alguns contadores mais sofisticados permitem diferenciar maiúsculas de minúsculas, detectar palavras com mesmo lexema ou desprezar deliberadamente

a acentuação. Algumas ferramentas que dispõem de contadores de palavras são: WordSmith (Seção 5.3), BNCWeb[46], TACT[47], CLAN[48] e Corsis[49]. Na Figura 5 é possível ver a lista de palavras do Corsis.

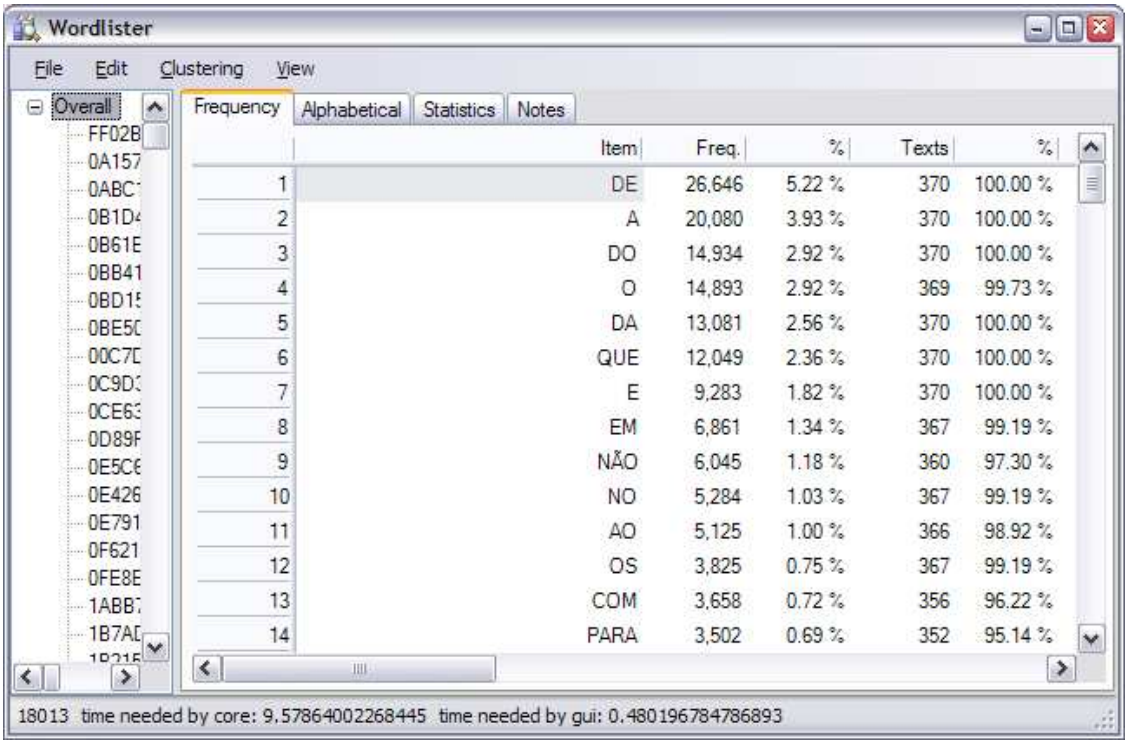


Figura 5 – Lista de palavras no Corsis

4.1.2. Concordanciadores

Concordâncias são listas de palavras ou seqüências de palavras dentro de um contexto. As concordâncias são muito utilizadas na lingüística de *corpus* por permitirem que importantes padrões de uso da língua sejam descobertos ou compreendidos.

Concordanciadores são ferramentas muito utilizadas na lingüística de *corpus*, pois constroem concordâncias de forma automática: o usuário entra com o termo a ser pesquisado e dispara a busca, como ocorre com qualquer outra ferramenta de busca. Porém, diferentemente de uma busca normal, em que os resultados são os arquivos ou páginas que contêm o termo pesquisado, concordanciadores apresentam trechos dos textos em que o termo pesquisado aparece.

[47] http://www.chass.utoronto.ca/cch/index.html
[48] http://childes.psy.cmu.edu/clan/
[49] http://sourceforge.net/projects/corsis/

Nos resultados, o termo pesquisado é sempre apresentado de forma destacada e alinhada no centro dos resultados, dentro de uma janela de contexto. Geralmente é possível configurar o tamanho da janela de contexto, ou seja, especificar o número de palavras que devem ser apresentadas no resultado, antes e depois do termo pesquisado.

Essa forma de visualização facilita o entendimento de padrões de uso, pois permite que um lingüista detecte diferentes usos de uma mesma palavra. É possível descobrir, por exemplo, se o termo pesquisado costuma ser seguido ou precedido de palavras ou classes de palavras específicas ou, ainda, se o termo costuma ser usado no início ou no final de frases. Alguns concordanciadores mais sofisticados permitem que a busca seja feita com expressões regulares, possibilitando, por exemplo, encontrar palavras com prefixos ou sufixos específicos. Quando utilizados em *corpora* anotados, algum concordanciadores permitem que as busca sejam aplicadas, por exemplo, sobre classes de palavras específicas.

Algumas ferramentas que dispõem de concordanciadores são citadas a seguir: Corpógrafo (Seção 5.1), WordSmith (Seção 5.3), Unitex (Seção 5.5), WebCorp (Seção 5.6), AntConc[50], WebCONC[51], Corsis, MonoConc[52], GlossaNet[53] e CorpusEye[54]. Na Figura 6 é possível ver o concordanciador do Corsis.

[50] http://www.antlab.sci.waseda.ac.jp/antconc_index.html
[51] http://www.niederlandistik.fu-berlin.de/cgi-bin/web-conc.cgi
[52] http://www.athel.com/mono.html
[53] http://ling.fltr.ucl.ac.be/index.php
[54] http://corp.hum.sdu.dk/

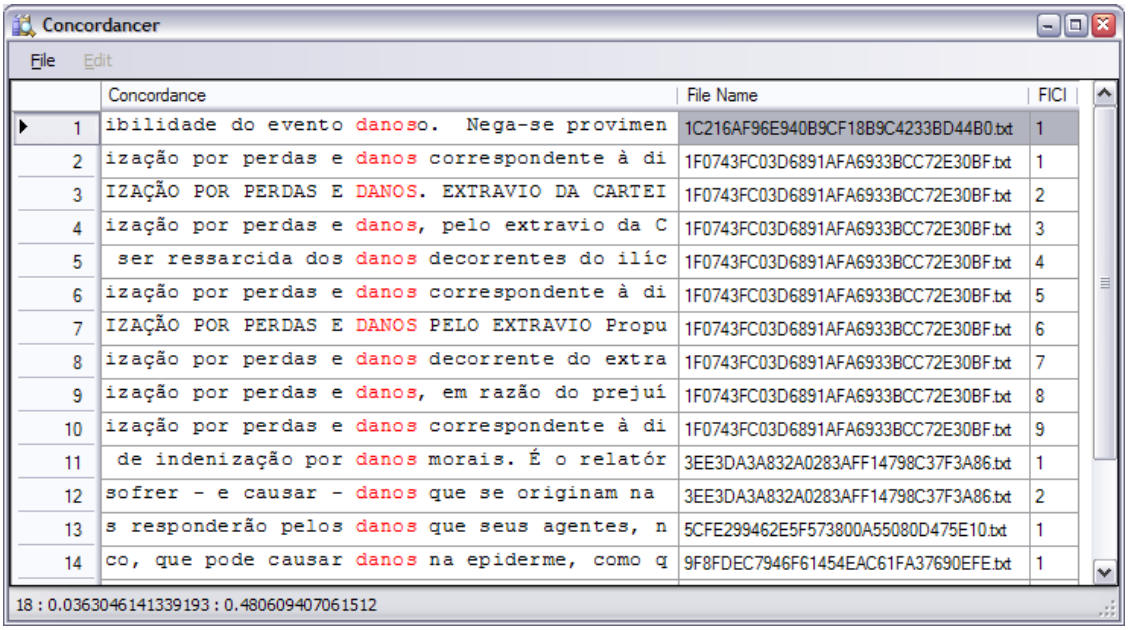


Figura 6 – Concordanciador do Corsis

4.1.3. Buscadores de colocações

Colocações (ALLEN, 1995) são expressões formadas por duas ou mais palavras que tendem a aparecer juntas – de forma consecutiva – ou próximas umas das outras, indicando combinações preferenciais ou usuais de palavras. Nessas combinações recorrentes, aparentemente livres, as colocações revelam maneiras comuns de organizar e posicionar as palavras em determinados contextos. Em colocações mais significativas, as palavras ocorrem com maior freqüência quando combinadas do que com outras palavras.

A identificação de colocações é importante, por exemplo, para a escrita de dicionários e o ensino de línguas e se dá através da observação das co-ocorrências. Gasperin e Strube de Lima (2001) descrevem algumas abordagens para encontrar colocações presentes em um texto, tais como: seleção das colocações por freqüência, seleção baseada em média e variância da distância entre a palavra foco e uma palavra vizinha, teste de hipótese e informação mútua.

O significado de uma colocação difere da simples combinação do significado das palavras que a compõem, pois apresenta um componente semântico que não poderia ser identificado se suas partes fossem observadas isoladamente.

4.2. Aplicações baseadas em *corpus*

Corpora são amplamente utilizados em diversas aplicações do processamento da linguagem natural. São exemplos de aplicações que podem ser baseadas em *corpus*: recuperação de informação, extração de informação, sistemas pergunta-resposta e sistemas de tradução automática (RUSSELL; NORVIG, 2003).

Recuperação de informação (RUSSELL; NORVIG, 2003) é a tarefa de encontrar documentos relevantes para as necessidades do usuário. Ferramentas de busca na Internet estão entre os melhores exemplos de sistemas de recuperação de informação. Na Internet o usuário pode pesquisar por uma palavra qualquer em uma ferramenta de busca e, em poucos segundos, ter uma lista de páginas relevantes. Um *corpus* pode servir como uma base de treino para este tipo de sistema.

Extração de informação (RUSSELL; NORVIG, 2003) é o processo de pesquisa em textos por ocorrências de uma classe particular de objeto ou evento e por relacionamentos entre estes objetos e eventos. Essas ocorrências oferecem aos lingüistas descrições da língua que antes não eram percebidas ou não podiam ser facilmente comprovadas.

Outra aplicação são sistemas pergunta-resposta (CALLISON-BURCH; OSBORNE, 2003), ou QA (*Question Answering*), um tipo de sistema de recuperação de informações inteligente que requer técnicas de PLN mais complexas. O principal objetivo deste tipo de sistema é recuperar, a partir de uma coleção de documentos, pequenas passagens de texto capazes de responder às perguntas formuladas em linguagem natural pelos usuários. Perguntas do tipo "Porquê?" ou "Como?", são mais complexas que perguntas do tipo "Quando?" ou "Onde?". Se o usuário, por exemplo, formular a pergunta "O que é um motor?", a resposta a ser procurada é do tipo "Um motor é X". Já perguntas do tipo "Quem?" e "Quando?", por exemplo, fazem com que, respectivamente, documentos do *corpus* com entidades nomeadas e datas, tornem-se candidatos a conter a resposta.

Outra importante aplicação em PLN com larga utilização de *corpora* é a tradução automática (MEGERDOOMIAN, 2003), ou *machine translation*, cujo objetivo é traduzir automaticamente textos de uma língua para outra. As principais abordagens para a tradução automática vão desde sistemas baseados

em conhecimento, com formalismos para representação de conhecimento, até abordagens puramente estatísticas. Atualmente os sistemas são uma combinação de módulos estatísticos e não estatísticos.

Especificamente para tradução automática são necessários *corpora* paralelos e alinhados. O objetivo do alinhamento é combinar sentenças, frases e palavras nos dois textos, fonte e alvo. Um *corpus* alinhado pode ser usado para criar dicionários bilíngües ou gramáticas paralelas. O alinhamento do texto não é uma tarefa simples, pois as traduções não refletem a estrutura original nem as mesmas palavras do texto traduzido. Os tradutores normalmente reorganizam o texto para obter um sentido melhor na língua alvo ou para converter expressões idiomáticas.

4.3. Considerações sobre este capítulo

Neste capítulo apresentamos três funcionalidades de exploração de *corpora*: contador de ocorrência, concordanciador e buscador de colocações. Apresentamos também, brevemente, algumas aplicações baseadas em *corpus*: recuperação da informação, extração da informação, sistemas pergunta-resposta e sistemas de tradução automática.

No próximo capítulo traremos informação sobre algumas ferramentas que realizam compilação e exploração de *corpus*, abordando seus principais recursos e funcionalidades.

Capítulo 5

Ferramentas para *corpora*

Neste capítulo apresentaremos ferramentas para *corpora* consideradas importantes, ou por serem muito citadas, ou por serem largamente utilizadas nas pesquisas em língua portuguesa. Observaremos os principais recursos e funcionalidades, assim como os principais aspectos positivos e negativos de algumas delas. São estudadas as seguintes ferramentas: Corpógrafo, Lácio-Web, Oxford WordSmith Tools, GATE, WebCorp, Unitex e Philologic. Destacam-se especialmente o Corpógrafo e o WordSmith, por serem entendidos como ferramentas similares à proposta no contexto desta dissertação.

5.1. Corpógrafo

O Corpógrafo[55] (MAIA; SARMENTO, 2003; SARMENTO; MAIA; SANTOS, 2004) é um ambiente integrado, em língua portuguesa, projetado para a lingüística de *corpus* e engenharia do conhecimento. Entende-se que o Corpógrafo se volta à engenharia do conhecimento por trabalhar com vocabulários e oferecer mecanismos para identificação de termos relevantes à compreensão dos domínios analisados. Nosso trabalho poderá, igualmente, ser entendido como uma ferramenta intermediária para aplicações de engenharia do conhecimento.

O sistema Corpógrafo provê, gratuitamente, diversas funcionalidades que permitem aos usuários compilar e explorar seus próprios *corpora*, mesmo sem muitos conhecimentos técnicos. O Corpógrafo é uma ferramenta *web*, portanto

[55] O Corpógrafo (http://www.linguateca.pt/corpografo/) foi desenvolvido pela Linguateca na Faculdade de Letras da Universidade do Porto (http://www.letras.up.pt/). Na época quando lançado, denominava-se Gestor de Corpora.

não requer a instalação de *software* específico, além de um *browser* na Internet, para ser usado.

O Corpógrafo facilita a compilação de *corpora* permitindo ao usuário a submissão e a extração de textos em diversos formatos. Os formatos de arquivo suportados são: PDF, HTML, DOC, PS, RTF e texto puro. Uma interface de edição permite que os textos submetidos possam ser limpos e segmentados em sentenças. Na Figura 7 é possível observar a tela para submissão de arquivos a partir do computador do usuário.

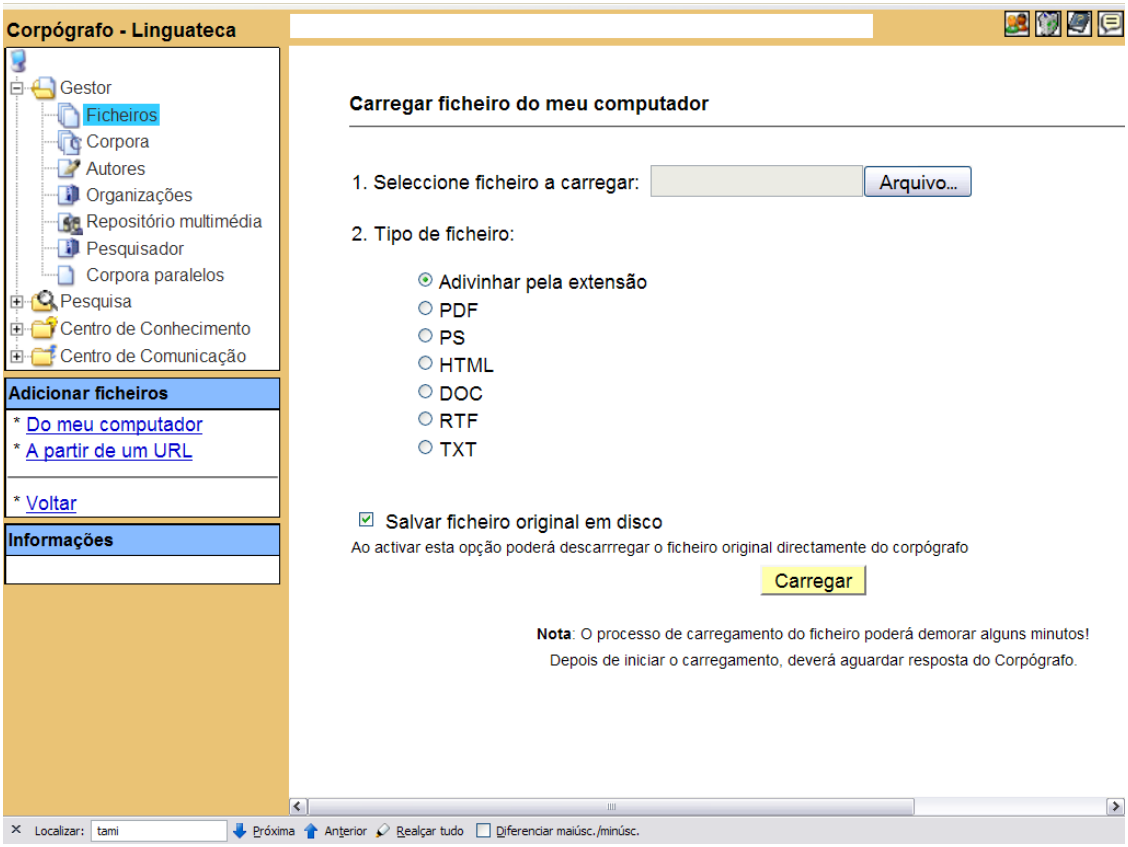


Figura 7 – Submissão de arquivos a partir do computador do usuário no Corpógrafo

O Corpógrafo permite também a submissão de textos através do *download* direto de documentos da Internet. Para submeter um texto, o usuário deve apenas indicar a URL do texto a ser adicionado. O usuário pode solicitar também que o sistema inspecione *links* dentro da página indicada, e filtre os tipos de documentos a serem adicionados. Na Figura 8 é possível ver a tela para submissão de arquivos a partir de URLs.

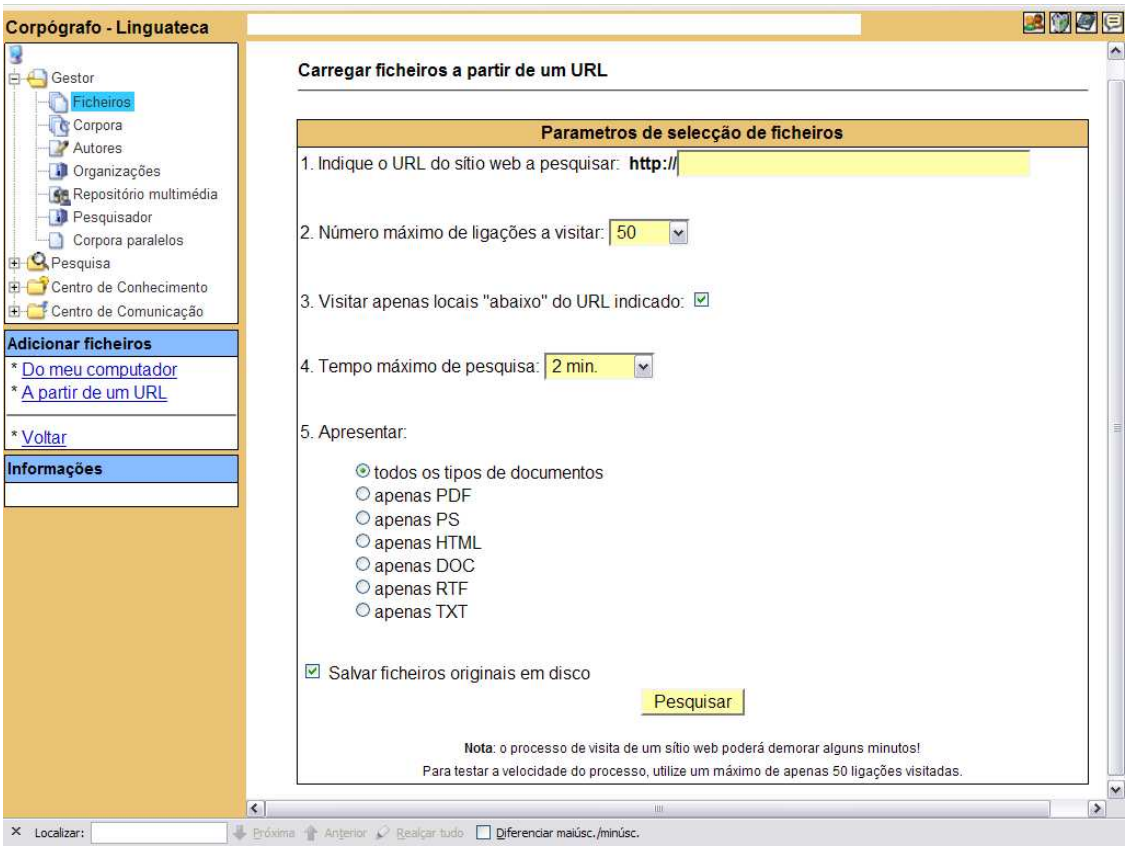


Figura 8 – Submissão de arquivos a partir de URLs no Corpógrafo

O usuário pode armazenar seus textos em um espaço privado no servidor, tendo também a opção de compartilhar seu trabalho com outros usuários cadastrados. A quota de espaço de armazenamento no servidor, para um usuário padrão, é de 10 *megabytes*.

É possível preencher e armazenar meta-dados específicos para cada um dos textos submetidos. Entre as informações que podem ser reunidas estão: título do documento, idioma, ano de publicação, autores, organização, área de domínio e gênero, entre outras. O preenchimento destas informações pode ajudar na organização e nas buscas sobre os *corpora*.

Os *corpora* compilados no sistema recebem o nome de *selection* e podem ser montados a partir dos textos submetidos e armazenados em seu espaço privado no servidor. Um *corpus* pode ser formado por vários textos, que por sua vez

podem integrar diferentes *corpora*. No entanto, o envio de grandes volumes de texto pode ser prejudicado pela velocidade de transmissão da Internet e limitações de espaço e de capacidade de processamento do servidor.

O Corpógrafo disponibiliza também funcionalidades para exploração de corpora: concordanciador com uso de expressões regulares, extrator de colocações, contagem de palavras e buscador de n-gramas. O sistema ainda permite a extração de informações terminológicas, relações semânticas e mapas conceituais.

5.2. Ferramentas do projeto Lácio-Web

O projeto Lácio-Web (ALUISIO et al., 2003), descrito na Seção 2.2.3, disponibiliza ferramentas lingüístico-computacionais aos usuários cadastrados. Dentre as ferramentas encontram-se concordanciadores, contadores de freqüência e etiquetadores morfossintáticos. Essas ferramentas, disponíveis após cadastro no site do projeto, são apresentadas a seguir:

- Contador de freqüência padrão: calcula a freqüência de ocorrência das palavras de um *corpus*, retornando os resultados em uma tabela de freqüência e informando o total de arquivos do *corpus* e a variação do vocabulário.

- Contador de freqüência por palavra: a partir do resultado ordenado de uma contagem de ocorrências de palavras em um *corpus*, a ferramenta busca a freqüência de uma palavra específica e um determinado número de palavras com freqüência maior e menor que a palavra escolhida.

- Concordanciador para *corpus* sem anotação: gera uma listagem de todas as ocorrências de uma palavra ou expressão.

- Concordanciador para *corpus* anotado morfossintaticamente: gera uma lista de ocorrências de uma determinada palavra, com determinada etiqueta.

- Etiquetadores morfossintáticos: A anotação morfossintática do *corpus* Lácio-Ref e seus *subcorpus* foi revisada manualmente e serviu de treinamento para três etiquetadores: MXPOST[56], TreeTagger[57] e Brill[58].

5.3. Oxford WordSmith Tools

O Oxford WordSmith Tools[59] (SCOTT, 2006; SARDINHA, 2004) é um conjunto integrado de ferramentas, bastante útil na preparação, manipulação, análise e descrição lingüística de um *corpus*. Entre os recursos disponíveis no WordSmith estão as ferramentas WordList, KeyWords e Concord.

Figura 9 – Tela principal do Oxford WordSmith Tools.

A ferramenta WordList permite, através da contagem de palavras, a criação de listas de palavras. As listas mostram a freqüência com que cada palavra foi encontrada nos textos e em quantos textos foi encontrada. Além

[56] http://www.inf.ed.ac.uk/resources/nlp/local_doc/MXPOST.html
[57] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
[58] http://www.cs.jhu.edu/~brill/
[59] O Oxford WordSmith Tools foi desenvolvido por Mike Scott e disponibilizado pela primeira vez pela Oxford University Press em 1996. Atualmente, encontra-se na versão 4.0. A licença da versão completa para um usuário custa aproximadamente €75. Uma versão de demonstração está disponível para *download* na página http://www.lexically.net/wordsmith/.

disso, é possível fazer a contagem usando lemas e organizar as listas por ordem alfabética ou por ordem de freqüência.

Com a KeyWords, é possível obter listas de palavras-chave de textos através da comparação da lista de freqüências de ocorrências de palavras dos textos em estudo com a lista de freqüências de ocorrências de palavras em um conjunto de textos de referência.

A ferramenta Concord (Figura 10) produz concordâncias em que todas as ocorrências de uma palavra ou de um conjunto de palavras são listadas. Na listagem, o contexto da palavra-chave é exibido, mostrando as palavras adjacentes à mesma. É possível produzir concordâncias diretamente a partir das ferramentas WordList e KeyWords, selecionando palavras das listagens geradas pelas duas ferramentas.

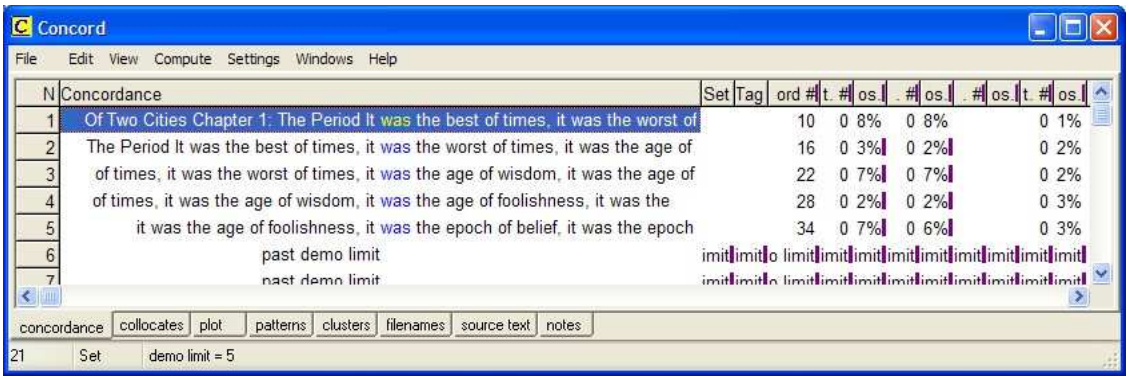


Figura 10 – Ferramenta Concord do Oxford WordSmith Tools

O Oxford WordSmith Tools conta ainda com outros pequenos programas bastante úteis na manipulação e visualização de arquivos de textos de um *corpus*. Entre eles, podemos destacar os utilitários Text Converter, Spliter e Viewer.

O utilitário Text Converter faz a busca e substituição de texto em arquivos. Ele é bastante útil quando um grande número de textos e arquivos necessita ser modificado. O Text Converter suporta conversões complexas, auxiliando quando muitos arquivos precisam ser reformatados.

O Splitter parte arquivos muito grandes em arquivos menores. Neste utilitário é possível escolher um símbolo ou termo para indicar os pontos onde

devem ser feitas as divisões, como por exemplo *"</text>"*. Cada termo ou símbolo de divisão encontrado no arquivo original irá gerar um novo arquivo texto.

O Viewer é um utilitário de visualização de arquivos texto que funciona de maneira integrada às demais ferramentas. Este utilitário permite que o usuário visualize o conteúdo de arquivos em vários formatos. O Viewer também produz saídas com sentenças ou parágrafos numerados, facilitando o alinhamento de duas versões de um texto.

5.4. GATE – General Architecture for Text Engineering

O GATE[60] (CUNNINGHAM et al., 2007), General Architecture for Text Engineering, provê uma infra-estrutura para o desenvolvimento de *software* de PLN. Desde seu lançamento o *GATE* tem sido usado por muitas organizações em diversos projetos de pesquisa e desenvolvimento.

O sistema oferece uma arquitetura (estrutura organizacional), um *framework* (biblioteca de classes) e um ambiente gráfico de teste e desenvolvimento. O ambiente gráfico pode ser visto na Figura 11 e permite que o *GATE* seja usado de forma completamente independente, mas não impede que o sistema seja integrado a outras aplicações.

A arquitetura do *GATE* define três tipos fundamentais de recursos ou componentes para a construção de sistemas de PLN: *Language Resources* (*LRs*), *Processing Resources* (*PRs*) e *Visual Resources* (*VRs*). Esses tipos assumem papéis completamente distintos dentro da arquitetura:

- *Language Resources* (*LRs*) são componentes que representam recursos lingüísticos como documentos, *corpora*, esquemas de anotação e ontologias.
- *Processing Resources* (*PRs*) são componentes que representam recursos de processamento fundamentalmente algorítmicos, como *parsers* e etiquetadores.
- *Visual Resources* (*VRs*) são componentes que representam recursos que oferecem funcionalidades de visualização e edição de dados.

[60] O GATE foi desenvolvido (1995) e é mantido pela University of Sheffield.

O *GATE* provê, na instalação padrão, um conjunto de recursos que podem ser acoplados a sua interface gráfica ou utilizados de forma integrada em outras aplicações. O conjunto de componentes integrados ao GATE é denominado *CREOLE* (*Collection of REusable Objects for Language Engineering*).

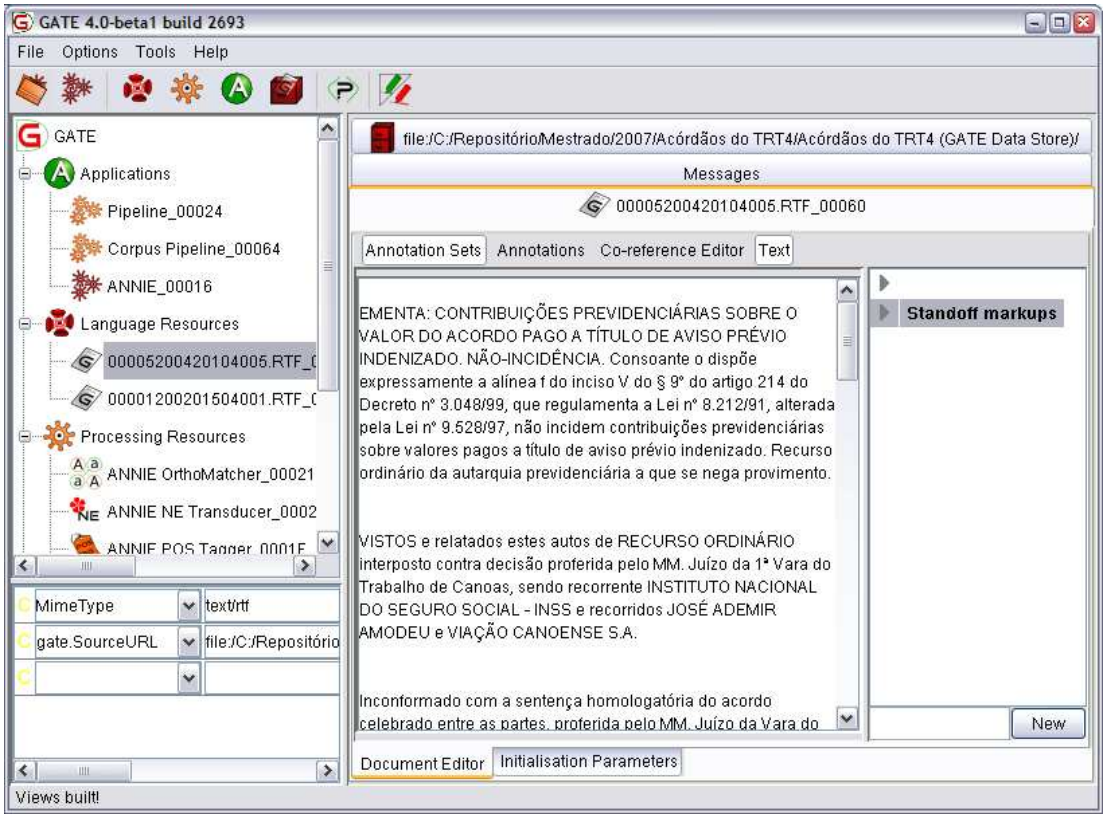


Figura 11 – Ambiente Gráfico do GATE

Além de sua arquitetura robusta, o *GATE* destaca-se também por possuir código fonte aberto e ter sido desenvolvido totalmente em Java, o que o torna independente de plataforma. Embora a persistência de seus recursos lingüísticos atenda ao padrão ISO TC37/SC4, o formato de codificação dos dados é próprio, o que faz com que o intercâmbio de recursos com outros sistemas exija a implementação de conversores.

Existe um grande número de componentes disponíveis para o *GATE*, a maioria deles para o tratamento de uma língua específica. Embora existam componentes disponíveis para diversas línguas, principalmente para a língua

inglesa, poucos são os recursos existentes específicos para o processamento da língua portuguesa.

5.5. Unitex

Unitex[61] (PAUMIER, 2006) é um conjunto de programas que possibilitam o tratamento de *corpora* utilizando recursos lingüísticos como gramáticas e dicionários eletrônicos DELA (Dictionnaires Électroniques du LADL). O Unitex surgiu como uma alternativa gratuita a outro sistema de processamento de *corpus*, o Intex[62].

Entre as funcionalidades disponíveis no sistema Unitex encontram-se: um gerador de concordâncias, um contador de freqüências, um gerenciador de dicionários DELA e um gerenciador de gramáticas. O Unitex permite também que o usuário faça buscas complexas utilizando expressões regulares.

Uma das características favoráveis do Unitex, é que ele utiliza a codificação de caracteres Unicode, suportando praticamente todos os caracteres de todos os idiomas. Outra característica positiva é o fato do Unitex poder ser livremente modificado e distribuído nos termos da licença LGPL (Lesser General Public License).

O Unitex, desenvolvido nas linguagens C e Java (J2SE 6), infelizmente não suporta o padrão XCES. Além disso, exige que os textos de um *corpus* estejam agrupados em único arquivo para que possam ser analisados, algo não usual.

5.6. Philologic

Philologic[63] é conjunto de ferramentas para processamento de *corpus*. A ferramenta suporta anotações TEI Lite (Text Encoding Initiative), usadas em buscas por critérios bibliográficos, tais como: título, autor e data de publicação. Além disso o Philologic pode ser configurado para processar outras versões do TEI ou outros padrões, como o XCES. O Philologic suporta a criação de *subcorpora* e dispõe de uma interface *web* que facilita sua utilização.

---

[61] A primeira versão do Unitex (http://www-igm.univ-mlv.fr/~unitex/), disponibilizada em 2002, foi desenvolvida por Sébastien Paumier do Institut d'Électronique et d'Informatique Gaspard-Monge (http://www-igm.univ-mlv.fr/), da universidade francesa Marne la Vallée (http://www.univ-mlv.fr/).

[62] http://intex.univ-fcomte.fr/

[63] Philologic (http://www.lib.uchicago.edu/efts/ARTFL/philologic/) foi desenvolvido no âmbito do projeto ARTFL (American and French Research on the Treasury of the French Language) da University of Chicago.

Para poder ser utilizado através de uma interface *web*, o Philologic requer a instalação de um servidor *web* e *software* adicionais em um ambiente Linux. Esses requisitos podem tornar a instalação complexa e de difícil execução para muitos usuários.

5.7. WebCorp

A WebCorp (KEHOE; RENOUF, 2002) é uma ferramenta de busca projetada para buscar e apresentar, de maneira adequada à análise lingüística, exemplos de uso de palavras em textos disponíveis na Web. A ferramenta refina as buscas na *Web* ao permitir o uso de caracteres curinga e a busca por padrões (realizando *pattern matching*), além de acrescentar outros recursos importantes para estudos lingüísticos.

WebCorp foi projetado para, internamente, buscar páginas relacionadas aos termos pesquisados pelo usuário fazendo uso de ferramentas de busca já existentes na *Web*. Feita essa busca, cada uma das páginas é acessada e analisada, e cada ocorrência do termo com seu respectivo contexto é extraída e apresentada ao usuário.

O formato de apresentação dos resultados é configurável. Entres as opções disponíveis ao usuário estão o formato de saída e o tamanho da janela de contexto do termo buscado. O tamanho da janela de contexto pode ser configurado entre uma e 50 palavras. Na Figura 12 é mostrado um exemplo de saída do WebCorp para o termo "*machine*", gerado com uma janela de 8 palavras para a esquerda e para a direita. As concordâncias foram geradas a partir de páginas publicadas na Internet.

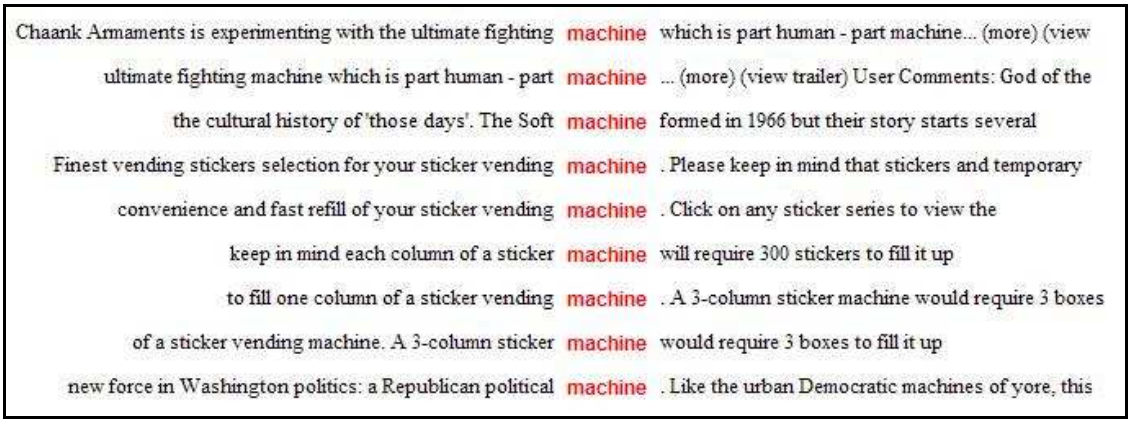


Figura 12 – Saída do WebCorp para o termo "machine"

5.8. Portal de Córpus

O Portal de Córpus (MUNIZ et al., 2007) é um portal para compilação, manutenção e disponibilização de *corpora* desenvolvido com recursos do projeto PLN-BR e FAROL[67]. O principal objetivo do projeto PLN-BR foi a construção e o compartilhamento de recursos e ferramentas lingüístico-computacionais entre sete importantes grupos de pesquisas brasileiros.

O Portal de Córpus abriga atualmente três *corpora* de textos jornalísticos extraídos do jornal Folha de São Paulo: o PLN-BR FULL, o PLN-BR CATEG e o PLN-BR GOLD. O *corpus* PLN-BR FULL contém 103.080 textos e conta com quase 30 milhões de *tokens*. Os outros dois *corpora* foram compilados a partir do primeiro. O *corpus* PLN-BR CATEG, compilado para a pesquisa e classificação de textos, contém 30 mil textos e quase 10 milhões de *tokens*, enquanto o PLN-BR GOLD contém 1.024 textos, 338.441 *tokens* e suas anotações lingüísticas adicionais.

O Portal de Córpus é inteiramente compatível com o padrão de codificação *XCES*, visto que as ferramentas do portal permitem o armazenamento e a recuperação de textos em conformidade com o formato. A

[67] FAROL (Fortalecimento e Integração das Competências do Processamento da Língua), financiado pela CAPES/PROCAD #0035050

estrutura de arquivos dos documentos do Portal é semelhante à estrutura adotada pelo American National Corpus (ANC), que também adere ao XCES.

A estrutura de arquivos para cada documento lógico existente num *corpus* do Portal é apresentada na Tabela 5. Essa estrutura faz com que as anotações sejam *stand-off*, separadas dos dados primários. Na estrutura, o arquivo de conteúdo, com os dados primários, é codificado em UTF-16 (16-bit Unicode Transformation Format), enquanto os demais arquivos são codificados em UTF-8 (8-bit Unicode Transformation Format).

Tabela 5 – Arquivos para cada documento no Portal

Sufixo no nome do arquivo	Tipo de arquivo / XML Schema	Descrição
`*.xces.xml`	`xcesHeader.xsd`	Cabeçalho do documento.
`*.txt`	Texto puro / *raw text*	Conteúdo do documento.
`*-s.xml`	`xcesAna.xsd`	Segmentação de sentenças.
`*-logical.xml`	`xcesAna.xsd`	Marcação lógica dos parágrafos.
`*.xml`	`xcesDoc.xsd`	Conteúdo e anotações em único arquivo.

O Portal de Córpus tem código-fonte aberto, bem como são abertas todas as tecnologias que ele utiliza. Além disso, a estrutura da base de dados e a documentação do Portal estão disponíveis gratuitamente, fazendo com que possa ser facilmente portado para outros servidores.

O Portal de Córpus é baseado numa arquitetura cliente-servidor (conforme a Figura 13). No lado do cliente, a interface do Portal pode ser acessada através de um navegador *web* qualquer, com suporte a *applets*[68]. O Portal foi desenvolvido utilizando as tecnologias Java 2 Platform Enterprise Edition (J2EE): Java Server Pages (JSP), Java Servlets e Java Standard Tag Library (JSTL). O lado servidor é composto por dois servidores: um servidor de aplicação J2EE, Apache Tomcat, que também funciona como servidor *web*; e um servidor de banco de dados, MySQL Server. Ambos são sistemas gratuitos, robustos, estáveis e amplamente usados pela comunidade de desenvolvedores, além de possuírem código-fonte aberto.

Para adicionar textos aos *corpora* do Portal, existe uma *applet* chamada Header Editor. O Header Editor, além de permitir a inserção de vários textos ao

[68] *Applets* são pequenos aplicativos escritos em Java, embutidos em páginas HTML, e executados na Java Virtual Machine (JVM) do *browser* da máquina cliente.

mesmo tempo, permite a edição e atualização dos cabeçalhos dos textos já inseridos.

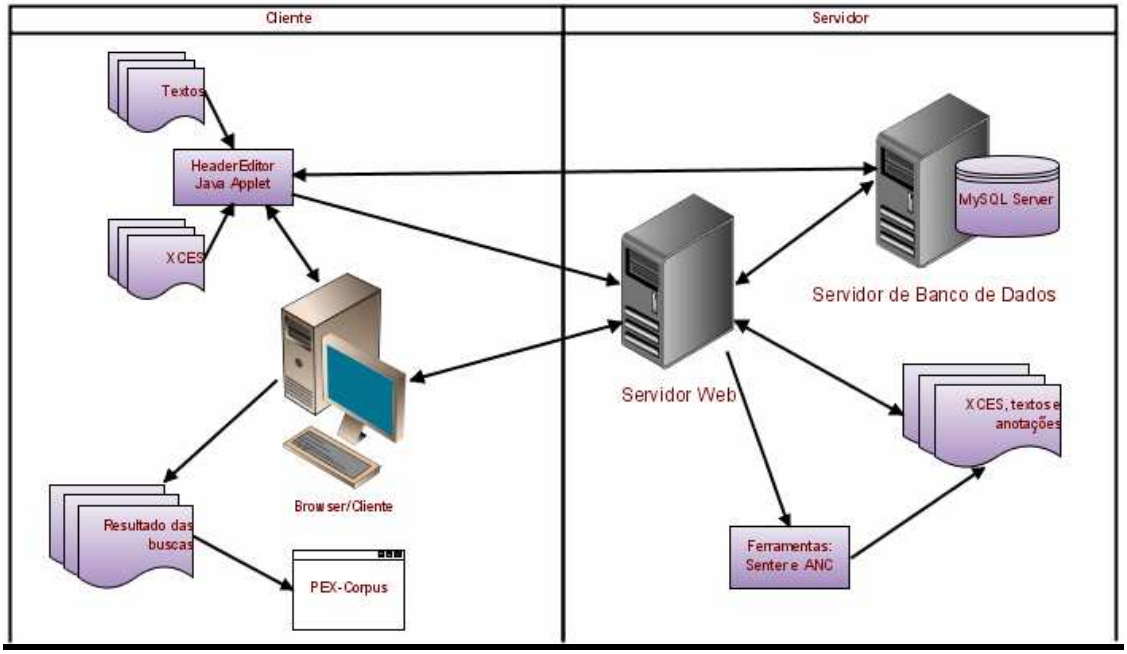


Figura 13 - Arquitetura cliente-servidor do Portal de Córpus

O Portal suporta o armazenamento de múltiplos *corpora*. Cada *corpus* é armazenado em uma base de dados diferente. Através de suas colunas e tabelas, essas bases de dados do portal mapeiam completamente o formato XCES e a estrutura de arquivos definida para os *corpora* armazenados no portal.

O Portal provê aos usuários poucas ferramentas para análise e exploração de um *corpus*. Entre as funcionalidades disponíveis podemos citar: a geração de *subcorpus* a partir de um *corpus* e a busca de textos baseada em informações armazenadas no cabeçalho dos documentos. O retorno das buscas e da geração de *subcorpus* é um arquivo compactado contendo os textos resultantes.

5.9. Considerações sobre este capítulo

Atualmente – apesar da maior disponibilidade de ferramentas para compilação, processamento e análise de *corpora* – muitos problemas ainda persistem. Muitas das ferramentas disponíveis são comerciais, dependem de plataformas específicas para serem executadas ou criam padrões próprios de

codificação de *corpus* e de anotações, dificultando a interoperabilidade e o compartilhamento de recursos lingüísticos entre aplicações. A seleção de ferramentas adequadas às necessidades de cada projeto representa hoje um desafio aos pesquisadores da área.

Outros problemas afetam especialmente os pesquisadores interessados no estudo da língua portuguesa. Além da falta de suporte a esquemas de codificação de caracteres adequados, enfrenta-se a escassez de recursos e de ferramentas específicos para nossa língua.

Ferramentas *web* como o Corpógrafo têm seu desempenho prejudicado pela velocidade de transmissão de grande volumes de texto, e limitações de espaço no servidor. Some-se a isso o fato de que a maioria das ferramentas exige grandes esforços para limpeza e conversão de dados, principalmente em etapas de pré-processamento. Esses esforços, muitas vezes realizados de forma totalmente manual, além de dificultar o uso das ferramentas por leigos, tornam mais onerosas as pesquisas com *corpus* e as deixam mais suscetíveis a erros.

Sendo assim, observamos a oportunidade de preencher algumas das lacunas deixadas por essas ferramentas e optamos por desenvolver uma nova, ferramenta apresentada no próximo capítulo.

Capítulo 6

A ferramenta Entrelinhas

Dadas as considerações apresentadas no capítulo anterior e considerando também a possibilidade de contribuir com o projeto PLN-BR através do Portal de Córpus, optamos por desenvolver e disponibilizar uma ferramenta para compilação e exploração de *corpora*. A ferramenta idealizada, que recebeu o nome de Entrelinhas, visa facilitar as atividades relacionadas à compilação de *corpora* e oferecer funcionalidades de exploração compatíveis com o padrão de codificação dos *corpora* disponibilizados pelo Portal de Córpus. Parte das funcionalidades é suportada através da integração de bibliotecas e ferramentas como GATE, Apache Lucene[69], Apache XMLBeans[70], ICU4J[71], JChardet[72] e Yahoo! Search Web Services[73]. A Figura 14 apresenta as bibliotecas internas e as relações de dependência com as bibliotecas integradas.

[69] http://lucene.apache.org/java/docs/
[70] http://xmlbeans.apache.org/
[71] http://icu-project.org/index.html
[72] http://jchardet.sourceforge.net/
[73] http://developer.yahoo.com/search/

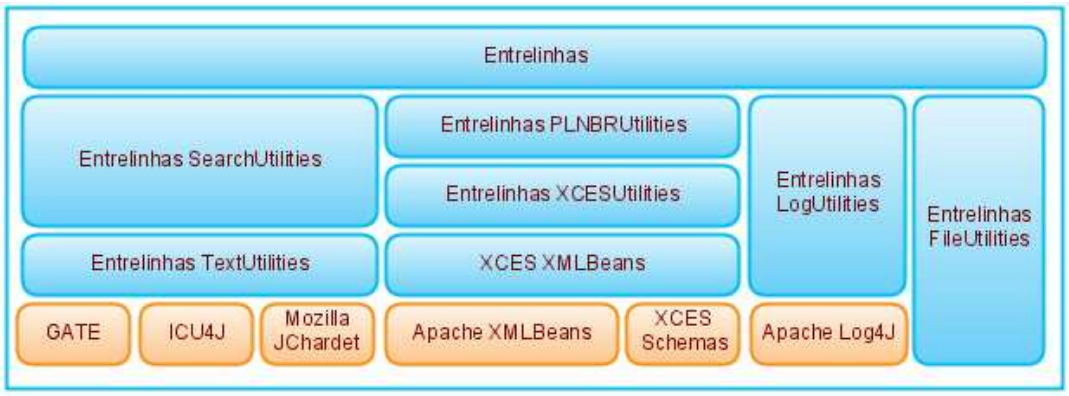


Figura 14 – Bibliotecas da ferramenta Entrelinhas

Entrelinhas foi implementada em Java, o que a torna independente de plataforma. Pode ser utilizada de duas formas: como uma aplicação independente, ou como uma biblioteca integrável a outras aplicações Java. Nas próximas seções descrevemos como um *corpus* é codificado e como as funcionalidades de compilação foram desenvolvidas e disponibilizadas dentro da ferramenta.

6.1. Codificação de um *corpus*

De maneira simplificada, poderíamos dizer que a ferramenta Entrelinhas compreende um *corpus* como um conjunto de arquivos XML gravados em um mesmo diretório. Esses arquivos devem poder ser validados conforme o tipo xcesDocType do *schema* xcesDoc do XCES (*revision* 0.4). Qualquer outro formato de arquivo não é considerado parte corpus, exceto se referenciado por um dos demais documentos.

A opção por utilizar um diretório como referência para o corpus em detrimento de um XML *schema* xcesCorpusType, também do esquema xcesDoc, é em razão do tipo não suportar *links* para xcesDocType. Essa característica obrigaria que todos os textos fossem armazenados em um único arquivo, fato que não consideramos desejável.

O tipo xcesDocType é utilizado para armazenar os dados primários com anotações de segmentação. O tipo suporta também anotações de cabeçalho (tipo xcesHeader do *schema* xcesHeader). No XCES, o header pode ser embutido dentro do xcesDocType ou em um arquivo separado, desde que seja referenciado no xcesDocType. Por praticidade, optou-se por manter o *header* dentro do documento. A Entrelinhas é capaz de ler e escrever três informações do *header*:

título do documento, nome do arquivo com a versão original do texto, nome do arquivo com a versão em texto puro.

Cada texto adicionado ao *corpus* recebe um código gerado a partir do *hash*[74] do texto original combinado com o *hash* da data e hora de inclusão. Essa combinação evita que dois documentos recebam o mesmo código quando textos idênticos forem deliberadamente adicionados. Muitos compiladores de *corpora* duplicam textos do *corpus* com a intenção de obter o balanceamento desejado.

Cada texto adicionado ao *corpus* gera um arquivo do tipo XML e outros dois arquivos: um no formato original, outro em texto puro. O código do texto é utilizado para a composição do nome destes arquivos conforme a Tabela 6. As diferenças em relação aos textos do Portal de Córpus não são percebidas pelo usuário, pois os textos são "convertidos" automaticamente na primeira leitura.

Tabela 6 – Nomes de arquivo em um *corpus* dentro da Entrelinhas

Nome do arquivo	Descrição
`<codigo>.xces.xml`	`xcesDocType` do *schema* `xcesDoc`
`<codigo>.txt`	Texto puro (UTF-16)
`<codigo>.original.<extensao original>`	Documento no formato original.

Sendo assim, supondo um *corpus* contendo dois textos, um em PDF e outro em RTF, com códigos 0A588D4BDA9EBC e 4A6C24A2ED7723, respectivamente, conteria em seu diretório seis arquivos:

```
0A588D4BDA9EBC.xces.xml
0A588D4BDA9EBC.original.PDF
0A588D4BDA9EBC.txt
4A6C24A2ED7723.xces.xml
4A6C24A2ED7723.original.RTF
4A6C24A2ED7723.txt
```

O arquivo xcesDocType e a versão em texto puro são codificados em *UTF-16* (*16-bit Unicode Transformation Format*). A adoção de esquema de codificação, que suporta adequadamente caracteres acentuados, se deve à preocupação em oferecer funcionalidades adequadas para estudos voltados para língua portuguesa. Os XMLs gerados pela Entrelinhas atendem ao requisito do nível 1[75] de conformidade do xcesDocType, que exige a segmentação dos parágrafos do texto. Um exemplo de documento do tipo xcesDocType gerado pelo XCES é apresentado no Apêndice A.

---

[74] Função matemática que permite obter uma representação única e relativamente curta de uma grande quantidade de dados.

[75] http://www.cs.vassar.edu/CES/CES1-4.html

6.2. Funcionalidades para a compilação de corpora

Entrelinhas oferece duas funcionalidades principais para a compilação de *corpora*: a compilação de *corpora* a partir de vários formatos de documentos e a compilação de corpora a partir de documentos disponíveis na Internet. As funcionalidades podem ser acessadas após a seleção da opção "Compilar corpus" ou "Editar corpus", na janela principal (Figura 15) e a seleção de um diretório adequado.

Figura 15 – Janela principal da Entrelinhas

Na janela de compilação de corpus (Figura 16), são oferecidas as opções para adicionar textos locais (detalhado na Seção 6.2.1), adicionar documentos da Internet (detalhado na Seção 6.2.2), editar ou remover um texto. A mesma janela permite visualizar em uma tabela o título, o código e a data e hora da inclusão ou última edição dos textos. Os textos da tabela podem ser ordenados em ordem crescente ou decrescente por qualquer um destes dados.

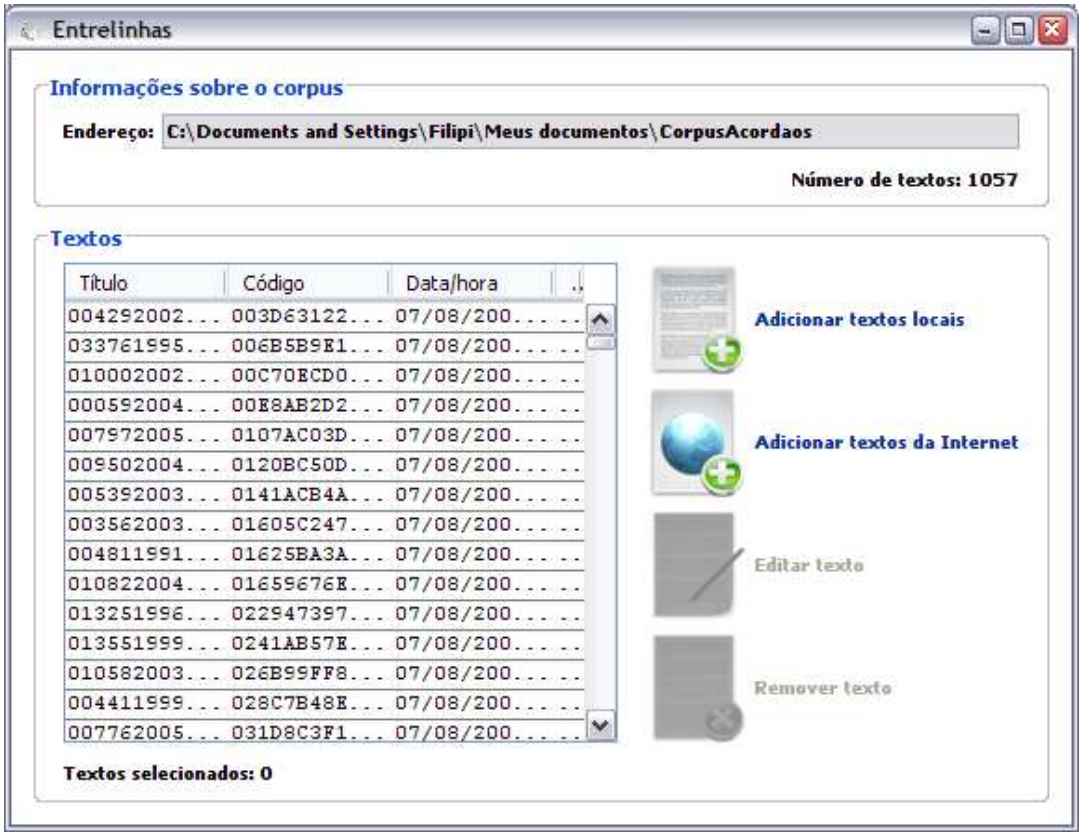


Figura 16 – Janela para compilação e edição de *corpus*

A janela de edição de textos (Figura 17) permite ao usuário, quando necessário, fazer a limpeza manual de um texto do *corpus* sem a necessidade de uma ferramenta externa.

A partir do editor também é possível abrir a versão original do documento. Essa funcionalidade visa facilitar a identificação de problemas durante a limpeza dos textos, pois em muitos casos o usuário precisa recorrer ao documento original para verificar alguma parte do texto. A abertura do documento original é condicionada a associação correta entre a extensão e o *software* adequado no sistema operacional do usuário.

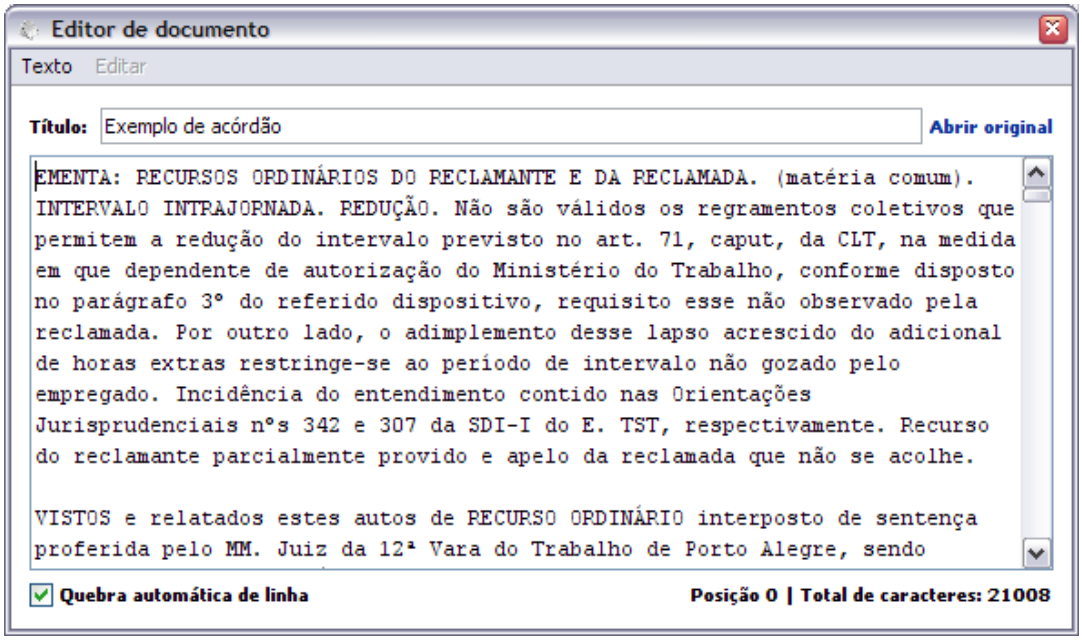


Figura 17 – Janela para edição de textos

6.2.1. Compilação a partir de vários formatos de documento

Atualmente, a compilação de *corpora* na maioria das ferramentas impõe restrições ao usuário, exigindo, por exemplo, que os arquivos contenham apenas texto puro. Essa restrição impede que o usuário utilize documentos nos formatos mais populares, sem que antes os arquivos sejam convertidos ou tenham seus textos extraídos manualmente.

A ferramenta Entrelinhas, através de componentes Language Resources do GATE, permite a compilação de um *corpus* a partir de vários tipos de arquivo. Estes arquivos são convertidos em texto puro e em seguida convertidos para o formato XML, compatível com o *schema* xcesDocType, do XCES. Os formatos suportados são:

- TXT (arquivos de texto puro, *plain text*)
- DOC (documentos do Microsoft Word)
- PDF (Portable Document Format)
- HTML (HyperText Markup Language)
- XML (Extensible Markup Language)
- RTF (Rich Text Format)
- EML (Electronic Mail)

O componente do GATE responsável pela conversão dos documentos exige como parâmetro de entrada o esquema de codificação do documento (*charset*). Para tal tarefa, a Entrelinhas utilizou duas bibliotecas *open source*, em Java, para detecção automática da codificação dos textos: Mozilla Chardet (utilizado pelo navegador *web* Mozilla FireFox[76]) e o ICU (International Components for Unicode, utilizado por diversos produtos da IBM[77]).

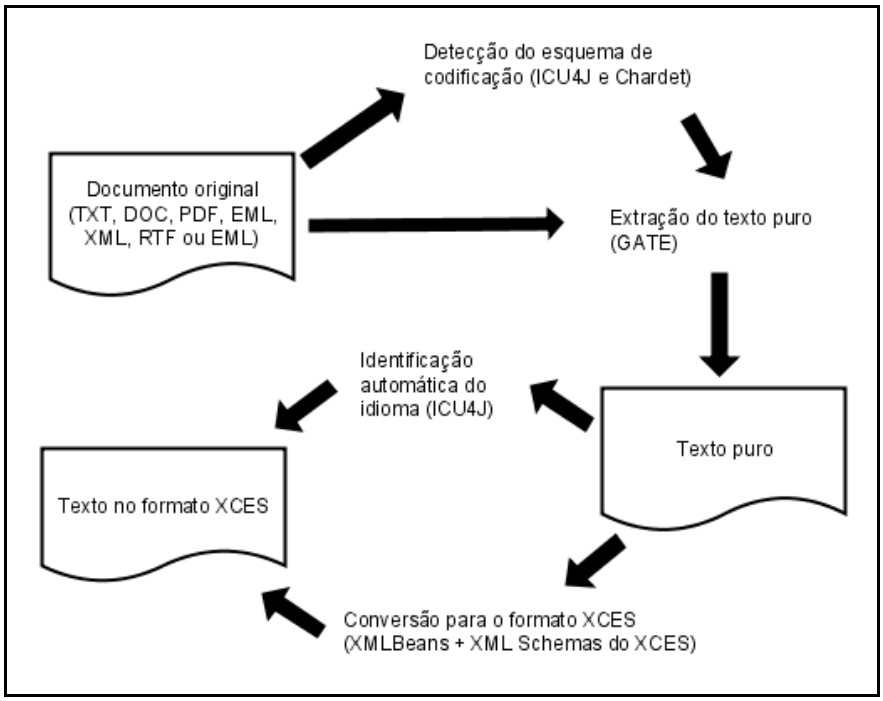


Figura 18 - Compilação de *corpora* a partir de vários formatos de documento

Alguns documentos do tipo DOC e do tipo PDF, dependendo da forma como foram criados e da aplicação que os gerou, podem não ser convertidos corretamente. Esses documentos podem ser editados ou removidos posteriormente dentro da Entrelinhas.

A conversão da versão em texto puro para o formato XCES é feita com auxílio da biblioteca Apache XMLBeans, que facilita a manipulação de arquivos XML. O uso dessa biblioteca foi fundamental devido à complexidade do formato XCES: são nove XML Schemas interligados, em que muitas marcações podem aparecer em *tags* com nomes diferentes. Essa característica se deve a tentativa do XCES de manter a compatibilidade com as especificações anteriores e algumas marcações do TEI.

[76] http://www.mozilla.com/firefox/
[77] http://www.ibm.com

Outra funcionalidade, disponível na biblioteca ICU e disponibilizada na ferramenta Entrelinhas, é a identificação automática do idioma dos textos. Essa funcionalidade permite que o idioma identificado seja informado no cabeçalho do documento.

6.2.2. Compilação a partir de documentos disponíveis na Internet

O grande número de documentos publicados na Internet torna esta rede uma fonte quase inesgotável de textos que podem ser utilizados como recursos lingüísticos. Através da biblioteca Yahoo! Search Web Services, a ferramenta Entrelinhas oferece ao usuário a opção de popular um *corpus* a partir de documentos na Internet.

Essa compilação de *corpora* a partir de documentos na Internet é feita através da janela mostrada na Figura 19. O usuário deve especificar os termos da pesquisa, o formato de documento e o idioma desejado. É possível especificar também o número máximo de resultados que devem ser retornados.

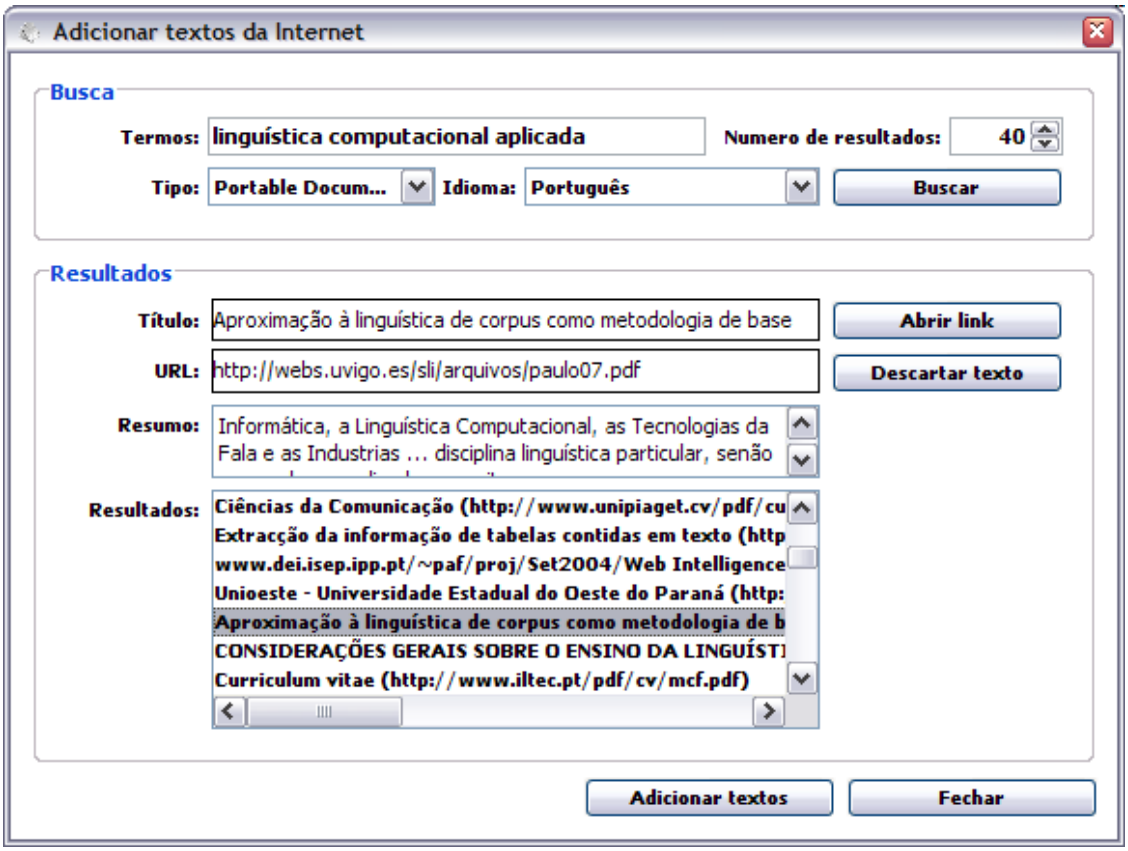


Figura 19 – Janela para a inclusão de textos da Internet

O usuário dispõe de quatro formatos de documentos na busca: DOC, HTML, PDF ou TXT. Além do português, é possível optar por oito outros idiomas: alemão, árabe, espanhol, francês, inglês, italiano, japonês ou russo.

Em muitas situações, a Internet não é adequada por não ser uma fonte confiável de documentos. Uma alternativa é tentar refinar os resultados dentro da Entrelinhas através de expressões que o Yahoo chama de *Meta Words*[78], presentes em várias ferramentas de busca na Internet. É possível, por exemplo, especificar um domínio para o qual a busca deve ser realizada, digitando "projeto de lei site:camara.gov.br" para pesquisar por "projeto de lei" no *site* da Câmara dos Deputados[79].

Após disparar a busca, a Entrelinhas envia os parâmetros informados para o serviço de busca do Yahoo através de sua biblioteca e, alguns segundos depois os resultados encontrados são apresentados. É possível clicar sobre um resultado e verificar a URL, o título da página, e um *snippet* com aproximadamente 20 palavras. O usuário pode optar por abrir para visualização e verificar qualquer um dos *links* retornados no botão "Abrir link". A URL será aberta no *browser* padrão.

Os resultados que não interessarem podem ser descartados através do botão "Descartar texto", que removerá o texto da lista. Ao finalizar a escolha dos textos o usuário deve clicar em "Adicionar textos" para fazer o *download* e adicionar os documentos presentes na lista ao *corpus*. Quando nenhum resultado interessar o usuário pode disparar uma nova busca com outros parâmetros.

6.3. Funcionalidades para exploração de *corpora*

Nas ferramentas estudadas, existem diversas funcionalidades que permitem ao usuário explorar um *corpus*. Nesse aspecto o Portal de Córpus ainda carece de ferramentas que proporcionem aos usuários a possibilidade de explorar e tirar melhor proveito dos *corpora* disponíveis no Portal.

Nesse sentido, duas funcionalidades de exploração de *corpora* foram disponibilizadas: um gerador de lista de palavras com contagem de ocorrências (Seção 6.3.1) e um concordanceador (Seção 6.3.2) com tamanho da janela de contexto configurável. Essas funcionalidades são acessíveis a partir da janela mostrada na Figura 20.

[78] http://help.yahoo.com/l/br/yahoo/ysearch/tips/tips-08.html
[79] http://www2.camara.gov.br/

Figura 20 - Janela exibindo funcionalidades de exploração de *corpora*

As funcionalidades citadas foram implementadas com auxílio da biblioteca Apache Lucene, um mecanismo de busca. Essa biblioteca foi escrita Java e é largamente utilizada na comunidade de desenvolvedores por possuir excelente desempenho.

A idéia inicial de utilizar a biblioteca Ngram Statistics Package[80] (NSP), escrita em Perl, na implementação das funcionalidades de exploração de corpora, foi descartada devido a dificuldades encontradas na integração das duas linguagens. Além disso, programas em Perl requerem a instalação adicional deum software interpretador denomiado ActivePerl[81] para serem executados, fato que poderia tornar mais complexa a instalação e distribuição da Entrelinhas. Existe uma iniciativa em andamento para criação de um *plugin*[82] do NSP para o GATE, porém os resultados disponibilizados ainda estão em versão beta.

O índice gerado pelo Lucene é gravado numa pasta denominada "index" dentro do diretório do *corpus* e é atualizado quando uma das ferramentas é acessada. Apesar de o Lucene, por padrão, remover as *stopwords*, optamos por configurar a indexação de modo que estas sejam também indexadas. A indexação utilizada diferencia palavras acentuadas de não acentuadas, mas não diferencia maiúsculas e minúsculas e segmenta palavras a cada ocorrência de caractere diferente de letra ou número. Por exemplo, a frase[84]:

```
O governo dos Estados Unidos confirmou nesta quarta-feira a realização no sábado de uma
reunião do G20, grupo atualmente presidido pelo Brasil.
```

[80] http://ngram.sourceforge.net/
[81] http://www.activestate.com/Products/activeperl
[82] http://sourceforge.net/projects/nspgate
[84] Extraída do site de notícias BBC Brasil.com (http://www.bbc.co.uk/portuguese/reporterbbc).

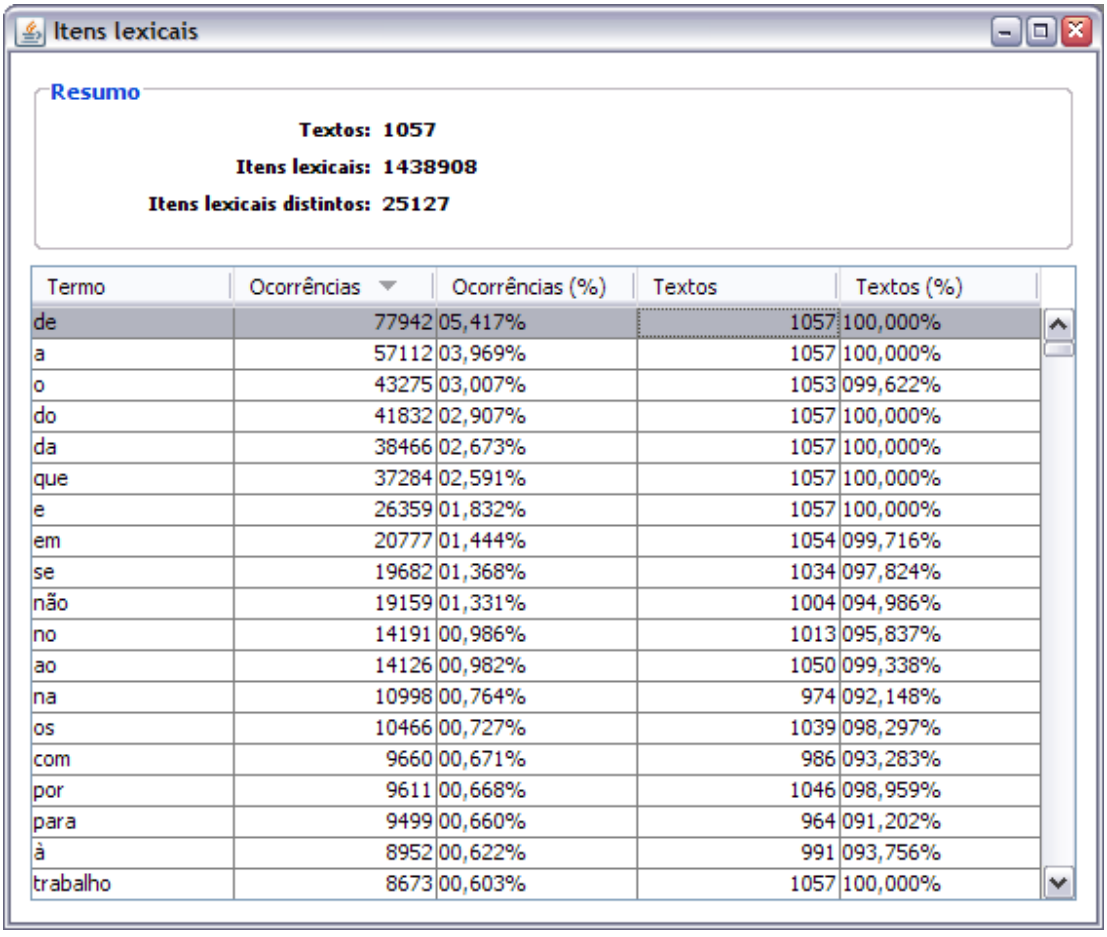
... será segmentada e indexada assim:

```
o|governo|dos|estados|unidos|confirmou|nesta|quarta|feira|a|realização|no|sábado|de|uma|
reunião|do|g20|grupo|atualmente|presidido|pelo|brasil
```

6.3.1. Lista de palavras

A lista de palavras (Figura 21) apresenta o número de ocorrências de cada item lexical armazenado no índice e sua freqüência em relação ao número total de itens lexicias. Além disso, é apresentado o número de textos em que cada item ocorre e sua freqüência em relação ao total de textos do *corpus*.

No topo da janela é exibido o total de textos do *corpus*, o total de itens lexicais e o total de itens lexicais distintos (que não se repetem). A lista pode ser ordenada em ordem alfabética, crescente ou decrescente por qualquer uma das colunas exibidas.



Termo	Ocorrências	Ocorrências (%)	Textos	Textos (%)
de	77942	05,417%	1057	100,000%
a	57112	03,969%	1057	100,000%
o	43275	03,007%	1053	099,622%
do	41832	02,907%	1057	100,000%
da	38466	02,673%	1057	100,000%
que	37284	02,591%	1057	100,000%
e	26359	01,832%	1057	100,000%
em	20777	01,444%	1054	099,716%
se	19682	01,368%	1034	097,824%
não	19159	01,331%	1004	094,986%
no	14191	00,986%	1013	095,837%
ao	14126	00,982%	1050	099,338%
na	10998	00,764%	974	092,148%
os	10466	00,727%	1039	098,297%
com	9660	00,671%	986	093,283%
por	9611	00,668%	1046	098,959%
para	9499	00,660%	964	091,202%
à	8952	00,622%	991	093,756%
trabalho	8673	00,603%	1057	100,000%

Figura 21 – Lista de palavras na Entrelinhas ordenada pelo número de ocorrências

6.3.2. Concordanciador

O concordanciador do Entrelinhas (Figura 22) apresenta as concordâncias existentes para um termo específico dentro de um *corpus*. O usuário pode escolher o tamanho do contexto que deseja visualizar para cada concordância, em caracteres. O termo pesquisado sempre aparece alinhado e em destaque nos resultados.

O Entrelinhas busca, através do Lucene, os textos em que os termos pesquisados ocorrem. Depois, utilizando funções nativas do Java, percorre cada um dos textos retornados, para extrair as concordâncias.

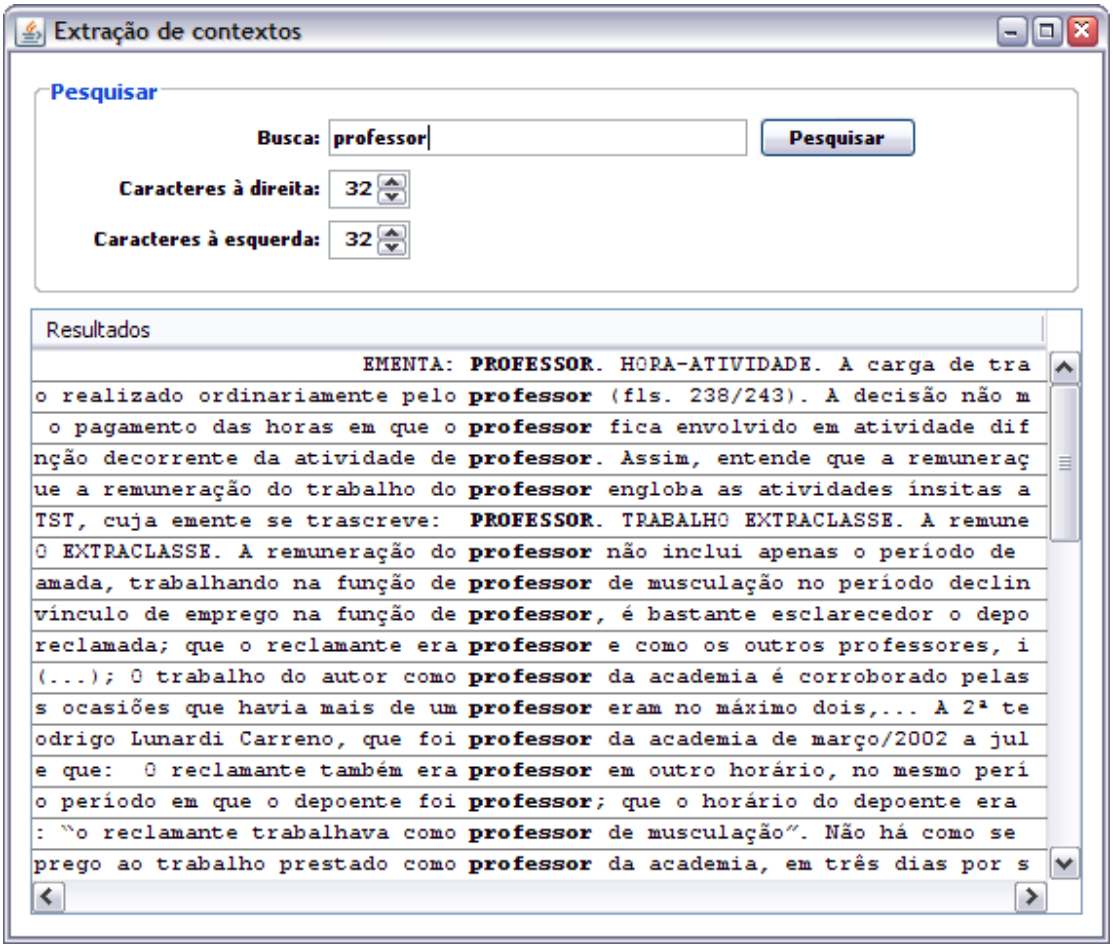


Figura 22 – Janela do concordanciador com concordâncias da palavra "professor"

6.4. Considerações sobre este capítulo

Neste capítulo apresentamos a Entrelinhas, uma ferramenta desenvolvida a partir de percepção de necessidades no âmbito da compilação e da exploração de *corpus* em língua portuguesa. Esta ferramenta a facilita a compilação e

disponibiliza funcionalidades essenciais para exploração de *corpora*. A ferramenta adere a um formato de codificação XCES, compatível com o Portal de Corpus, contribuindo com o intercâmbio de recursos no Projeto PLN-BR. No próximo capítulo, apresentaremos algumas experiências realizadas e mostraremos o relato de um usuário especializado quanto às funcionalidades do programa, em decorrência de suas percepções no uso da Entrelinhas.

Capítulo 7

Experiências de uso

Visando obter indicação da confiabilidade e da usabilidade do *software* desenvolvido, realizamos iniciativas de uso e acompanhamos os usuários através de relatos de uso. O objetivo dessas experiências foi conhecer a opinião do usuário especializado quanto às funcionalidades do programa. Para tanto identificamos um usuário com experiência em lingüística de *corpus* e habituado a utilizar ferramentas para compilação e exploração de *corpora*. Este usuário foi convidado a analisar se as funcionalidades apresentam os resultados esperados, verificar se o desempenho é aceitável e tentar revelar falhas ainda não descobertas. Por falta de tempo hábil, optamos neste momento por não incluir a integração com o Portal de Córpus nos experimentos.

O usuário especializado ficou livre para seguir seu próprio padrão de utilização de ferramentas dessa natureza, e definir a estratégia de como avaliar, os tipos de testes a serem aplicados, as técnicas e os critérios adotados. Para não influenciar no resultado, optou-se por não oferecer e não especificar uma coleção de textos a ser utilizada.

Os requisitos mínimos para o pleno funcionamento da Entrelinhas foram informados:

- computador conectado à Internet;
- Java Plataform, Standard Edition 6 instalado;
- Gate 4.0 (build 2752) instalado;

Não disponibilizamos ao usuário especializado ferramentas de teste automatizadas ou qualquer outro *software* além dos listados acima.

7.1. Primeira experiência de uso

Entrelinhas foi utilizada pela Profa. Susana de Azeredo[85]. Ela recebeu um breve treinamento e solicitou-se que avaliasse as funcionalidades de compilação e exploração de *corpus* da Entrelinhas com textos de sua preferência. Solicitamos que a avaliação fosse redigida na forma de um parecer relatando sua experiência, impressões, sugestões, dificuldades, problemas encontrados, pontos positivos e aspectos a serem melhorados. Ao longo dessa seção apresentaremos trechos do parecer, seguidos imediatamente pelos comentários do autor referentes ao trecho exposto. Trechos do parecer que se referiam a falhas na realização do experimento ou erros que foram imediatamente corrigidos antes da realização do segundo experimento, foram suprimidos desta seção. O parecer completo pode ser lido no Anexo A.

> *ENTRELINHAS: UM TESTE*
>
> *Susana de Azeredo*
>
> *O objetivo aqui é fazer um relato de um teste com o software ENTRELINHAS.*
>
> *O teste foi dividido em dois momentos. Em um primeiro momento, foi utilizado um corpus previamente montado de cerca de 100 mil palavras (todos os arquivos desse corpus eram de formato .txt). O objetivo neste primeiro momento foi testar a ferramenta de contagem de palavras e o concordanciador. Em um segundo momento, foi feito a montagem de um corpus através do ENTRELINHAS. O objetivo no segundo momento, foi testar a eficiência do Entrelinhas com relação à montagem do corpus.*
>
> *A seguir, detalhamos os dois momentos e as observações feitas.*
>
> *PRIMEIRO MOMENTO:*
>
> *Neste primeiro momento de teste, utilizamos um corpus previamente montado, o qual chamaremos aqui de corpus QUIM. Esse*

[85] Susana de Azeredo possui graduação em Letras (Bacharelado) Habilitação Português-Inglês pela Universidade Federal do Rio Grande do Sul (2004). Possui mestrado em Teorias do Texto pela mesma universidade. Tem experiência na área de Lingüística, com ênfase em Lingüística de Corpus e Terminologia, atuando principalmente nos seguintes temas: textos de química, corpus, lingüística de corpus, coesão e expressões anunciadoras de paráfrase. Está habituada a utilizar a ferramenta Oxford WordSmith Tools na compilação e exploração de *corpora*. Atualmente é professora substituta na Faculdade de Letras da UFRGS.

> *corpus compreende textos de Química, retirados de dois manuais de Química Geral. O corpus QUIM é composto de cerca de 100 mil palavras e os arquivos estavam todos em formato .txt.*
>
> *O objetivo aqui foi testar as ferramentas de contagem de palavras e o concordanciador.*
>
> *Com relação à contagem de palavras:*
>
> *c) A ferramenta de contagem de palavras revela o número total de palavras do corpus (tanto os itens lexicais quanto os itens lexicais distintos). Esse informação é muito relevante para um lingüista. No entanto, seria útil revelar também o número de palavras de cada texto. Essa é uma informação importante para o lingüista, pois é possível fazer uma comparação entre os textos do corpus, traçando diferentes perfis de textos. Além disso, não tem como selecionar apenas um dos textos do corpus para fazer uma análise específica ou uma comparação desse texto específico com o restante do corpus.*

Na versão atual, esse tipo de comparação pode ser feito se os textos a comparar forem compilados em corpora distintos; o mesmo procedimento vale para a análise da contagem de palavras em um texto específico. A apresentação da contagem de palavras, de forma individualizada para cada texto ou conjunto de textos, foi registrada como sugestão de melhoria futura.

> *d) Uma informação interessante para um lingüista e que não consta no Entrelinhas é a relação entre o número total de palavras do corpus e o número de palavras diferentes do corpus. Essa relação permite ver a variedade vocabular de um texto.*

A variedade vocabular pode ser calculada a partir das contagens apresentadas no topo da janela de contagem de palavras. A apresentação mais clara dessa informação foi registrada como sugestão de melhoria futura.

> *e) Com relação ao tempo que leva para finalizar a operação, pode-se dizer que foi bem eficiente. Há uma barra que indica que o Entrelinhas está fazendo o processamento estatístico, o que é bem útil. Uma única observação é que essa barra, às vezes, tranca e ficamos sem saber se o processo já terminou ou não.*

O problema foi identificado e corrigido.

> *f) As informações que aparecem nas colunas da janela de contagem de palavras são bem relevantes e úteis, principalmente, a coluna "textos". Essa coluna indica em quantos dos textos do corpus aparece a palavra.*

Clicando no cabeçalho das colunas, é possível ordenar os dados em ordem crescente ou decrescente.

> *Com relação ao concordanciador:*
>
> *a) No concordanciador, não aparece o total de vezes que aparece, no corpus, a palavra buscada. Para isso, é necessário voltar para a janela de contagem de palavras. Seria interessante aparecer junto com as concordâncias da palavra "tabela", por exemplo, quantas vezes ela aparece.*

O concordanciador retorna todos os resultados encontrados. A apresentação do total de concordâncias foi registrada como sugestão para melhoria futura.

> *b) Não aparece de que texto é cada concordância. Seria interessante colocar ao lado da concordância, o texto onde ela aparece.*

A identificação dos textos junto às concordâncias foi registrada como sugestão para melhoria futura.

> *c) Quando se abre o concordanciador, há uma formatação de 30 caracteres à direita e à esquerda da palavra. Quando se amplia esse horizonte para 50 ou mais, o Entrelinhas se perde, não mostrando mais a palavra buscada. Para algumas pesquisas com corpus, o horizonte mostrado ali é muito reduzido. Seria necessário que o usuário pudesse aumentar o horizonte para, no mínimo, ver a frase inteira. Quando se tenta buscar o horizonte da frase inteira, no Entrelinhas, ele se perde.*

A concordância pode não ser exibida corretamente quando o número de caracteres não cabe em uma única linha da tabela, por exceder o tamanho da janela. O usuário pode redimensionar a janela para aumentar o espaço disponível. Para facilitar a visualização do contexto, estuda-se a possibilidade de levar ao texto a partir da concordância, registrada como sugestão para melhoria futura.

> *d) Algo bem útil seria poder acessar o texto original a partir da concordância mostrada. Por exemplo, clicando na concordância, aparece o texto original com a palavra buscada onde ela se encontra no texto.*

> *Assim, é possível visualizar a frase inteira e até mesmo o parágrafo em que a palavra buscada aparece.*

A proposta foi identificada no item anterior e registrada como sugestão para melhoria futura.

> *SEGUNDO MOMENTO:*
>
> *O segundo momento envolveu a montagem de um corpus, utilizando os recursos do Entrelinhas. Esse corpus, que será chamado aqui de TESTE, contou com 23 arquivos. Entre esses arquivos, há textos de formato .doc, .rtf, .pdf, .txt e .html.*
>
> *Com relação à compilação do corpus, adicionando textos locais:*
>
> *a) Os arquivos .doc, .rtf e .txt foram incorporados ao corpus sem problemas maiores. A única coisa é que os gráficos, as tabelas e as figuras que constavam no arquivo foram transformados em símbolos. Isso exigiu uma limpeza dos textos antes que fossem processados no contador de palavras e no concordanciador.*

Entrelinhas utiliza os conversores embutidos no GATE. A limpeza dos textos é uma atividade muito comum na compilação de *corpus*, e o usuário pode realizá-la de dentro do Entrelinhas. Estuda-se a utilização de outros conversores em trabalho futuro, ou a aplicação de algoritmos que realizem parte da limpeza de maneira automática.

> *b) Os arquivos .pdf também foram incorporados sem maiores problemas. Eles tiveram uma incidência maior de símbolos do que os outros tipos de arquivo, exigindo uma tempo maior na limpeza dos textos. É ótimo que há a possibilidade de limpeza do texto dentro do Entrelinhas.*

Arquivos .PDF são bastante complexos porque podem ser gerados de diferentes formas por diferentes aplicações. Essa dificuldade pode ser observada até mesmo com um simples copiar e colar de parte de um texto em PDF, principalmente quando o texto inclui imagens ou fórmulas. Além disso, alguns documentos em PDF estão protegidos contra cópia e a conversão pode apresentar resultados inesperados.

> *c) Depois que os textos foram adicionados, todos os tipos de arquivo exigiram uma verificação da formatação do texto e da existência (e sua remoção) de símbolos.*

Conforme dito anteriormente, a limpeza dos textos faz parte da preparação dos textos e é uma atividade muito comum na compilação de *corpus* (ver Seção 3.3), e o usuário pode realizá-la dentro do Entrelinhas.

> *Com relação à compilação do corpus, adicionando textos da Internet:*
>
> *a) A primeira coisa que aparece na janela "Adicionar textos da Internet" é o campo "TERMO". Essa palavra não seria a mais adequada para uso. A definição de "TERMO" em Lingüística é muito controversa. Sugiro a utilização de "PALAVRA DE BUSCA" ou "PALAVRA-CHAVE".*

A sugestão resultou na alteração de "Termo" para "Palavra-chave".

> *b) A partir da colocação do "Termo", aparecem, no mínimo, 10 sites. O resumo que aparece no campo "resumo" é muito reduzido, tornando-se pouco confiável para que um texto possa ser selecionado e adicionado ao corpus. Seria muito bom que o botão "abrir link" estivesse ativo. Assim, seria possível verificar se o conteúdo da página é realmente útil para ser adicionado ao corpus.*

O número máximo de resultados apresentados pode ser escolhido pelo usuário na mesma tela. É possível inclusive retornar menos resultados através da entrada de mais palavras no campo de busca, restringindo a procura e obtendo resultados mais relevantes, como ocorre com qualquer outro mecanismo de busca na Internet.

O botão "Abrir *link*" pode não funcionar adequadamente quando o endereço retornado for muito extenso e apresentar muito parâmetros. Mesmo assim, nestes casos os documentos poderão ser adicionados ao corpus e o usuário poderá decidir mais tarde por mantê-los ou não.

O tamanho do resumo, cerca de 20 palavras, é uma limitação da biblioteca utilizada: Yahoo Search API. Essa limitação é idêntica ao sistema de busca do Yahoo disponível na web e semelhante à do Google.

> *c) Muitas vezes, o lingüista já tem um site de onde ele quer retirar textos para incorporar ao seu corpus ou até mesmo para montar seu próprio corpus. No Entrelinhas não é possível buscarmos um determinado site.*

Quanto à busca em site específico, ela pode ser feita através da utilização da expressão "site:" no campo de procura. Por exemplo, para procurar páginas relacionadas a "trabalho" em sites do governo, poderíamos digitar: "trabalho site:gov.br". Essas expressões são chamadas de *meta words* de busca e estão disponíveis na página de dicas do Yahoo[86].

> *d) Se os 10 primeiros sites retornados não são úteis, se faz uma nova busca. No entanto, os sites que aparecem continuam sendo os mesmos. Inclusive, se aumentarmos os números dos sites para serem mostrados (ao invés de 10, colocamos 20), o Entrelinhas retorna os mesmos primeiros 10 e mais outros 10 diferentes. Se aumentarmos para 30, aparecem sempre os mesmos primeiros 20 e mais 10 diferentes. Assim, ficamos sempre com os mesmos sites aparecendo.*

É esperado que buscas idênticas retornem resultados e posições idênticas. Estuda-se a possibilidade de paginar e navegar nos resultados para evitar esse comportamento inconveniente.

> *e) Não consigo selecionar apenas um dos textos da busca na Internet para adicionar ao corpus. Foi preciso adicionar todos ao corpus e só depois selecionar os necessários. Os arquivos .pdf e .doc que continham figuras e gráficos ou tabelas também precisaram de uma limpeza antes que fossem processados. No mais, o texto estava perfeito.*

Os textos podem ser removidos antes de serem adicionados ao *corpus* através do botão "Descartar texto". Registramos, para trabalhos futuros, o estudo de um maneira de melhorar a interface, deixando-a mais intuitiva.

> *f) Aplicam-se aqui as mesmas considerações feitas acima no primeiro momento da pesquisa com relação à contagem das palavras e ao concordanciador.*

Os mesmos comentários feitos anteriormente, relacionados ao uso de uma versão mais antiga, são válidos aqui.

> *ALGUMAS OBSERVAÇÕES EXTRAS E SUGESTÕES:*
>
> *a) Não é possível selecionar um diretório vazio para montar meu corpus. O diretório sempre precisa conter algum texto. Assim, foi necessário colocar um texto dentro do diretório para que o trabalho pudesse começar (mesmo que esse texto não fosse ser utilizado no meu corpus). Fica a pergunta: Esse primeiro texto que foi necessário colocar é*

[86] http://help.yahoo.com/l/br/yahoo/ysearch/tips/tips-08.html

> *contado como parte do meu corpus? Se sim, o resultado do número de palavras do corpus pode ser irreal, pois o Entrelinhas está contando um texto que eu não queria que estivesse ali.*

O problema foi identificado e solucionado. O texto que necessitou ser adicionado antes da compilação não interferiu na contagem nem no concordanciador. Para ser "visto" pela ferramenta ele deveria estar no formato XCES.

> *b) A barra "procurando concordâncias" tranca seguidamente.*

A janela de *status* fecha automaticamente após finalizada a operação. Em algumas situações, a ferramenta não conseguia fechar janela ao terminar e era preciso clicar no botão de fechar para prosseguir o uso normal, sem nenhuma concordância perdida. O problema no encerramento foi identificado e solucionado.

> *c) Na janela de compilação do corpus, os textos aparecem em uma determinada seqüência. No entanto, eles saem dessa ordem se selecionamos algum para editá-lo. Em que ficar arrumando sempre que se volta para essa janela. Fica confuso.*

Assim como na lista de palavras, nessa janela de compilação e edição de *corpus* também é possível ordenar os textos clicando no cabeçalho das colunas. É possível ordenar os textos pelo código, pelo título, ou pela data da última modificação, no entanto após uma edição a ordenação pelo código é sempre restaurada porque as demais colunas podem ter sido alteradas na edição.

> *d) Uma sugestão com relação ao concordanciador. Algumas pesquisas lingüísticas focalizam uma palavra e suas derivações. Por exemplo, uma pesquisa que focaliza o verbo PODER, talvez queira observar as derivações PODERÁ, PODERIA, PODEREMOS, PODE, etc. Para isso, se coloca o radical POD e um asterisco ao lado (POD*). Essa forma retorna o verbo PODER e seus derivados. O Entrelinhas não retorna essa informação. Tentei fazer isso com o verbo PODER (POD*) e com advérbios terminados em –mente (*MENTE), mas não obtive resposta.*

O uso de expressões regulares ainda não é suportado. Registramos a sugestão para melhoria futura.

> *e) O Entrelinhas faz uma separação entre palavras no singular e plural. Se eu coloco a palavra "TABELA" no singular, aparece apenas a palavra no singular. O mesmo ocorre com o plural. Isso é muito bom.*

O concordanciador e a contagem de palavras diferenciam palavras acentuadas e não acentuadas. As palavras "para" (preposição), "pára" (verbo) ou "Pará" (o Estado), por exemplo, apresentarão resultados diferentes.

7.2. Segunda experiência de uso

Para verificar se os problemas encontrados anteriormente na contagem de palavras repetiam-se na versão mais recente da Entrelinhas, solicitou-se à Susana Azeredo uma nova utilização. A avaliadora complementou o parecer anterior com o relato a seguir, acompanhado dos comentários do autor:

> *1) AS PALAVRAS "GRÁFICO" E "TABELA" QUE NÃO APARECIAM NA LISTA DE CONTAGEM: Eu selecionei novamente o corpus QUIM e, realmente, agora as duas palavras apareceram. Talvez tenha dado algum problema na minha máquina quando fiz a exploração do corpus anteriormente.*
>
> *2) INTERFACE: A Interface do Entrelinhas é muito elegante, agradável e de muito bom gosto. Também, não há uma poluição visual que confunde o usuário.*
>
> *3) USABILIDADE: A usabilidade do Entrelinhas é boa. No início, o Filipe deu algumas dicas para o primeiro acesso. Mas, acredito ser possível o acesso sem preparação antecipada sobre o programa. Penso ser interessante que o "botão" "explorar corpus" possa aparecer na janela de "compilação do corpus", o que tornaria a atividade que está sendo realizada mais dinâmica. A escolha do diretório é bastante recorrente e, em alguns casos, parece desnecessária.*

Registramos a observação e estudaremos uma maneira de tornar a interface mais fácil de ser utilizada.

> *4) INSTALAÇÃO: A instalação do Entrelinhas foi fácil e rápida. Também, não foi necessário instalar outros programas para que o Entrelinhas pudesse ser utilizado.*

Os programas necessários (Java 6 e Gate) já haviam sido instalados no experimento anterior.

7.3. Terceira experiência de uso

Documentos jurídicos são em geral muito extensos e pouco estruturados. A área jurídica produz um grande volume de textos e carece de ferramentas específicas para processá-los. Esses documentos – redigidos por Juízes, Desembargadores, advogados e seus assessores – são freqüentemente consultados pelos mesmos, na busca de jurisprudência que fundamente suas decisões ou sustente seus apelos.

Nesse sentido, obtivemos junto ao Egrégio Tribunal Regional do Trabalho da 4ª Região decisões judiciais publicadas entre junho de 1993 e abril de 2007. Os documentos, redigidos por Desembargadores e disponibilizados na Internet, são acórdãos e representam decisões colegiadas na segunda instância da Justiça do Trabalho. Os acórdãos, em formato RTF, totalizam aproximadamente 500 mil documentos e 13 *gigabytes* de dados e foram gravados em 5 DVDs.

Selecionamos uma amostra destes documentos para o terceiro estudo de caso, realizado pelo próprio autor. Esse terceiro uso envolve 1057 acórdãos publicados entre 1º e 7 de dezembro de 2005, e serviu para obtenção de dados acerca da performance da ferramenta. Um resumo da coleção é apresentado na Tabela 7.

Tabela 7 – Resumo da coleção de documentos para o terceiro estudo de caso

Número de textos:	1057
Formato dos arquivos:	RTF
Tamanho total:	39,4 MB
Tamanho médio dos arquivos:	38,16 KB

Os documentos foram compilados em um *corpus* na ferramenta Entrelinhas. O volume de dados gerado e os tempos de compilação e indexação foram registrados (Tabela 8).

Tabela 8 – Volume de dados gerado e tempos registrados

Tempo de compilação dos textos:	12 minutos e 44 segundos
Tempo de indexação:	35 segundos
Velocidade da compilação:	83,01 docs/min. ou 3,09 MB/min.
Tamanho total do corpus (indexado):	83,1 MB
Tamanho do índice:	3,14 MB

Coletamos os tempos de processamento da lista de palavras e duas buscas no concordanciador. A lista de palavras foi gerada em quatros segundos, e contabilizou o total de 1.438.908 itens lexicais, e 25.127 itens lexicais distintos. Os 50 itens mais freqüentes encontram-se no Apêndice B. O tempo das buscas e o número de resultados retornados no concordanciador são apresentados na Tabela 9.

Tabela 9 – Resultados e tempos registrados no concordanciador

	Concordâncias para "inconstitucional"	Concordâncias para "danos morais"
Concordâncias encontradas:	14	29
Tempo de resposta:	4 segundos	14 segundos

7.4. Considerações sobre este capítulo

Além da utilização por Susana Azeredo, as mestrandas Lilian Figueiró Teixeira[87] e Josiane Fountoura Brandolt[88], em estágio mais incipiente do trabalho, contribuiram efetuando os primeiros relatos de uso que permitiram corrigir falhas importantes e identificar os desejos de potenciais usuários da ferramenta (SILVEIRA, STRUBE DE LIMA, no prelo 2008). As sugestões coletadas proporcionaram uma visão do que ainda precisa ser feito e poderão guiar trabalhos futuros. Nas considerações finais, retomaremos os assuntos que foram abordados ao longo do trabalho e lições aprendidas, e indicaremos trabalhos fututos sobre a ferramenta Entrelinhas e sobre o tema estudado.

[87] Mestranda em Lingüística Aplicada na Universidade do Vale do Rio dos Sinos (UNISINOS) e graduada em Letras Licenciatura - Português e Inglês pela mesma universidade (2006). Tem experiência na área de Lingüística, com ênfase em Lingüística Computacional, atuando principalmente nos seguintes temas: ontologia, semântica lexical, compostos nominais, lingüística de corpus e ensino de língua estrangeira.

[88] Possui graduação em Informática (2000) e especialização em Educação (2006) pela Universidade da Região da Campanha. Atualmente é mestranda em Ciência da Computação pela Pontifícia Universidade Católica do Rio Grande do Sul, bolsista Capes. Tem interesse em Processamento da Linguagem Natural, Ontologias, Web Semântica e Informática na Educação.

Capítulo 8

Considerações finais

Neste trabalho apresentamos um estudo sobre a compilação e a exploração de *corpora*. Iniciamos apresentando o conceito de *corpus* e em seguida introduzimos as noções de tamanho e representatividade de um *corpus*. Vimos que essas características são bastante subjetivas e dependem muito do contexto em que se inserem. Em seguida, apresentamos o modo como *corpora* costumam ser classificados quanto ao propósito para o qual foram criados, o gênero, as áreas de domínio, o idioma, e a época em que os textos foram produzidos. Percebemos que a taxionomia de um *corpus* não é um assunto consolidado e que a mesma classificação pode aparecer sob diferentes nomes.

Fizemos referência a alguns dos maiores *corpora* em língua inglesa e língua portuguesa hoje existentes e abordamos aspectos sobre uso da *web* como um *corpus*. Notamos o crescimento no número de instituições que disponibilizam *corpora* cada vez maiores e mais ricos em anotações. É notável o potencial de uso da *web* como um *corpus*, revelado no crescente número de publicações e projetos sobre o tema.

Abordamos as principais etapas envolvidas na compilação de um *corpus*: o projeto, a coleta dos documentos, a preparação dos textos e a codificação do *corpus*. Discutimos questões importantes sobre critérios relacionados ao tamanho, ao balanceamento e à representatividade dos textos no projeto de um corpus. Descrevemos o modo como tarefas tais como conversão de formatos, limpeza de textos, uso de meta-dados, inserção de anotações e codificação dos textos impactam nos projetos e são afetadas pela disponibilidade de ferramentas adequadas.

Apresentamos brevemente algumas aplicações do uso de um *corpus* e funcionalidades utilizadas pelos lingüistas para a exploração de *corpora*. Descrevemos ferramentas para *corpora* que dispõem de parte dessas funcionalidades, apresentando suas principais características.

Observamos alguns problemas em aberto nessas ferramentas, como a dificuldade em criar um *corpus* com documentos em formatos que estamos habituados a utilizar, e o intercâmbio de recursos. Assim, enxergamos a possibilidade de contribuir com o Portal de Córpus através de uma nova ferramenta, a Entrelinhas.

Apresentamos as funcionalidades da Entrelinhas, que utiliza uma codificação compatível com textos do Portal. Relatamos experiências de uso da Entrelinhas e tivemos a oportunidade de receber o retorno de uma usuária qualificada e experiente no uso desse tipo de ferramenta. Os relatos confirmaram, a nosso ver, a contribuição esperada com a ferramenta e nos permitiram que coletássemos diferentes sugestões de melhorias.

Entre estas melhorias, destacamos: a paginação dos resultados da busca de textos na Internet, seleção de textos para a lista de palavras, comparação de listas de palavras, suporte ao uso de expressões regulares no concordanciador, acesso aos textos completos a partir do concordanciador e alterções na interface no intuito de deixá-la mais intuitiva e mais fácil de usar. Sabemos que, além das melhorias apontadas, há muitos outros pontos em que a Entrelinhas precisa evoluir, avançando rumo a uma maior integração com o Portal de Córpus e oferecendo mais funcionalidades.

Lembramos que este trabalho resultou, além da própria ferramenta Entrelinhas[89], na apresentação de um pôster no VI Encontro de Linguítica de Corpus[90], ocorrido em setembro de 2007, em São Paulo, e a publicação de um artigo nos anais[91] do mesmo evento e em livro (SILVEIRA, STRUBE DE LIMA, no prelo 2008).

Consideramos importante também a possibilidade de explorar e melhor aproveitar em trabalhos futuros os textos disponibilizados pelo Tribunal Regional do Trabalho 4ª Região, permitindo novas publicações na área.

---

[89] A ferramenta está disponível em http://sites.google.com/site/entrelinhaspln/
[90] http://www.nilc.icmc.usp.br/EncontroCorpora/index.htm
[91] http://www.nilc.icmc.usp.br/viencontro/Anais

Referências bibliográficas

ALLEN, James. **Natural language understanding**. 2. ed. Menlo Park, CA: Benjamin/Cummings, 1995. 654 p.

ALUISIO, Sandra M.; PINHEIRO, Gisele; FINGER, Marcelo; NUNES, Maria das Graças V.; TAGNIN, Stella E. The Lacio-Web Project: overview and issues in brazilian portuguese corpora creation. In: CORPUS LINGUISTICS 2003, 2003, Lancaster, UK. **Proceedings of the Corpus Linguistics 2003 Conference**: UCREL technical paper number 16. UCREL, Lancaster, UK: Lancaster University, 2003. v. 16, pp. 14-21.

ARNOLD, Douglas; BUCKLEY, Justin. **Corpus linguistics**. London: W3Corpora / IGE Project. 1998. Disponível em: <http://www.essex.ac.uk/linguistics/clmt/w3c/corpus_ling/content/index.html>. Acesso em: jun. 2007.

ASTON, Guy. **The british national Corpus as a language learner resource.** In: CONFERENCE ON TEACHING AND LANGUAGE CORPORA, 2., 1996, Lancaster. Proceedings. Lancaster: Lancaster University, UK, 1996. pp. 178-191.

BICK, Eckhard. **The parsing system palavras**: automatic grammatical analysis of portuguese in a constraint grammar framework. 2000. Tese (Doutorado) - Aarhus University Press, Aarhus, 2000.

BUITELAAR, Paul; DECLERCK, Thierry; RAILEANU, Diana et al. A multi-layered, XML-based approach to the integration of linguistic and semantic annotations. In: EACL 2003 WORKSHOP ON LANGUAGE TECHNOLOGY AND THE SEMANTIC WEB (NLPXML'03), 2003, Budapeste. **Proceedings of EACL 2003 Workshop on Language Technology and the Semantic Web (NLPXML'03**). Cunningham: EACL, 2003. Disponível em: <http://www.dfki.de/dfkibib/publications/docs/eacl03-xmlnlp.ps>. Acesso em: jun. 2008.

BURNAGE, Gavin; DUNLOP, Dominic. Encoding the British National Corpus. In: AARTS J.; HAAN, P. de; OOSTDIK, N. (Ed.). **English language corpora**: design, analysis and exploitation. Nijmegen, Netherlands: Rodopi, 1992. pp. 79-95.

CALLISON-BURCH, Chris; OSBORNE, Miles. Statistical natural language processing. In: FARGHALY, Ali Ahmed Sabry (Ed.). **Handbook for language engineers**. Standford : CSLI, 2003. pp. 269-297.

CHKLOVSKI, Timothy; PANTEL, Patrick. VerbOcean: mining the web for fine-grained semantic verb relations. In: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, 2004, Barcelona. **Proceedings of Conference on Empirical Methods in Natural Language Processing**. Barcelona, SSIGDAT, 2004. pp. 33-40.

CHOMSKY, Noam. **Syntactic structures**. The Hague: Mouton & Co., 1957.

CRPC: Corpus de Referência do Português Contemporâneo. Lisboa: Centro Lingüístico da Universidade de Lisboa, 2006. Disponível em: <http://www.clul.ul.pt/sectores/linguistica_de_corpus/projecto_crpc.php> Acesso em: junho de 2008.

CUNNINGHAM, Hamish et al. **Developing language processing components with GATE version 4 (a user guide)**. Sheffield: Gate, 2007. Disponível em: <http://gate.ac.uk/sale/tao/index.html>. Acesso em: jul. 2007.

EVANS, David. **Information about Corpus building and investigation**: aon-line information pack about corpus investigation techniques for the Humanities. Birmingham: Centre for Corpus Research/University of Birmingham, 2008. Disponível em: <http://www.humcorp.bham.ac.uk/humcorp/information/corpusintro.html>. Acesso em: jul. 2008.

FRANCIS, W. Nelson; KUČERA, Henry. **Brown Corpus manual**: manual of information to accompany a standard Corpus of present-day edited american english, for use with digital computers. Ed. rev. Providence: Department of Linguistics, Brown University, 1979.

GASPERIN, Caroline; LIMA, Vera Strube de. **Fundamentos do processamento estatístico da linguagem natural**. Porto Alegre: Faculdade de Informática/PUCRS, 2001. Relatório Técnico Nº 021. Disponível em: <http://www.inf.pucrs.br/relatorios/tr021.pdf>. Acesso: jun. 2006.

GREFENSTETTE, Gregory. The world wide web as a resource for example-based machine translation tasks. In: INTERNATIONAL CONFERENCE ON TRANSLATING AND THE COMPUTER , 20., 1999, London. **Proceedings of the**

ASLIB Conference on Translating and the Computer. London: Aslib, 1999. v. 21, pp. 110-116.

HALTEREN, Hans van (Ed.). **Syntactic wordclass tagging**. Dordrecht: Kluwer Academic, 1999. 334 p.

HERNÁNDEZ, Chantal Pérez. Explotación de los córpora textuales informatizados para la creación de bases de datos terminológicas basadas en el conocimiento. **Estudios de Lingüística Española**, Málaga, v. 18, 2002. Disponível em: <http://elies.rediris.es/elies18/index.html>. Acesso em: jun. 2006.

IDE, Nancy; MACLEOD, Catherine. The American National Corpus: a standardized resource of american english. In: CORPUS LINGUISTICS 2001, 2001, Lancaster, UK. **Proceedings of Corpus Linguistics 2001**. Lancaster, UK: Lancaster University, 2001. pp. 274-280.

IDE, Nacy; ROMARY, Laurent; DE LA CLERGERIE, Eric. International standard for a linguistic annotation framework. **Journal of Natural Language Engineering**, Cambridge, v. 10 n. 3-4, pp. 307-326, Sept. 2004.

IDE, Nancy; SUDERMAN, Keith. The American National Corpus 1rst release. In: LANGUAGE RESOURCES AND EVALUATION CONFERENCE, 4., 2004, Lisboa. **Proceedings of the Fourth Language Resources and Evaluation Conference (LREC)**. Lisboa: LREC, 2004. pp. 1681-1684.

IDE, Nancy; SUDERMAN, Keith. Integrating linguistic resources: the american national Corpus model. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 5., 2006, Génova. **Proceedings of the 5th International Conference on Language Resources and Evaluation**. Génova: LREC, 2006. Disponível em: <http://www.cs.vassar.edu/~ide/papers/ANC-LREC06.pdf>. Acesso em: jun. 2007.

IDE, Nancy; BONHOMME, Patrice; ROMARY, Laurent. XCES: an XML-based enconding standard for linguistic Corpora. In: INTERNATIONAL LANGUAGE RESOURCES AND EVALUATION CONFERENCE, 2., 2000, Atenas. **Proceedings of the Second International Language Resources and Evaluation Conference**. Paris: European Language Resources Association, 2000. pp. 825-830.

KEHOE, Andrew; RENOUF, Antoinette. WebCorp: applying the web to linguistics and linguistics to the web. In: INTERNATIONAL WIDE WORLD WEB CONFERENCE, 12., 2002, Honolulu, Hawaii. **Proceedings of the WWW2002 Conference**. [S.l.: s.n.], 2002. Disponível em: <http://www2002.org/CDROM/poster/67/> Acesso em: jun 2007.

KENNEDY, Graeme. **An introduction to Corpus linguistics**. London: Longman, 1998. 315 p.

KILGARRIFF, Adam; GREFENSTETTE, Gregory. Introduction to the special issue on the web as Corpus. **Computational Linguistics**, Cambridge, v. 29, n. 3, pp. 333-347, 2003.

KILGARRIFF, Adam; GREFENSTETTE, Gregory. Web as Corpus. In: CORPUS LINGUISTICS 2001 CONFERENCE, 2001, Lancaster, UK. **Proceedings of the Corpus Linguistics 2001 Conference**. Lancaster, UK: University of Lancaster, 2001. pp. 342-344.

KYTÖ, Merja. **Manual to the diachronic part of the Helsinki Corpus of english texts**: coding conventions and lists of source texts. Helsinski: Department of English/University of Helsinki, 1996.

LINGUATECA. **Acesso a corpora de português**: projecto AC/DC. Oslo, 2006. Disponível em: <http://www.linguateca.pt/ACDC/>. Acesso em: jun. 2006.

MAIA, Belinda; SARMENTO, Luís. Gestor de corpora: um ambiente Web integrado para lingüística baseada em corpora. In: ALMEIDA, José João (Ed.). **Corpora paralelos, aplicações e algoritmos associados (CP3A)**. Braga: Universidade do Minho, jun. 2003. pp. 25-30.

MCENERY, Tony; WILSON, Andrew. **Supplement the book Corpus linguistics**. Edinburgh: Edinburgh University Press, 1996. Disponível em: <http://www.lancs.ac.uk/fss/courses/ling/corpus/>. Acesso: jun. 2006

MANNING, Christopher; SCHÜTZE, Hinrich. **Foundations of statistical natural language processing**. Cambridge: The Mit Press, 1999. 680 p.

MCENERY, Tony; GABRIELATOS, Costas. English Corpus linguistics. In: AARTS, Bas; MCMAHON April (Ed.). **The handbook of english linguistics**. Malden: Blackwell Publishing, 2006. pp. 33-71.

MEGERDOOMIAN, Karine. Text mining, Corpus building, and testing. In: FARGHALY, Ali Ahmed Sabry (Ed.). **Handbook for language engineers**. Standford : CSLI, 2003. pp. 213-268.

MUNIZ, Marcelo et al. Taming the tiger topic: an XCES compliant corpus Portal to generate subcorpus based on automatic text topic identification. In: CORPUS LINGUISTICS 2007 CONFERENCE, 2007, Birmingham. **Proceedings of the Corpus Linguistics 2007 Conference**. Birmingham: University of Birmingham, 2007. Disponível em: <http://ucrel.lancs.ac.uk/publications/CL2007/>. Acesso em: nov. 2007.

NASCIMENTO, Maria Fernanda Bacelar. O corpus de referência do português contemporâneo e os projectos de investigação do Centro de Linguística da Universidade de Lisboa sobre variedades do português falado he escrito. In: GÄRTNER, E. et al. (Ed.). **Estudos de gramática portuguesa (I)**. Frankfurt am Main: Biblioteca Luso-Brasileira/Centro do Livro e do Disco de Língua Portuguesa, 2000. pp. 185-200.

PAUMIER, Sébastien. **Unitex 1.2**: user manual. Paris: university of Paris, 2006. Disponível em: <http://www-igm.univ-mlv.fr/~unitex/UnitexManual.pdf>. Acesso em: jul. 2007.

PINHEIRO, Gisele Montilha; ALUÍSIO, Sandra Maria. **Córpus Nilc**: descrição e análise crítica com vistas ao projeto Lacio-Web. SãoPaulo: USP, 2003. Apresentado no 51º Seminário do Grupo de Estudos Lingüísticos do Estado de São Paulo - GEL 2003 em maio 2003, UNITAU/São Paulo. Disponível em: <http://www.nilc.icmc.usp.br/lacioweb/downloads/NILC-TR-03-03.zip>. Acesso em: jun. 2006.

ROCHA, Paulo; SANTOS, Diana. CETEMPúblico: um corpus de grandes dimensões de linguagem jornalística portuguesa. In: ENCONTRO PARA O PROCESSAMENTO COMPUTACIONAL DA LÍNGUA PORTUGUESA ESCRITA E FALADA, 5., 2000, Atibaia, SP. **V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR 2000).** São Paulo: ICMC/USP, 2000. pp. 131-140.

REHBEIN, Ines. **Web as Corpus**: using web data for linguistic purposes. Dublin: NCLT/Dublin City University, 2006-2007. Disponível em: <http://www.nclt.dcu.ie/2006-2007NCLTSlides/irehbein_webcorpora.pdf>. Acesso em: jul. 2008.

RUSSELL, Stuart J.; NORVIG, Peter. **Artificial intelligence**: a modern approach. 2nd ed. Upper Saddle River, NJ : Prentice Hall, 2003. 1080 p.

SANTAMARÍA, Celina; GONZALO, Julio; VERDEJO, Felisa. Automatic association of web directories with word senses. **Computational Linguistics**, Arlington, v. 29, n. 3, pp. 485-502, 2003.

SANTOS, Diana. Disponibilização de corpora de texto através da WWW. In: WORKSHOP DA APL SOBRE LINGÜÍSTICA COMPUTACIONAL, 1., 1998, Lisboa. **Actas do I Workshop da APL sobre Lingüística Computacional**. Lisboa: Colibri, 1999. pp. 323-346.

SANTOS, Diana; ROCHA, Paulo. Evaluating CETEMPúblico, a free resource for Portuguese. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 39., 2001, Toulouse. **Proceedings of the 39th**

Annual Meeting of the Association for Computational Linguistics. Toulouse: ACL, 2001. pp. 442-449.

SARDINHA, Tony Berber. **Lingüística de Corpus**. Barueri : Manole, 2004. 410 p.

SARMENTO, Luís; MAIA, Belinda; SANTOS, Diana. The corpógrafo: a web-based environment for corpora research. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 4., 2004, Lisboa. **Proceedings of the IV International Conference on Language Resources and Evaluation**: LREC 2004. Lisbon: ELRA, 2004. pp. 449-452.

SCOTT, Mike. **Oxford wordsmith tools**: version 4.0. Oxford: Oxford University Press, c2004-2006. Disponível em: <http://www.lexically.net/downloads/version4/wordsmith. pdf> Acesso em: nov. 2006.

SERETAN, Violeta; NERIMA, Luka; WEHRLI, Eric. Using the web as a Corpus for the syntactic-based collocation identification. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 4., 2004, Lisboa. **Proceedings of the IV International Conference on Language Resources and Evaluation**: LREC 2004. Lisbon: ELRA, 2004. pp. 1871-1874.

SILVEIRA, Filipe; STRUBE DE LIMA, Vera Lúcia. Entrelinhas integração de ferramentas para compilação e exploração de corpora. In: TAGNIN, Stella; VALE, Oto (Ed.). **A lingüística de Corpus no Brasil**: pesquisa e crítica. São Paulo: Humanitas, 2008. pp. 247-268.

STEFANOWITSCH, Anatol. **Corpus compilation**. Bremen: Faculty of Linguistics and Literature/University of Bremen, 2003. Disponível em: <http://www-user.uni-bremen.de/~anatol/docs/corp_compilation.pdf>. Acesso em: jun. 2008.

SOUZA, José Guilherme et al. **Proposta de um esquema para integração anotação lingüística em XML**. São Leopoldo, 2006. Relatório Técnico Projeto PLN-BR (RT1-Unisinos-PLNBR).

VINTAR, Špela; BUITELAAR, Paul; SACALEANU, Bogdan et al. **MUCHMORE annotation format**. [Germany]: Muchmore, 2001. Disponível em: <http://muchmore.dfki.de/pubs/D4.1.pdf>. Acesso em: nov. 2006.

Apêndice A – Documento xcesDocType do XCES

```
<?xml version="1.0" encoding="UTF-16"?>
<ns:cesDoc version="" xmlns:ns="http://www.xces.org/schema/2003">
    <ns:cesHeader version="">
        <ns:fileDesc>
            <ns:titleStmt>
                <ns:title>00776200500804001.RTF</ns:title>
            </ns:titleStmt>
            <ns:sourceDesc>
                <ns:biblStruct>
                    <ns:monogr>
                        <ns:title/>
                        <ns:author/>
                        <ns:edition/>
                        <ns:respStmt/>
                        <ns:biblNote/>
                    </ns:monogr>
                </ns:biblStruct>
            </ns:sourceDesc>
        </ns:fileDesc>
        <ns:profileDesc>
            <ns:annotations>
                <ns:annotation type="ORIGINAL"
ann.loc="031D8C3F19154BCF948FACB8493E38ED.original.RTF"/>
                <ns:annotation type="CONTENT"
ann.loc="031D8C3F19154BCF948FACB8493E38ED.txt"/>
            </ns:annotations>
        </ns:profileDesc>
    </ns:cesHeader>
    <ns:text>
        <ns:body>
            <ns:p/>
            <ns:p>RECORRENTE(S): TRACTEBEL ENERGIA S.A.</ns:p>
            <ns:p>RECORRIDO(S):  CRISTÓVÃO DE ARAÚJO TORRADA</ns:p>
            <ns:p>ORIGEM: 8ª VARA DO TRABALHO DE PORTO ALEGRE</ns:p>
            <ns:p/>
            <ns:p>CERTIDÃO DE JULGAMENTO</ns:p>
            <ns:p>Processo TRT 00776-2005-008-04-00-1 ROPS</ns:p>
            <ns:p/>
            <ns:p>CERTIFICO  e dou fé que, em sessão realizada nesta data pela
Eg. 2ª Turma do Tribunal Regional do Trabalho da 4ª Região, sob a presidência
do Exmo. Juiz JOÃO GHISLENI FILHO, presentes os Exmos. Juízes JURACI GALVÃO
JÚNIOR, VANDA KRINDGES MARQUES e o Exmo. Procurador do Trabalho, Dr. LUIZ
FERNANDO MATHIAS VILAR, sendo Relator o Exmo. Juiz JOÃO GHISLENI FILHO, decidiu
a Turma, à unanimidade de votos, negar provimento ao recurso ordinário da
reclamada, mantendo a sentença por seus próprios fundamentos.</ns:p>
            <ns:p/>
            <ns:p>Porto Alegre, 9 de novembro de 2005.</ns:p>
            <ns:p/>
            <ns:p>Ceci Dal Mass Coser</ns:p>
            <ns:p>Secretária da 2ª Turma</ns:p>
        </ns:body>
    </ns:text>
</ns:cesDoc>
```

Apêndice B – 50 itens mais freqüentes do *corpus* de acórdãos

	Item	Ocorrências	Ocorrências (%)	Textos	Textos (%)
1	de	77942	5,42%	1057	100,00%
2	a	57112	3,97%	1057	100,00%
3	o	43275	3,01%	1053	99,62%
4	do	41832	2,91%	1057	100,00%
5	da	38466	2,67%	1057	100,00%
6	que	37284	2,59%	1057	100,00%
7	e	26359	1,83%	1057	100,00%
8	em	20777	1,44%	1054	99,72%
9	se	19682	1,37%	1034	97,82%
10	não	19159	1,33%	1004	94,99%
11	no	14191	0,99%	1013	95,84%
12	ao	14126	0,98%	1050	99,34%
13	na	10998	0,76%	974	92,15%
14	os	10466	0,73%	1039	98,30%
15	com	9660	0,67%	986	93,28%
16	por	9611	0,67%	1046	98,96%
17	para	9499	0,66%	964	91,20%
18	à	8952	0,62%	991	93,76%
19	trabalho	8673	0,60%	1057	100,00%
20	reclamante	8127	0,57%	766	72,47%
21	das	7760	0,54%	949	89,78%
22	dos	7429	0,52%	934	88,36%
23	as	7123	0,50%	937	88,65%
24	pelo	6348	0,44%	943	89,22%
25	é	6274	0,44%	997	94,32%
26	reclamada	6182	0,43%	719	68,02%
27	recurso	5933	0,41%	905	85,62%
28	como	5766	0,40%	903	85,43%
29	pela	5328	0,37%	973	92,05%
30	horas	5054	0,35%	376	35,57%
31	fls	4988	0,35%	919	86,94%
32	pagamento	4785	0,33%	764	72,28%
33	sentença	4590	0,32%	884	83,63%
34	fl	4555	0,32%	874	82,69%
35	ou	4374	0,30%	822	77,77%
36	nº	4179	0,29%	722	68,31%
37	art	4154	0,29%	772	73,04%
38	nos	4107	0,29%	891	84,30%
39	º	4090	0,28%	798	75,50%
40	provimento	3962	0,28%	988	93,47%
41	ª	3954	0,28%	1057	100,00%
42	autos	3904	0,27%	966	91,39%
43	lei	3624	0,25%	701	66,32%
44	extras	3559	0,25%	352	33,30%
45	ser	3513	0,24%	802	75,88%
46	às	3459	0,24%	773	73,13%
47	autor	3453	0,24%	501	47,40%
48	decisão	3324	0,23%	845	79,94%
49	aos	3276	0,23%	799	75,59%
50	sobre	3060	0,21%	709	67,08%

Anexo A – Parecer sobre a Entrelinhas

ENTRELINHAS: UM TESTE

Susana de Azeredo

O objetivo aqui é fazer um relato de um teste com o software *ENTRELINHAS.*

O teste foi dividido em dois momentos. Em um primeiro momento, foi utilizado um corpus previamente montado de cerca de 100 mil palavras (todos os arquivos desse corpus eram de formato .txt). O objetivo neste primeiro momento foi testar a ferramenta de contagem de palavras e o concordanciador. Em um segundo momento, foi feito a montagem de um corpus através do *ENTRELINHAS.* O objetivo no segundo momento, foi testar a eficiência do Entrelinhas com relação à montagem do corpus.

A seguir, detalhamos os dois momentos e as observações feitas.

PRIMEIRO MOMENTO:

Neste primeiro momento de teste, utilizamos um corpus previamente montado, o qual chamaremos aqui de corpus QUIM. Esse corpus compreende textos de Química, retirados de dois manuais de Química Geral[92]. O corpus QUIM é composto de cerca de 100 mil palavras e os arquivos estavam todos em formato .txt.

O objetivo aqui foi testar as ferramentas de contagem de palavras e o concordanciador.

Com relação à contagem de palavras:

A contagem das palavras é bem eficiente. No entanto, coloco abaixo algumas observações:

a) No final da lista das palavras contadas, aparece o código dos arquivos como palavras:

12175BAF919F639E7916C13C6D1FBB5B	1	00,001%	1	012,500%
22CA38DB811AD6C5BD3789707F363CDF	1	00,001%	1	012,500%
75A8FC63272AD621384C2D4A05EF54FE	1	00,001%	1	012,500%
786716DDDFAB182E3F05C583E09B0350	1	00,001%	1	012,500%
8193DED1336D013693B43810F6DA4702	1	00,001%	1	012,500%
AFBCD200DCA5373217EEE5CE0EDCA1C7	1	00,001%	1	012,500%

[92] Esse corpus é uma amostra do corpus utilizado no projeto TEXTQUIM (www.ufrgs.br/textquim) na UFRGS, onde o objetivo, dito de uma forma bem ampla é o estudo da linguagem da Química. Para a elaboração desse corpus, no projeto, foi necessário o escaneamento dos textos e a revisão, linha por linha, de cada texto. Portanto, a margem de erros de grafia é mínima. Além disso, o corpus está todo preparado com algumas marcações. Por exemplo, onde havia uma tabela, há a marcação: <TABELA>. Os sinais gráficos de maior e menor foram utilizados, pois o software utilizado no projeto TEXTQUIM (Wordsmith Tools) não lê o que está entre esses sinais. Isso é muito útil, uma vez que, para um lingüista, é relevante saber onde estão as tabelas e os gráficos no texto, mesmo que eles não apareçam.

D2E17123AEDD5A8246CA8515B397F292	1	00,001%	1	012,500%
EADA94F81E56CE56E112841372172888	1	00,001%	1	012,500%

Acima, pode-se ver que a palavra “últimos” é a última palavra lexical que aparece na lista. Após ela, vêm os códigos dos arquivos. Parece-me que esses códigos também são contados como palavras. *Isso daria um resultado irreal do número de palavras constantes no corpus.*

b) Um outro fator é que algumas palavras encontradas no concordanciador não aparecem na contagem de palavras. Por exemplo: A palavra “tabela” e a palavra “gráfico” não aparecem na lista de contagem de palavras. No entanto, se procurarmos essas palavras no concordanciador, elas aparecem. Então, a pergunta é: Porque elas não aparecem na contagem? Elas não foram contadas como palavras no corpus?

Aqui, se percebe que não há a palavra TABELA na lista de contagem:

sítios	1	00,001%	1	012,500%
sólidas	1	00,001%	1	012,500%
tabelados	1	00,001%	1	012,500%
tamanho	1	00,001%	1	012,500%
tamponados	1	00,001%	1	012,500%

Aqui, se percebe que não há a palavra GRÁFICO na lista de contagem

governam	1	00,001%	1	012,500%
gr	1	00,001%	1	012,500%
gradual	1	00,001%	1	012,500%
gradualmente	1	00,001%	1	012,500%
gramas	1	00,001%	1	012,500%
grave	1	00,001%	1	012,500%
graves	1	00,001%	1	012,500%
gravitacional	1	00,001%	1	012,500%

Aqui, aparece a palavra TABELA no concordanciador

anto Lewis tentava explicar a **tabela** periódica aos alunos do prime
átomo, porém constatou que a **Tabela** periódica poderia ser explica
cos seguem uma certa ordem na **tabela** periódica, Lewis supôs que os
encial, ou número atômico, na **Tabela** periódica. Apesar desses avan
o cúbico de oito elétrons. Na **Tabela** 6.2 está mostrada uma parte d
.2 está mostrada uma parte da **tabela** periódica para os primeiros 2
ria ao seu número atômico. Na **Tabela** 6.2, o número atômico de cada

Aqui, aparece a palavra GRÁFICO no concordanciador

ilidade do iodo é indicar num **gráfico** as concentrações de iodo diss
4 é perceptível. O tratamento **gráfico** dos equilíbrios que analisamo
mentos baseados no tratamento **gráfico** de sistemas simples fornecem
ma ao equilíbrio seguindo, no **gráfico**, uma linha de inclinação +1.

c) A ferramenta de contagem de palavras revela o número total de palavras do corpus (tanto os itens lexicais quanto os itens lexicais distintos). Esse informação é muito relevante para um lingüista. No entanto, seria útil revelar também o número de palavras de cada texto. Essa é uma informação importante para o lingüista, pois é possível fazer uma comparação entre os textos do corpus, traçando diferentes perfis de textos. Além disso, não tem como selecionar apenas um dos textos do corpus para fazer uma análise específica ou uma comparação desse texto específico com o restante do corpus.

d) Uma informação interessante para um lingüista e que não consta no Entrelinhas é a relação entre o número total de palavras do corpus e o número de palavras diferentes do corpus. Essa relação permite ver a variedade vocabular de um texto[93].

e) Com relação ao tempo que leva para finalizar a operação, pode-se dizer que foi bem eficiente. Há uma barra que indica que o Entrelinhas está fazendo o processamento estatístico, o que é bem útil. Uma única observação é que essa barra, às vezes, tranca e ficamos sem saber se o processo já terminou ou não.

[93] Para verificar como se calcula essa relação, ver: BERBER, Tony (2004). Lingüística de Corpus. Manole, São Paulo.

Anexo B – Complemento do parecer sobre a Entrelinhas

1) AS PALAVRAS "GRÁFICO" E "TABELA" QUE NÃO APARECIAM NA LISTA DE CONTAGEM: Eu selecionei novamente o corpus QUIM e, realmente, agora as duas palavras apareceram. Talvez tenha dado algum problema na minha máquina quando fiz a exploração do corpus anteriormente.

2) INTERFACE: A Interface do Entrelinhas é muito elegante, agradável e de muito bom gosto. Também, não há uma poluição visual que confunde o usuário.

3) USABILIDADE: A usabilidade do Entrelinhas é boa. No início, o Filipe deu algumas dicas para o primeiro acesso. Mas, acredito ser possível o acesso sem preparação antecipada sobre o programa. Penso ser interessante que o "botão" "explorar corpus" possa aparecer na janela de "compilação do corpus", o que tornaria a atividade que está sendo realizada mais dinâmica. A escolha do diretório é bastante recorrente e, em alguns casos, parece desnecessária.

4) INSTALAÇÃO: A instalação do Entrelinhas foi fácil e rápida. Também, não foi necessário instalar outros programas para que o Entrelinhas pudesse ser utilizado.

Susana de Azeredo

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo