Criação de um ambiente para o processamento de córpus de Português Histórico

Arnaldo Candido Junior

Criação de um ambiente para o processamento de córpus de Português Histórico

Arnaldo Candido Junior

Orientador: ***Profa. Dra. Sandra Maria Aluísio***

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Ciências - Ciências de Computação e Matemática Computacional.

USP – São Carlos
Fevereiro/2008

Agradecimentos

Agradeço a Deus por esta oportunidade de estudo e desenvolvimento pessoal.

Agradeço ao meu pai Arnaldo, à minha mãe Maria e à minha irmã Amanda, pessoas fundamentais na minha vida.

Agradeço à minha família toda, pelo convívio, exemplo e educação, pelos laços que foram criaram e que não podem ser desfeitos.

Agradeço à Michelle, pelo companheirismo e carinho dos últimos 12 meses, muito importantes para mim.

Agradeço à minha orientadora Sandra, pela amizade, pela orientação exemplar, pela paciência e pela dedicação nos últimos dois anos.

Agradeço à professora Maria Tereza, pela sua dedicação incansável que possibilitou o desenvolvimento desta pesquisa e pela orientação durante do trabalho.

Agradeço a todos os pesquisadores do projeto DHPB, pelo apoio de grande importância na realização desta pesquisa.

Agradeço aos amigos do NILC Ariani, Carmen, Eliane, Eloize, Helena, Leandro, Livia, Luiz, Marcelos, Ricardo, e todos os outros pela convivência e amizade.

Agradeço aos professores do NILC Gladis, Graça, Oto, Thiago e todos os outros, pelo exemplo e pela orientação.

Agradeço aos amigos do LCAD II Dalcimar, Marco e Jarbas pelos momentos divertidos.

Agradeço a todos os colegas de mestrado, aos amigos do ICMC e da USP em geral.

Agradeço às professoras Vera, Maria José e Renata e a todos os amigos do Rio Grande do Sul.

Agradeço ao professor Adriano, pelo empenho, dedicação e orientação durante a minha graduação.

Agradeço ao CNPq, cujo apoio foi muito importante para a realização deste trabalho.

Agradeço a todas as pessoas não foram mencionadas nestas palavras mais do que breves, mas que foram e são pessoas importantes na minha vida.

"Há apenas um bem, o conhecimento; e um mal, a ignorância"
(Sócrates)

Sumário

Índice de figuras

Índice de tabelas

Lista de abreviaturas

AJAX: Asynchronous JavaScript and XML (XML e JavaScript Assíncronos)

ANC: American National Corpus (Córpus Nacional Americano).

BNC: British National Corpus (Córpus Nacional Britânico).

CES: Corpus Encoding Standard (Padrão de Codificação de Córpus).

CSS: Cascading Style Sheets (Folhas de Estilo em Cascata).

DELA: Dictionnarie Electronique du LADL (Dicionários Eletrônicos do LADL).

DHPB: Dicionário Histórico do Português do Brasil.

DTD: Document Type Definition (Definição de Tipo de Documento).

EAGLES: Expert Advisory Group on Language Engineering Standards (Grupo Especialista em Recomendações sobre Padrões de Engenharia da Linguagem).

HTML: HiperText Markup Language (Linguagem de Marcação de HiperTexto).

LADL: Laboratoire d'Automatique Documentaire et Linguistique.

LW: Lácio-Web.

NILC: Núcleo Interinstitucional de Lingüística Computacional.

OCR: Optical Character Recognition (Reconhecimento Óptico de Caracteres).

Procorph: Processador de Córpus Históricos.

Protej: Processador de Textos Históricos em Java.

Protew: Processador de Textos Históricos em MS-Word.

PLN: Processamento de Língua Natural.

RE: Recuperação de Informação.

ReGra: Revisor Gramatical.

RSNSR: Rule-Based Search in Text Data Bases with Non-standard Orthography.

TEI: Text Encoding Initiative

VARD: VARiant Detector.

XCES: XML CES.

Resumo

A utilização de córpus tem crescido progressivamente em áreas como Lingüística e Processamento de Língua Natural. Como resultado, temos a compilação de novos e grandes córpus e a criação de sistemas processadores de córpus e de padrões para codificação e intercâmbio de textos eletrônicos. Entretanto, a metodologia para compilação de córpus históricos difere das metodologias usadas em córpus contemporâneos. Outro problema é o fato de a maior parte dos processadores de córpus proverem poucos recursos para o tratamento de córpus históricos, apesar de tais córpus serem numerosos. Da mesma forma, os sistemas para criação de dicionários não atendem satisfatoriamente necessidades de dicionários históricos. A motivação desta pesquisa é o projeto do Dicionário Histórico do Português do Brasil (DHPB) que tem como base a construção de um córpus de Português do Brasil dos séculos XVI a XVIII (incluindo alguns textos do começo do século XIX). Neste trabalho são apresentados os desafios encontrados para o processamento do córpus do projeto DHPB e os requisitos para redação de verbetes do dicionário histórico. Um ambiente computacional para processamento de córpus, criação de glossários e redação de verbetes foi desenvolvido para o projeto DHPB, sendo possível adaptá-lo para ser aplicado a outros projetos de criação de dicionários históricos.

Abstract

Corpora have been increasingly used within the areas of Linguistics and Natural Language Processing. As a result, new and larger corpora have been compiled and processing systems and standards for encoding and interchange of electronic texts have been developed. However, when it comes to compilation of historical corpora, the methodology is different from the ones used to compile corpora of contemporary language. Another drawback is the fact that most corpus processing systems provide few resources for the treatment of historical corpora, although there are numerous corpora of this type. Similarly, the systems for dictionary creation do not satisfactorily meet the needs of historical dictionaries. The present study is part of a larger project – the Historical Dictionary of Brazilian Portuguese (HDBP) – which aims to compile a dictionary on the basis of a corpus of Brazilian Portuguese texts from the sixteenth through the eighteenth centuries (including some texts from early nineteenth century). Here, we present the challenges for processing the corpus of the HDPB project and establish the criteria for creating the entries of a historical dictionary. This study has developed a computational environment for processing the corpus, building glossaries as well as for creating the entries of the HDPB. This system can be easily adapted to the needs and scope of other historical dictionary projects.

1 Introdução

1.1 Contextualização

Córpus[1] podem ser definidos como uma coleção de dados lingüísticos (sejam eles textos ou partes de textos escritos ou a transcrição de fala) de uma determinada língua, escolhidos segundo um determinado critério, representando uma amostra dessa língua ou uma variedade lingüística (SARDINHA, 2004). Aluísio e Almeida (2006) apresentam diversas definições de córpus e as agrupam sob duas perspectivas diferentes: da Lingüística e da Lingüística de Córpus. A principal diferença entre as duas é relativa ao formato do córpus. Para a Lingüística, os córpus podem ser compostos por documentos impressos, enquanto que para Lingüística de Córpus, os córpus devem, obrigatoriamente, estar em formato eletrônico. Neste trabalho, a perspectiva da Lingüística de Córpus será utilizada.

A construção e uso de córpus eletrônicos ainda estão em sua infância, embora um grande progresso com respeito a projeto de córpus tenha ocorrido. Existem diversos projetos de córpus, citados por diferentes autores (ATKINS; CLEAR; OSTLER, 1992; MCENERY; WILSON, 1996; SARDINHA, 2004; ALUÍSIO et al., 2003b, 2004; MUNIZ et al 2007). Cada projeto possui objetivos e finalidades distintas. A seguir, serão descritos alguns destes projetos.

Projetos de córpus internacionais:

- *Brown Corpus of Standard American English*: projeto de córpus para o idioma inglês criado em 1964 (o primeiro córpus desenvolvido). O córpus foi lançado com 1 milhão de palavras, o que é um volume de informação considerável para a tecnologia da época.
- ICE (*International Corpus of English*): projeto iniciado em 1990 com o objetivo de coletar material para estudos comparativos do Inglês no mundo. Para execução desse projeto, foram planejados subcórpus com 1 milhão de palavras para 18 variantes nacionais (ou regionais) do Inglês com textos produzidos após 1989. Parte dos subcórpus já está completa, e parte ainda em construção.
- BNC (*British National Corpus*) (BURNAGE; DUNLOP, 1992): córpus do Inglês

1 Neste texto, foi utilizada a variação portuguesa da palavra "córpus" ao invés de sua variação em latim "*corpus*".

britânico falado e escrito com um total de 100 milhões de palavras. O projeto se iniciou em 1991 e terminou em 1994.

- *The Bank of English*: inclui diferentes tipos de textos falados e escritos, a maior parte produzida após 1990. O projeto se iniciou em 1991 na Universidade de Birmingham. Em 2005, o córpus possuía mais de 524 milhões de palavras.
- CNC (*Czech National Corpus*): córpus para a Língua Tcheca contemporânea de 100 milhões de palavras com uma versão disponibilizada publicamente com mais de 20 milhões de palavras. O projeto de construção desse córpus data de 1994 com a criação de um instituto para seu gerenciamento.
- FRANTEXT (primeiramente conhecido como *Trésor de la Langue Française*): córpus para a Língua Francesa constituído por 2 mil textos, totalizando mais de 150 milhões de palavras. São incluídos textos dos séculos XVI até XX. O projeto se iniciou em 1998 unindo centros de pesquisadores da Europa, Austrália, Canadá e Japão.
- ANC (*American National Corpus*) (MACLEOD; IDE; GRISHMAN, 2000; IDE; SUDERMAN, 2004): córpus em construção do Inglês americano com textos posteriores a 1990, com 100 milhões de palavras no total (textos falados e escritos). A proposta de construção do ANC data de 1998 e sua primeira disponibilização pública se deu em 2003.

Projetos de córpus do Português:

- AC/DC (Acesso Corpora / Disponibilização Corpora) (SANTOS; BICK, 2000): o projeto AC/DC faz parte da Linguateca, um centro de recursos distribuído para o processamento computacional da Língua Portuguesa. O projeto agrega textos de 22 projetos de córpus diferentes, que somados totalizam 371 milhões de palavras.
- Córpus do NILC (Núcleo Interinstitucional de Lingüística Computacional) (PINHEIRO; ALUÍSIO, 2003): córpus desenvolvido durante a criação da ferramenta ReGra (Revisor Gramatical) (NUNES; OLIVEIRA JR., 2000), um corretor gramatical construído para Língua Portuguesa e integrado ao *MS-Word*. O córpus do NILC possui mais de 41 milhões de palavras com textos de vários gêneros, com uma grande participação de textos jornalísticos obtidos a partir de edições do jornal A

Folha de São Paulo de 1994.

- Córpus do Português: reúne textos das variantes brasileira e européia do Português. Os textos foram criados entre os séculos XIV e XX. Atualmente, o córpus conta com 45 milhões de palavras e inclui textos do córpus Tycho-Brahe e do *Lácio-Web* (LW), entre outros projetos.
- DICIWeb: um projeto de córpus lexicográfico concebido por Alexandre Miguel Moura Maia Fernandes Moreira e Sérgio Paulo Cardoso Barbosa na Universidade de Aveiro. Diferentemente de projetos de córpus para tarefas lexicográficas, esse córpus é constituído por dicionários de Português e textos lexicográficos relacionados. A versão do córpus disponibilizada para acesso público conta com 378 mil palavras.
- *Lácio-Web* (LW) (ALUÍSIO et al., 2003a, 2003b, 2004): O projeto LW tem como objetivo a compilação de córpus do Português do Brasil e a implementação de ferramentas para análises lingüísticas. Quatro córpus foram desenvolvidos durante o projeto: (a) um córpus aberto e de referência de Português Contemporâneo (Lácio-ref); (b) um córpus fechado e manualmente anotado morfossintaticamente, composto de 1,2 milhões de palavras (Mac-morpho); (c) um córpus de textos em Inglês e Português com originais e traduções (Par-c); (d) um córpus de textos originais em Português e Inglês com conteúdos comparáveis (Comp-c).
- PLN-BR: O projeto PLN-BR tem como objetivo a construção de recursos e ferramentas para a recuperação de informação em bases textuais em Português do Brasil. O projeto foi aprovado pelo CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) no âmbito do edital CTInfo/MCT/CNPq nº 011/2005. Possui 3 córpus de trabalho sendo um deles o PLN-BR GOLD, distribuído publicamente, contando com 1024 mil textos jornalísticos da Folha de São Paulo, entre 1994 e 2005, anotado em vários níveis lingüísticos, seguindo uma amostra de um ano construído.
- Tycho-Brahe (GALVES; BRITTO, 1999; ALVES; FINGER, 1999; BRITTO; FINGER, 1999): um córpus de Português Histórico composto por 40 textos criados entre os séculos XVI e XIX, totalizando 2,3 milhões de palavras. Os textos estão disponibilizados publicamente e possuem anotação morfossintática e sintática.

Grandes córpus contribuem para a descrição de uma língua, por exemplo, através da construção de recursos como dicionários e gramáticas. O que pode ser menos visível numa primeira análise é que eles também impulsionam o desenvolvimento de padrões internacionais de anotação e codificação, bem como de diversas ferramentas de Processamento de Língua Natural (PLN), amplamente utilizadas para o processamento de córpus. As ferramentas desenvolvidas (por exemplo, lematizadores[2] e etiquetadores morfossintáticos[3]), por sua vez, permitem a própria construção das anotações lingüísticas desses grandes recursos.

Apesar da disponibilidade de grandes córpus, como os citados acima, novos córpus continuam sendo construídos para finalidades diversas, com o objetivo de cobrirem novos gêneros como o da *Web*, de serem maiores, mais balanceados e/ou mais flexíveis que os atuais. Os córpus e suas ferramentas são um auxílio importante para se progredir de maneira rápida e confiável na compreensão das línguas. Nesses projetos, é importante o uso de padrões internacionais de anotação, pois facilitam o reuso do córpus em pesquisas diferentes e o seu uso nas ferramentas de manipulação de córpus.

O projeto Dicionário Histórico do Português do Brasil (DHPB[4]) aprovado no âmbito do Programa Institutos do Milênio (edital MCT/CNPq nº 01/2005) consiste na criação de um dicionário de Português do Brasil entre os séculos XVI e XIX (até 1808) a partir de um córpus constituído por documentos históricos desses séculos. O projeto conta com 41 pesquisadores pertencentes a 11 universidades, dos quais 18 são doutores.

Uma diferença entre o DHPB e o Tycho-Brahe é o fato de o segundo ter como objetivo a criação de um córpus de domínio público e com anotação em níveis morfossintático e sintático. O córpus DHPB não será publicamente disponibilizado inicialmente, pois é necessária a obtenção de direitos autorais de distribuição para parte dos textos pertencentes ao córpus. Como o córpus do projeto DHPB não é anotado morfossintaticamente, a compilação do córpus tende a ser mais rápida, o que permite a obtenção de um córpus maior. O tamanho do córpus é um fator importante para a criação de dicionários, pois tarefas lexicográficas requerem córpus de grandes proporções que tragam os vários sentidos de uma palavra. O DHPB difere do Córpus do Português devido ao período de tempo estudado e ao fato de o DHPB conter apenas textos escritos por brasileiros (ou por portugueses que viveram um período de tempo grande no Brasil).

2 Ferramentas capazes de encontrar a forma canônica de uma palavra a partir de uma de suas flexões
3 Ferramentas que identificam automaticamente a categoria gramatical de uma dada palavra
4 Siga não oficial

O córpus foi totalmente compilado, contando como 2.458 textos e 7.5 milhões de formas simples. Os textos selecionados para o córpus incluem cartas dos missionários jesuítas, documentos dos bandeirantes, relatos dos sertanistas, documentos da inquisição católica, inventários e testamentos, entre outros.

O projeto DHPB é o primeiro dicionário histórico voltado para o Português do Brasil. Nos primeiros séculos da nossa história, o Português do Brasil era semelhante ao Português de Portugal. Entretanto, as duas variantes do Português começaram a diferir com o passar do tempo. A Figura 1.1 mostra quatro grandes etapas relativas à criação do dicionário e do córpus do projeto DHPB. As etapas de coleta e conversão de textos e de geração de córpus estão relacionadas às etapas de compilação e anotação do córpus (descritas no Capítulo 2).

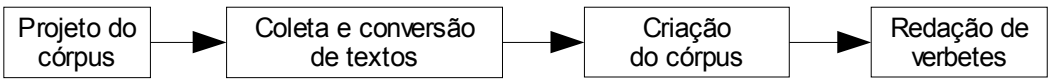

Figura 1.1: Etapas para construção do córpus e do dicionário do projeto DHPB

A **etapa de projeto** envolve a coleta e seleção dos textos que farão parte do córpus usado na confecção do dicionário. Os textos, em sua maioria, são levantados a partir de pesquisas bibliográficas em arquivos públicos e bibliotecas brasileiras e portuguesas.

A **etapa de coleta e conversão dos textos** envolve a obtenção de textos e conversão para o formato digital, especificamente, para o formato de texto com formatação[5]. Os textos são obtidos de três maneiras distintas: impressos, manuscritos e arquivos computacionais em formato PDF (*Portable Document Format*) (que, por sua vez, foram digitalizados a partir de documentos inacessíveis provenientes de acervos únicos). Textos impressos são digitalizados (convertidos para arquivos em formato de imagem). Da mesma forma, os arquivos em PDF são convertidos para imagens. As imagens são convertidas em texto com formatação através do processo de Reconhecimento Óptico de Caracteres (*Optical Character Recognition*). Manuscritos são analisados e transcritos para o formato eletrônico através de digitação. O processo é ilustrado na Figura 1.2.

5 A formatação está relacionada à exibição visual dos textos e trata de recursos como negrito, tamanhos diferentes de fonte, sobrescrito, entre outros.

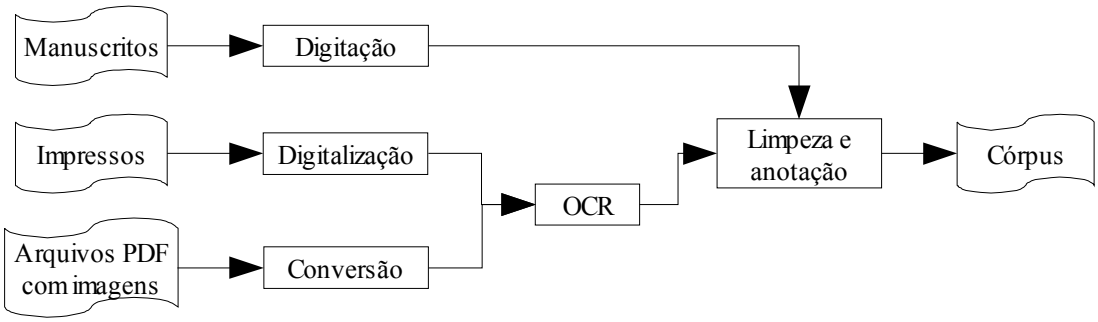

Figura 1.2: Conversão dos textos do projeto DHPB

A Figura 1.3 mostra um pequeno trecho da carta do Padre Manuel da Nóbrega ao Padre Simão Rodrigues escrita em 1549 (em formato de imagem). A Figura 1.4 mostra o mesmo trecho da carta após a conversão para o formato de texto com formatação. É possível observar algumas alterações no número de linhas e na distribuição das palavras.

Os arquivos de textos criados são então revisados manualmente na busca de erros ocorridos durante a conversão dos textos para o formato eletrônico. A seguir, cada texto recebe uma ficha catalográfica com informações bibliográficas (Fonte), tipológicas (Tipologias), sobre a revisão do texto (Revisão) e dados do arquivo final (Formato do Arquivo Final) como mostra a Tabela 1.1. As informações 2.1, 2.2, 3a, 3b, 22 e 25 da tabela não foram inseridas para este texto. A etapa de coleta termina com o tratamento manual de hifenização no documento. Durante o processo, hífens inseridos para quebra de palavras em fins de linha são removidos e hífens usados em lexias compostas são preservados. Os documentos resultantes (em formato de texto com formatação) são estruturalmente semelhantes às suas versões originais (impressos, manuscritos ou PDFs).

A **etapa de criação do córpus** envolve a conversão dos textos com formatação para um formato capaz de ser tratado por meio de ferramentas de processamento de córpus. O primeiro passo consiste na remoção de formatação (caso necessário, é possível utilizar etiquetas para denotar recursos de formatação). A seguir, é feita a limpeza do texto, na qual algumas informações são removidas para melhorar o desempenho de ferramentas de processamento de córpus (por exemplo, numeração de linhas ou de parágrafos). Após a limpeza, informações importantes sobre o texto (por exemplo, nome do autor e notas de rodapé) são anotadas também com o objetivo de melhorar o desempenho dos processadores de córpus. Nessa etapa, é possível analisar os textos novamente buscando novos erros ou erros ocorridos durante a

coleta e conversão que não foram corrigidos. Adicionalmente, é possível gerar glossários[6] para o apoio ao processamento do córpus (por exemplo, um glossário de abreviaturas ou de variações de grafia). O córpus gerado pode então ser processado automaticamente para geração de diferentes versões para uso em diferentes processadores de córpus. Os textos resultantes são estruturalmente diferentes de suas versões com formatação e, conseqüentemente, de suas versões originais. Entretanto, o conteúdo entre os textos do córpus e os textos originais é idêntico. A Figura 1.5 mostra o resultado da conversão do trecho de uma carta do Padre Manuel da Nóbrega para o formato TEI (*Text Encoding Initiative*). Somente alguns campos da ficha catalográfica foram convertidos no cabeçalho TEI (autor, título, data de produção e data de edição), entretanto, outros dados serão adicionados em trabalhos futuros. O formato TEI será discutido em detalhes no Capítulo 2. A figura também inclui informações da ficha catalográfica etiquetadas no cabeçalho do arquivo.

A **etapa de redação de verbetes** consiste na criação do dicionário propriamente dito, o objetivo principal do projeto DHPB. Durante a elaboração de verbetes o córpus gerado é analisado por lexicógrafos com o amparo de ferramentas computacionais.

6 Os termos "léxico" e "léxico computacional" podem ser utilizados no lugar de "glossário" em textos da literatura.

convertidos, onde estaremos Vicente Rodriguez e eu, e hum soldado[19] que se meteo comnosco para nos servir, e está agora em Exercicios, de que eu estou muy contente. Faremos nossa igreja, onde insinemos os nossos novos christãos, e aos domingos e festas visitarey a Cidade e pregarey.

O Padre Antonio Pirez e o P.e Navarro estaram em outras Aldeas longe, onde já lhes fazem casas. E portanto hé necessario V. R. mandar officiaes, e am-de vir já com a paga, porque cá diz ho Governador que, ainda que venha Alvará de S. A. para nos dar o necessario, que nom o averá hi para isto. Os officiaes que cá estão tem muito que fazer, e que o nom tenhão, estão com grande saudade do Reyno, porque deixão lá suas molheres e filhos, e nom aceitaram a nossa obra depois que cumprirem com S. A., e tambem ho trabalho que tem com as viandas e o mais os tira disso. Portanto me parece que avião de vir de lá, e, se possivel fosse, com suas molheres e filhos, e alguns que fação taipas e carpinteiros. Cá está hum Mestre para as obras, que hé hum sobrinho[20] de Luis Diaz, mestre das obras d'El-Rey, ho qual veo con trinta mil reis de partido. Este nom hé necessario porque abasta ho tio para as obras de S. A.; a este avião de dar o cuidado do nosso collegio; hé bom official. Serão cá muito necessarias pessoas que teção algodão, que há muito, e outros officiaes.

12. Trabalhe V. R. por virem a esta terra pessoas casadas, porque certo hé mal empregada esta terra em degradados, que cá fazem muyto mal, e já que cá viessem avia de ser para andarem afferrolhados nas obras de S. A.

13. Tambem peça V. R. algum petitorio para roupa, para entretanto cubrirmos estes novos convertidos, ao menos huma camisa a cada molher, polla honestidade da religião christã, porque vem todos a esta Cidade à missa aos domingos e festas, que faz muita devação, e vem rezando as ora-

19 Simão Gonçalves. LEITE I 573.

20 Este «bom oficial», sobrinho de Luís Dias, era Diogo Peres. LEITE I 22.

Figura 1.3: Exemplo de texto convertido para imagem

7. - BAÍA 9 DE AGOSTO DE 1549 127

convertidos, onde estaremos Vicente Rodriguez e eu, e hum soldado[19] que se meteo comnosco para nos servir, e está agora em Exercicios, de que eu estou muy contente, Faremos nossa igreja, onde insinemos os nossos novos christãos, e aos domingos e festas visitarey a Cidade e pregarey. 205

O Padre Antonio Pirez e o P.[e] Navarro estaram em outras Aldeas longe, onde já lhes fazem casas. E portanto hé necessario V. R. mandar officiaes, e am-de vir já com a paga, porque cá diz ho Governador que, ainda que venha Alvará de S. A. para nos dar o necessario, que nom o averá 2I0 hi para isto. Os officiaes que cá estão tem muito que fazer, e que o nom tenhão, estão com grande saudade do Reyno, porque deixão lá suas molheres e filhos, e nom aceitaram a nossa obra depois que cumprirem com S. A., e tambem ho trabalho que tem com as viandas e o mais os tira disso. 2I5 Portanto me parece que avião de vir de lá, e, se possivel fosse, com suas molheres e filhos, e alguns que fação taipas e carpinteiros. Cá está hum Mestre para as obras, que hé hum sobrinho[20] de Luis Diaz, mestre das obras d'El-Rey, ho qual veo con trinta mil reis de partido. Este nom hé 220 necessario porque abasta ho tio para as obras de S. A.; a este avião de dar o cuidado do nosso collegio; hé bom official. Serão cá muito necessarias pessoas que teção algodão, que há muito, e outros officiaes.

12 Trabalhe V. R. por virem a esta terra pessoas casadas, 225
porque certo hé mal empregada esta terra em degradados,
que cá fazem muyto mal, e já que cá viessem avia de
ser para andarem afferrolhados nas obras de S. A.

13 Tambem peça V. R. algum petitorio para roupa,
para entretanto cubrirmos estes novos convertidos, ao menos 230
huma camisa a cada molher, polla honestidade da religião
christã, porque vem todos a esta Cidade à missa aos domingos
e festas, que faz muita devação, e vem rezando as orações

I9 Simão Gonçalves. LEITE I 573.
20 Este «bom oficial», sobrinho de Luís Dias, era Diogo Peres. LEITE I 22.

Figura 1.4: Texto da Figura 1.3 convertido para o formato de texto com formatação

Tabela 1.1: Ficha catalográfica do texto da Figura 1.4

Tipologias
1. Tipo da Fonte: EDIÇÃO IMPRESSA
2.1 Domínio Discursivo/Subdomínio Discursivo:
2.2 Gênero/Subgênero:
3a. Tipologia de Assuntos:
3b. Características Sociolingüísticas do Autor:
4. Descrição: CARTAS JESUÍTICAS DISPOSTAS EM ORDEM CRONOLÓGICA, ORGANIZADAS E, QUANDO PRECISO, TRADUZIDAS E ANOTADAS PELO P.e SERAFIM LEITE (1538-1553) - 3 VOLUMES
5: Localização da Obra: UNESP - CAMPUS DE ARARAQUARA
Fonte
6 Nome do Autor do Texto: P. MANUEL DA NÓBREGA
7: Título do Texto: CARTA DO P. MANUEL DA NÓBREGA AO P. SIMÃO RODRIGUES, BAÍA 9 DE AGOSTO 1549
8. Data em que o Texto foi produzido pelo Autor: BAÍA 9 DE AGOSTO 1549
9. Amostra: INTEGRAL
10. Título da Obra: CARTAS DOS PRIMEIROS JESUÍTAS DO BRASIL
11. Editor: SERAFIM LEITE S. J
12. Organizador/Coordenador (coletânea/livro): SERAFIM LEITE S. J
13. Editora: COMISSÃO DO IV CENTENÁRIO DA CIDADE DE SÃO PAULO
14. Local da Edição: SÃO PAULO
15: Data da Edição: 1956
16: Número da Edição:
17: Volume: I
18: Tipografia: TIPOGRAFIA DA ATLÂNTIDA - COIMBRA
19: Número de Páginas da Obra: 577
20: Número de Páginas Escaneadas da Obra: 65
21: Número de Páginas do Texto: DA P. 118 A P. 132
22: Identificador (ISBN/ISSN/DOI):
Revisão
23: Revisor(a) (responsável pela revisão da digitalização): ISABELA
Formato do Arquivo Final (txt)
24: Codificação: UTF-16LE (Unitex); UTF-8 (Philologic)
25: Data da Integração do Arquivo de Texto ao Córpus:
26: Tamanho do Texto (número de palavras ortográficas): calculado automaticamente

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<!DOCTYPE TEI.2 SYSTEM "http://docsouth.unc.edu/dtds/teixlite.dtd">
<TEI.2>
<teiHeader>
  <fileDesc>
    <titleStmt>
      <title>  CARTA DO P. MANUEL DA NÓBREGA AO P. SIMÃO RODRIGUES, BAÍA 9 DE AGOSTO
1549</title>
      <author>
        <name> P. MANUEL DA NÓBREGA</name>
        <date>  BAÍA 9 DE AGOSTO 1549</date>
      </author>
      <respStmt>
        <resp>Arquivo preparado por</resp>
        <name>varios pesquisadores do Projeto DHPB</name>
      </respStmt>
    </titleStmt>
    <publicationStmt>
      <distributor>Projeto do Dicionario Historico do Portugues do Brasil (DHPB), UNESP, Araraquara</distributor>
      <date>2006-2007</date>
    </publicationStmt>
    <sourceDesc>
      <biblStruct>
        <monogr>
          <title>  CARTA DO P. MANUEL DA NÓBREGA AO P. SIMÃO RODRIGUES, BAÍA 9 DE AGOSTO
1549</title>
          <author> P. MANUEL DA NÓBREGA</author>
          <imprint>
            <pubDate> 1956</pubDate>
          </imprint>
        </monogr>
      </biblStruct>
    </sourceDesc>
  </fileDesc>
</teiHeader>
<text>
  <body>
(...)
<p> {7. - BAÍA 9 DE AGOSTO DE 1549 127 - A00_0002.txt,.N} </p>
<p> convertidos, onde estaremos Vicente Rodriguez e eu, e hum soldado  <note place="foot"n="19">  Simão Gonçalves.
LEITE I 573.  </note> que se meteo comnosco para nos servir, e está agora em Exercicios, de que eu estou muy contente,
Faremos nossa igreja, onde insinemos os nossos novos christãos, e aos domingos e festas visitarey a Cidade e pregarey. </p>
<p> O Padre Antonio Pirez e o P.^e Navarro estaram em outras Aldeas longe, onde já lhes fazem casas. E portanto hé
necessario V. R. mandar officiaes, e am-de vir já com a paga, porque cá diz ho Governador que, ainda que venha Alvará de S.
A. para nos dar o necessario, que nom o averá hi para isto. Os officiaes que cá estão tem muito que fazer, e que o nom tenhão,
estão com grande saudade do Reyno, porque deixão lá suas molheres e filhos, e nom aceitaram a nossa obra depois que
cumprirem com S. A., e tambem ho trabalho que tem com as viandas e o mais os tira disso. Portanto me parece que avião de
vir de lá, e, se possivel fosse, com suas molheres e filhos, e alguns que fação taipas e carpinteiros. Cá está hum Mestre para as
obras, que hé hum sobrinho <note place="foot"n="20">  Este «bom oficial», sobrinho de Luís Dias, era Diogo Peres. LEITE
I 22.      </note> de Luis Diaz, mestre das obras d'El-Rey, ho qual veo con trinta mil reis de partido. Este nom hé necessario
porque abasta ho tio para as obras de S. A.; a este avião de dar o cuidado do nosso collegio; hé bom official. Serão cá muito
necessarias pessoas que teção algodão, que há muito, e outros officiaes. </p>
<p> 12 Trabalhe V. R. por virem a esta terra pessoas casadas,  porque certo hé mal empregada esta terra em degradados, que
cá fazem muyto mal, e já que cá viessem avia de ser para andarem afferrolhados nas obras de S. A. </p>
(...)
  </body>
</text>
</TEI.2>
```

Figura 1.5: Texto da Figura 1.4 convertido para o formato TEI

1.2 Motivação e relevância

Projetos de córpus históricos são uma das formas de preservação da história de um povo. As pesquisas nesses córpus resultam em diversos produtos, alguns com importância cultural e histórica como no caso do dicionário DHPB. No Brasil, iniciativas para a construção de córpus históricos são particularmente importantes, pois há poucos projetos com esse perfil.

As ferramentas computacionais são um importante auxílio na construção de córpus e de produtos derivados da pesquisa em córpus, pois fornecem aos pesquisadores um ganho de produtividade nas atividades de construção e uso do córpus, além de viabilizarem tais produtos. Com isso, os córpus podem ser construídos em um tempo menor, com um volume maior de textos e tornam-se menos suscetíveis a erros gerados em sua construção. Essas ferramentas também contribuem para a disponibilização do córpus para outros tipos de pesquisa.

Durante o desenvolvimento do projeto DHPB foi constatado que apesar da ampla gama de ferramentas computacionais disponíveis para processamento de córpus, poucas delas cobrem de maneira desejável os requisitos para a construção de córpus históricos do Português. Devido ao número de processadores de córpus existentes, torna-se difícil a escolha do conjunto de ferramentas ideal para atender às necessidades de um projeto de córpus histórico. Além disso, muitas vezes é necessário o desenvolvimento de ferramentas para atender a necessidades específicas de um projeto. Verificou-se também a necessidade da criação de ferramentas para redação de verbetes, pois as ferramentas atuais para essa tarefa são mais focadas em dicionários contemporâneos. Outra necessidade levantada foi a construção de glossários para apoiar a tarefa lexicográfica e são geralmente comerciais. Por exemplo, um glossário de abreviaturas é útil para encontrar a forma expandida de uma abreviatura desconhecida no córpus.

1.3 Objetivos

Esta pesquisa consistiu em três objetivos principais: (a) levantar as necessidades do projeto DHPB, (b) desenvolver uma metodologia e um ambiente computacional para atender essas necessidades e (c) generalizar o ambiente para uso em outros projetos de córpus históricos para o Português. Espera-se que o ambiente proposto seja útil não apenas ao projeto

DHPB, mas também a outros projetos baseados em córpus históricos em Português.

Especificamente, o objetivo deste trabalho foi fornecer recursos computacionais ao projeto DHPB, com a expectativa de aumentar a produtividade dos pesquisadores envolvidos. Para isso, glossários de apoio a tarefa lexicográfica foram criados, ferramentas de processamento de córpus existentes foram adaptadas às necessidades do projeto e novas ferramentas foram desenvolvidas. As ferramentas são utilizadas nas etapas de geração de córpus e redação de verbetes mostradas na Figura 1.1, e executam três tipos de função: (a) processamento automático (por exemplo, conversão da ficha catalográfica para etiquetas), (b) processamento semi-automático, ou seja, revisado por humanos (por exemplo, tratamento de notas de rodapé) e (c) apoio a tarefas manuais (como edição e gerenciamento de verbetes). Além do recurso do córpus, bastante ênfase foi dada a dois glossários de apoio a tarefa lexicográfica: glossário de variação de grafias e glossário de abreviaturas.

1.4 Organização da monografia

O Capítulo 2 trata das etapas para a construção e uso de córpus e da tipologia de córpus, segundo a visão de diferentes autores. O Capítulo 3 apresenta ferramentas úteis para o processamento de córpus, classificando-as de acordo com sua função. Além disso, uma análise comparativa de parte das principais ferramentas existentes é feita. O Capítulo 4 apresenta problemas comuns em projetos de córpus históricos e possíveis soluções. O Capítulo 5 contém a metodologia empregada para processamento do córpus, desenvolvimento das ferramentas propostas e geração dos glossários de apoio à tarefa lexicográfica. O Capítulo 6 faz a avaliação do córpus, das ferramentas e dos glossários. O Capítulo 7 apresenta uma generalização do ambiente utilizado no projeto DHPB para outros projetos de córpus de Português histórico. Por fim, o Capítulo 8 traz as conclusões do trabalho.

2 Projeto e compilação de córpus

2.1 Considerações iniciais

Este capítulo contrasta os tipos de córpus existentes e as várias tipologias de córpus propostas desde o trabalho pioneiro de Atkins, Clear e Ostler (1992) (Seção 2.2). Além disso, uma metodologia para construção de córpus é apresentada (Seção 2.3). A seguir, as quatro fases envolvidas na construção de córpus (projeto, compilação, anotação e uso) são detalhadas (seções 2.4, 2.5, 2.6 e 2.7, respectivamente).

2.2 Tipologia de córpus

Diferentes tipologias de córpus foram propostas. Entretanto, ainda não existe um consenso sobre a tipologia mais adequada para classificação dos diversos projetos de córpus existentes. Este trabalho apresenta algumas das principais tipologias, incluindo o trabalho de Atkins, Clear e Ostler (1992), a proposta do grupo de pesquisadores EAGLES (*Expert Advisory Group on Language Engineering Standards*) (SINCLAIR, 1996), a tipologia de Sardinha (2004) e o trabalho de Giouli e Piperidis (2007).

No trabalho proposto por Atkins, Clear e Ostler (1992), os córpus podem ser classificados como:

- Com Textos completos, com amostras ou monitor: referente à forma pela qual os textos são incluídos no córpus. Nos primeiros, os textos são incluídos na íntegra, enquanto que, nos segundos, são incluídas apenas amostras dos textos originais. Já em córpus do tipo monitor, os textos são incluídos e removidos dinamicamente. Estes últimos são utilizados para detecção de modificações na língua estudada.
- Finalizado ou em construção: relativo a situação da fase de compilação do córpus está (encerrada ou construção).
- Sincrônicos ou diacrônicos: tipo em que é essencial o período histórico de que procedem os textos. Nos córpus sincrônicos, os textos pertencem a um intervalo de tempo bem delimitado. Os córpus diacrônicos englobam diferentes períodos de tempo.
- Gerais ou terminológicos: define a abrangência dos textos (pertencentes à língua geral

ou a um domínio de conhecimento específico). Córpus de domínios específicos são utilizados por terminólogos para tarefas como a construção de dicionários terminológicos.

- Monolíngües, bilíngües ou plurilíngües: refere-se ao número de idiomas presentes no córpus. Um córpus pode ser constituído por uma língua (monolíngüe), duas (bilíngüe) ou várias (plurilíngüe). Tais córpus podem ser constituídos apenas por textos originais ou por originais e suas traduções. Córpus monolíngües podem ainda ser subdivididos em de língua geral e de língua regional (ou de dialeto).
- Simples, 2-paralelo, n-paralelo: refere-se ao número de traduções para cada texto. Não se aplica a córpus monolíngües.
- Central ou exterior[7]: um córpus central é constituído por textos coletados para pesquisa geral. Um córpus exterior contém textos extras. Estes não foram incluídos no córpus central para não afetar seu balanceamento, mas podem servir para fins de pesquisa específicos.
- Núcleo ou periferia: refere-se aos tipos textuais incluídos no córpus. O núcleo contém tipos gerais da língua, enquanto que a periferia contém tipos textuais de um subcórpus específico.

Para Sinclair (1996), existem quatro tipos principais de córpus:

- De referência: projetado para fornecer informação abrangente e vasta sobre a língua. Um córpus de referência deve representar todas as variantes/dialetos de uma língua e pode ser aplicado para criação de gramáticas, dicionários, tesauros e outros materiais de referência. Quatro parâmetros são propostos para analisar o equilíbrio de textos no córpus: formalidade, espontaneidade, modo (falado ou escrito) e assunto. Um córpus é balanceado quando possui textos que assumem diversos valores para cada parâmetro.
- Monitor: textos podem ser adicionados e removidos dinamicamente do córpus (semelhante à definição de Atkins). Permite a identificação de palavras novas e de evolução no uso da língua. Um problema com esse tipo de córpus são as dificuldades para comparação de experimentos, pois podem fornecer resultados diferentes no

7 do original "*shell*"

decorrer do tempo.

- Comparáveis: possuem textos similares em diversos idiomas ou dialetos.
- Paralelos: possuem textos em diversas línguas e versando sobre a mesma matéria de forma que cada texto apareça em sua versão original e, a seguir, traduzido para os demais idiomas.

Quatro características são propostas para análise de córpus: quantidade, qualidade, simplicidade e documentação. Existe um valor esperado que um córpus deve apresentar para cada uma das características. Com relação à quantidade, um córpus deve apresentar o tamanho grande, ou seja, possuir um número de textos tão grande quanto possível. Com relação à qualidade, um córpus deve ser autêntico, ou seja, os textos devem ser realizações verbais ou escritas genuínas dos falantes da língua, sem interferência do compilador do córpus. Com relação à simplicidade, um córpus deve ser em texto puro (sem anotação). Com relação à documentação, um córpus deve ser documentado com informações referentes à sua procedência. A documentação pode ser separada do texto original ou feita no cabeçalho do arquivo.

Um córpus é considerado especial quando assume valores diferentes dos apresentados acima para as uma ou mais das quatro características. Córpus especiais não contribuem para a descrição da língua geral por conterem uma proporção incomum de recursos pouco usuais. Por exemplo, córpus de textos de crianças, de falantes não nativos ou de usuários de dialetos incomuns são considerados especiais.

Para Sardinha (2004), os córpus podem ser classificados pelos critérios:

- Modo: falado e escrito.
- Tempo: sincrônico (um único período de tempo), diacrônico (vários períodos de tempo), contemporâneo (tempo corrente) e histórico (tempo passado).
- Seleção: de amostragem, monitor, estático, dinâmico e equilibrado. Córpus de amostragem são projetados para representar uma amostra da língua e são compostos por porções de textos. Córpus monitores mudam sua composição para refletir o estado atual de uma língua (em contraste com córpus de amostragem). Em córpus estáticos, não são permitidos acréscimo e remoção de textos, o que caracteriza os córpus de amostragem. Em córpus dinâmicos, a alteração no número de textos é

permitida, o que caracteriza os córpus monitores. Por fim, em córpus equilibrados, os componentes (por exemplo, gêneros e textos) são distribuídos em quantidades semelhantes.

- Conteúdo: especializado (textos de tipos específicos), regional/dialetal (textos de uma ou mais variedades sociolingüísticas específicas) e multilingüe (idiomas diferentes)
- Autoria: de aprendiz (falantes não nativos) e de falantes nativos da língua.
- Disposição interna: paralelo (textos similares ou originais e traduções) e alinhado (traduções são alinhadas sentença a sentença com o original). O critério de disposição interna se aplica apenas a córpus multilingües.
- Finalidade: de estudo (córpus que se pretende descrever), de referência (para contraste com o córpus de estudo) e de treinamento/teste (usado para o desenvolvimento de aplicações ou ferramentas de PLN).

A classificação de Giouli e Piperidis (2007) é mais completa e abrange muitos itens levantados por outros autores. Um córpus pode ser classificado segundo os critérios:

- Modalidade: córpus falado (em formato de áudio), escrito ou multímodo (falado e escrito).
- Tipos de texto: córpus falado (fala transcrita para texto), escrito ou híbrido.
- Mídia: relativo a mídia original de publicação do texto (por exemplo: córpus de livro ou de periódico).
- Cobertura da língua: córpus de língua geral ou de sub-língua. Córpus de sub-língua cobrem uma variedade particular do idioma como um dialeto ou um assunto específico (por exemplo, o domínio financeiro).
- Gênero: córpus literário, técnico, não-ficção ou híbrido. Este critério está relacionado aos gêneros presentes no córpus. Os gêneros serão discutidos na Seção 4.6.1.
- Quantidade de línguas: córpus monolíngües ou plurilíngües. Córpus plurilíngües podem ainda ser classificados em córpus de tradução (originais e traduções), córpus paralelos (semelhante a córpus de tradução, mas o original não precisa estar presente) ou córpus comparáveis (textos originais similares em idiomas diferentes).
- Comunidade produtora: córpus de falantes nativos ou de aprendizes.

- Anotação: córpus sem anotação (texto cru) ou anotado.
- Mutabilidade[8]: córpus fechado (estático) ou monitor.

Giouli e Piperidis (2007) também propõem a classificação de córpus pelos seguintes critérios:

- Variedade nacional: relativo à nacionalidade dos autores dos textos do córpus (por exemplo: córpus britânico, americano ou internacional).
- Variações históricas: córpus sincrônicos, diacrônicos ou que cobrem apenas um estágio da história da língua.
- Variações geográficas ou de dialetos: córpus de dialetos ou mistos.
- Idade dos autores: córpus de adultos ou de crianças.
- Disponibilidade: córpus comercial ou não comerciais, disponíveis via *Web*, servidores FTP, disquetes ou CD-ROMs.

2.3 Etapas na construção e uso de córpus

A vida útil de um córpus pode ser dividida em quatro etapas: projeto, compilação, anotação e uso. A etapa de projeto consiste na definição dos objetivos do córpus e na tomada de decisões a respeito de sua constituição. A etapa de compilação envolve a estratégia de coleta de textos, conversão para o formato digital (caso ainda não estejam) e pré-processamento desses textos. Na etapa de anotação (opcional), os metadados dos textos (por exemplo, informações estruturais de parágrafos e capítulos ou informações lingüísticas nos níveis morfossintático e sintático) são identificados e anotados para uso em ferramentas de processamento de córpus. Por fim, o córpus é então usado para as pesquisas para as quais foi originalmente concebido.

As quatro etapas não são totalmente independentes, pois dificuldades na coleta de textos podem implicar em mudanças no projeto e novas necessidades de uso podem implicar em mudanças na compilação e na anotação dos textos. Atkins, Clear e Ostler (1992) recomendam que o processo de construção do córpus seja feito iterativamente de forma a obter um córpus o mais balanceado possível. Biber (1993a) compartilha da mesma idéia, conforme ilustrado na

8 Do original *open-endedness*

Figura 2.1.

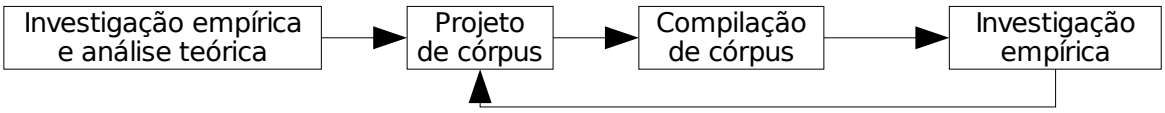

Figura 2.1: Construção iterativa de córpus (BIBER, 1993a)

2.4 Projeto

As decisões de projeto do córpus estão diretamente relacionadas com os objetivos de pesquisa que o córpus deve atender. Ide e Brew (2000) colocam a reusabilidade e a extensibilidade como dois aspectos a serem considerados em projetos de córpus. A reusabilidade é característica de um córpus ser usável em mais de um projeto de pesquisa e por mais de um grupo de pesquisadores. A extensibilidade é a capacidade de o córpus ser melhorado em várias direções, por exemplo, com o acréscimo de um nível a mais de análise lingüística.

Aluísio e Almeida (2006) levantam algumas questões iniciais de projeto, entre elas o modo do córpus: falado, escrito ou híbrido. No último caso, é necessário definir também a proporção de textos falados e escritos, uma vez que a compilação de textos falados é cara. Kennedy (1998) apresenta três importantes decisões que devem ser tomadas durante a fase de projeto: a forma de inclusão de textos (estaticamente ou dinamicamente), o tamanho do córpus e seu balanceamento e representatividade.

Córpus estáticos são mais simples de serem construídos e se aplicam a diversas pesquisas (por exemplo, fonologia, lexicografia, morfologia e análise sintática). Córpus dinâmicos (ou monitores) podem ser usados para as mesmas pesquisas aplicadas a córpus estáticos e, além disso, são particularmente úteis em pesquisas diacrônicas. Por exemplo, podem ser utilizados para detecção de neologismos na língua e de palavras que caíram em desuso. A etapa de compilação de córpus dinâmicos é permanente, uma vez que os textos que compõem o córpus estão continuamente mudando.

O tamanho do córpus é uma decisão importante, pois implica diretamente no custo de compilação dos textos. Kennedy (1998) afirma que:

> "(...) qualquer córpus, não importa o quão grande seja, não pode representar mais do que uma minúscula amostra de toda fala e escrita emitida ou recebida pelos falantes

da língua em um único dia[9]. ”

Esse fato pode ser observado em córpus com um milhão de palavras, nos quais de 40 a 50% das formas aparecem uma única vez, o que dificulta pesquisas lexicográficas ou terminológicas. Embora Sinclair (1991) sugira um córpus com tamanho mínimo entre 10 a 20 milhões de palavras, o tipo de pesquisa realizada deve ser levado em conta. Por exemplo, para o estudo da prosódia, um córpus com 100 mil palavras é mais do que o suficiente (KENNEDY, 1998, p. 68). Já para pesquisas lexicográficas, é desejável córpus de grandes dimensões. Por exemplo, o córpus monitor *Bank of English*, criado para tarefas lexicográficas, possuía 450 milhões de palavras em 2004.

Outros fatores importantes são o balanceamento e a representatividade do córpus. Aluísio e Almeida (2006) definem um córpus como balanceado quando possui:

> “(...) um equilíbrio entre gêneros discursivos (informativo, científico, religioso, etc), ou de tipos de textos (artigo, editorial, entrevista, dissertação, carta, etc.), ou de títulos, ou de autores, ou de todos esses itens juntos (...).”

Biber (1993a, p. 243) define representatividade como:

> “ Representatividade refere-se à extensão sobre a qual uma amostra inclui todo o intervalo de variação da população. No projeto de córpus, a variação pode ser considerada de perspectivas situacionais e lingüísticas, ambas importantes para definir a representatividade. Dessa forma, um projeto de córpus deve avaliar a extensão, o que inclui: (1) o intervalo de textos em uma língua e (2) o intervalo de distribuição lingüística em uma língua[10].”

Representatividade e balanceamento são fatores que, quando ausentes, podem comprometer a pesquisa sobre o córpus. Por conseguinte, esses critérios são mais importantes que o tamanho do córpus em si.

A Estatística fornece subsídios para a definição de amostras populacionais representativas e balanceadas para diversas áreas de conhecimento. Entretanto, essa ciência não tem sido aplicada ostensivamente na Lingüística de Córpus devido à dificuldade em definir e estimar a população que constitui a língua. A população pode ser definida, por

9 Do original: “(...) *any corpus, however big, can never be more than a minuscule sample of all the speech or writing produced or received by all of the users of a major language on even a single day*”.

10 Do original: “*Representativeness refers to the extent to which a sample includes the full range of variability in a population. In corpus design, variability can be considered from situational and from linguistic perspectvies, and both of these are important in determining representativeness. Thus a corpus design can be evaluated for the extent to which it includes: (1) the range of text types in a language, and (2) the range of linguistic distributions in a language*”.

exemplo, como o número total de palavras, de sentenças ou de textos da língua. Mas não é possível definir o tamanho ideal de um córpus (ou amostra da língua), pois não é possível fazer uma estimativa precisa sobre a quantidade de textos produzidos por todos os falantes da língua em um determinado período de tempo. Mesmo que a população pudesse ser estimada corretamente, seu tamanho implicaria em córpus excessivamente grandes para as técnicas de compilação atuais. Devido a esses problemas, é sempre possível mostrar que um dado córpus não é representativo para alguma característica lingüística (ATKINS; CLEAR; OSTLER, 1992, p. 4).

Os pontos levantados mostram que não é possível construir um córpus completamente representativo e balanceado. Cabe então ao projetista do córpus selecionar cuidadosamente os textos de forma que estes tornem o córpus o mais balanceado e representativo possível.

2.5 Compilação

É possível definir três etapas de compilação do córpus: a obtenção de permissão de uso de textos protegidos por direitos autorais, a coleta dos textos e a limpeza.

A obtenção de **permissão de uso** é uma etapa não-técnica e geralmente trabalhosa, dado que um córpus pode ser constituído por textos de diversos autores. Dificuldades na obtenção de permissão de uso podem acarretar mudanças no projeto de córpus. Uma medida que pode minimizar o número de permissões de uso necessárias para a compilação do córpus é a utilização de muitos textos de poucos autores. Contudo, vale ressaltar que essa medida pode afetar negativamente o balanceamento do córpus. Em textos históricos, geralmente, os direitos autorais já expiraram. Entretanto, muitas vezes o projetista do córpus precisa usar uma versão editada do texto (e protegida por direitos autorais) por não ter acesso aos textos originais.

Kennedy (1998) apresenta estratégias para a **coleta de textos** falados e escritos.

Para textos falados, a coleta envolve a obtenção dos textos e sua transcrição para o formato eletrônico. O compilador do córpus pode ser o responsável direto pela obtenção dos textos através de equipamento próprio como gravadores de áudio e vídeo. Nesse caso, é recomendável o uso de equipamento digital em detrimento a equipamentos analógicos. No caso de córpus com textos de comunicação em massa (como programas de rádio ou televisão) é possível obter material de qualidade diretamente com os responsáveis pela transmissão de tais programas. Uma hora de texto falado contém cerca de 7 a 9 mil palavras e sua transcrição deve demorar cerca de 10 a 25 horas dependendo do nível de anotação utilizado. A transcrição

pode ser dificultada por trechos inaudíveis causados por fatores como ruídos no ambiente, problemas na mídia de gravação e conversas paralelas. Além disso, o transcritor pode ter dificuldade para conhecer a grafia correta de nomes próprios (principalmente de estrangeiros).

Para textos escritos, as possíveis estratégias de captura são: digitação, digitalização e processamento de textos eletrônicos (*parsing*). A digitação é útil para coleta de manuscritos ou de textos impressos com má qualidade ou rasuras. Um digitador médio é capaz de digitar cerca de 10 mil palavras ao dia.

O processo de digitalização envolve o uso das técnicas de OCR (*Optical Character Recognition*). Esse processo é mais rápido que a digitação, mas não se aplica eficientemente a documentos com muitas rasuras, além de ser inviável em manuscritos. Além disso, o processo não é isento de falhas e uma revisão ortográfica manual faz-se necessária. A revisão automática feita por ferramentas como o *MS Word* não é totalmente confiável. Por exemplo, a palavra "mato" pode ser incorretamente reconhecida como "rato", situação não detectada via revisão ortográfica automática. Além disso, em córpus históricos há abundância de variações de grafia das palavras (como "pharmacia") e a presença de palavras que caíram em desuso. Em ambos os casos, a revisão ortográfica não pode ser aplicada. Alguns erros comuns na digitalização de textos em inglês são a troca de "o" por "a", "m" por "in", "ni" ou "ir", "c" por "e", "ij" por "y", entre outros. Alguns desses erros também ocorreram durante a digitalização de dos textos do projeto DHPB (como a troca de "c" por "e"). Além disso, houve trocas de "0" (zero) por "O" (o maiúsculo), e de "1" (um) por "l" (L minúsculo) ou por "I" (i maiúsculo). Outro problema durante a digitalização foi a formatação dos textos históricos, pois estes contêm muitas ocorrências de abreviaturas com sobrescrito (como em "sr.º") que muitas vezes não são reconhecidas pelo software de digitalização.

Outra possibilidade para a coleta é o acesso a versões eletrônicas dos textos que constituirão o córpus, por exemplo, textos disponíveis via Web ou versões eletrônicas liberadas pela editora em acordos de permissões de uso. A compilação de textos em formato eletrônico é mais rápida que o uso de OCR. Nesses casos, o texto geralmente encontra-se com formatação (com negrito, itálico, variações de fonte, etc) e pode conter anotação estrutural (como marcadores de capítulos, seções, etc). Dessa forma, esses textos possuem particularidades que precisam ser tratadas durante a etapa de limpeza dos textos.

A **limpeza** envolve o tratamento de dados pessoais, de metadados[11] e de formatação.

11 Dados estruturados sobre dados

Dados pessoais como nome ou endereço para correspondência podem estar presentes em determinados tipos de córpus (por exemplo: córpus de redações de alunos do ensino fundamental). Esse é um procedimento não-técnico em que dados pessoais são removidos do córpus para preservar a privacidade dos autores dos textos. Metadados estão presentes em praticamente todos os tipos textuais e podem interferir na pesquisa realizada sobre o córpus. Por exemplo, em textos extraídos de livros é comum a presença de títulos de capítulos mostrados página a página (gerando distorções na contagem de freqüência de algumas palavras) e notas do editor (gerando distorções no estudo de estilo de um determinado autor). Em (WYNNE, 2005), os metadados são agrupados em quatro categorias: (a) administrativos (informações sobre a compilação do córpus), (b) editoriais (informações sobre a edição do texto), (c) analíticos (unidades de textos como parágrafos e capítulos e informações lingüísticas) e (d) descritivos (informações sobre o contexto social dos textos).

O tratamento feito para metadados consiste na remoção ou na anotação desses recursos, de forma que possam ser corretamente analisados por processadores de córpus. A mesma estratégia se aplica à formatação. Para textos que já contenham alguma estratégia de anotação estrutural (por exemplo, documentos HTML – *HyperText Markup Language*), é possível converter diretamente suas etiquetas para o padrão de anotação utilizado no córpus. O trabalho de limpeza deve ser menor em textos digitados ou transcritos, pois os elementos de texto indesejados podem ser descartados pelo digitador/transcritor.

Uma decisão importante para a compilação do córpus é a escolha da codificação de caracteres utilizada. Algumas das codificações de caracteres são discutidas na Seção 2.5.1.

2.5.1 Codificação de caracteres

A codificação de caracteres define a representação computacional na qual o córpus é convertido durante sua digitalização. Basicamente, uma codificação consiste de um conjunto composto de representações visuais de símbolos (por exemplo, as letras do alfabeto romano) e por códigos associados a esses símbolos. Os símbolos precisam ser convertidos para seus respectivos códigos para armazenamento e processamento em sistemas computacionais, uma vez que estes são capazes apenas de processar informação em formato numérico. Por exemplo, o símbolo “A” pode ser associado ao decimal 64 ou ao hexadecimal 40. Por conveniência, usa-se códigos hexadecimais ao invés de decimais.

A primeira codificação de caracteres amplamente usada foi a ASCII (*American*

Standard Code for Information Interchange), uma proposta criada para a unificação de representação de informação em diversos sistemas computacionais. Contudo, a codificação engloba poucos idiomas (em sua maioria, similares ao inglês), pois apenas 128 símbolos são permitidos. Então novas codificações foram criadas para atender a outros idiomas, diferindo entre si pelo conjunto de símbolos permitidos e pelo código associado a cada símbolo. Exemplos de codificações são o padrão ISO-8859-1 (para idiomas diversos idiomas ocidentais) e ISO-8859-3 (para os idiomas Turco, Maltês e Esperanto).

A escolha da codificação de caracteres é importante para o projeto de córpus, pois define os símbolos que poderão ser codificados. Além disso, também é importante para o desenvolvimento de ferramentas computacionais para compilação e processamento de córpus, pois define quais linguagens poderão ser processadas.

O *Unicode* (UNICODE CONSORTIUM, 2006) é um esforço para a criação de um padrão que compatível com todos os idiomas contemporâneos. Entre os alfabetos permitidos estão o romano, o árabe e diversos alfabetos asiáticos. Além disso, alfabetos de línguas que caíram em desuso também são permitidos (por exemplo, o Hebraico). O *Unicode* define diferentes versões de codificação. O mesmo conjunto de caracteres é permitido em todas. Entretanto, de uma versão para outra, a representação em formato digital de cada caractere pode variar. As versões de codificação são:

- UTF-7: codificação de tamanho variável, na qual símbolos ASCII possuem representação de 7 *bits*.
- UTF-8: codificação de tamanho variável, com tamanho mínimo de 8 *bits*. O UTF-8 pode gerar economia de espaço para representar documentos nos quais a maior parte dos caracteres são ocidentais. Essa é a codificação *Unicode* mais utilizada na *Web*.
- UTF-16: define códigos de tamanho variável, com tamanho mínimo de 16 *bits*. A maior parte dos símbolos *Unicode* pode ser representada em 16 *bits*. Possui duas versões: *little-endian* e *big-endian*, que diferem pela ordem na qual os dígitos são armazenados. Em *little-endian*, o símbolo “A” é representado como “00 40”. Já em *big-endian* sua representação é “40 00” (pode ser processado mais rapidamente em algumas arquiteturas computacionais).
- UCS-2: define códigos de tamanho fixo de 16 *bits*. Esse padrão é considerado obsoleto, pois não é possível representar todos os símbolos *Unicode* com apenas 16

bits.

- UTF-32 (UCS-4): representação fixa de 32 *bits*.

O utilizador pode se referir a um símbolo pelo seu código universal. Por exemplo, é suficiente referir-se ao símbolo "A" pelo código 0040. O sistema computacional converte automaticamente esse código para a codificação utilizada, como 40 no caso de UTF-8 ou 00000040 no caso de UTF-32. No caso do símbolo "A", a conversão é imediata, mas para outros símbolos, os códigos entre uma codificação e outra podem ser completamente distintos. Por exemplo, o símbolo "Ã" pode ser associado aos códigos C383 (UTF-8) e 00C3 (UTF-16).

2.6 Anotação

A anotação envolve a inserção de etiquetas nos textos para preservação dos quatro tipos de metadados levantados por (WYNNE, 2005) (administrativos, editorias, analíticos e descritivos). Exemplos de metadados **administrativos** incluem a tipologia textual utilizada para a classificação dos textos e informações sobre catalogação do córpus.

Metadados **editoriais** incluem informações sobre mudanças no texto durante a compilação do córpus. Três eventos podem ocorrer durante a compilação do córpus: inclusão, remoção e alteração de informações. Esses eventos precisam ser documentados. Por exemplo, etiquetas podem ser inseridas para denotar blocos de texto ilegíveis no documento original por motivos como a presença de rasuras (remoção de informações).

Metadados **analíticos** variam de acordo com o modo do texto (falado ou escrito). Para textos escritos é possível citar capítulos, parágrafos, títulos, notas de rodapé, tabelas, figuras, sentenças, citações, palavras, abreviações, referências, hifenização, ênfase e formatação em geral. Para textos falados, é possível citar mudanças de orador, interrupções, sobreposições de fala, pausas, entonação das palavras, sons não funcionais (como riso ou tosse), entre outros. Os metadados analíticos também incluem informações nos diversos níveis lingüísticos (por exemplo, morfossintático, sintático, semântico e discursivo). A anotação dessas informações pode ser realizada de três formas: manualmente (por lingüistas), automaticamente (por ferramentas de PLN) ou semi-automaticamente (correção manual da saída de outras ferramentas). A vantagem desta última é que a revisão permite eliminar o erro das ferramentas e é mais rápida que anotar pela primeira vez.

Os metadados **descritivos** contém informações bibliográficas sobre os textos e sobre o contexto social no qual foram produzidos. Alguns exemplos de metadados descritivos incluem o nome do autor, data de criação do texto, o nome da editora (para o caso de impressos), entre outros.

As etiquetas podem ser processadas computacionalmente e permitem a realização de pesquisas no córpus sobre os metadados. Por exemplo, é possível buscar textos de um dado autor, de um certo período de tempo ou que possuam capítulos ou seções com nomes específicos. Para que o córpus seja reusado em diversas pesquisas e processado por diferentes ferramentas computacionais é desejável que o córpus seja anotado com o máximo de informações possível. Entretanto, anotar um córpus pode ser um procedimento caro, dependendo do nível de detalhamento desejado. Faz-se necessária então uma análise de custo-benefício para determinar o nível de detalhamento utilizado.

Uma grande parte dos padrões de anotação são baseados em XML (*eXtensible Markup Language*) (W3C CONSORTIUM, 2006). XML é um subconjunto da linguagem de marcação SGML (*Standard Generalized Markup Language*), e tem se popularizado por ser mais simples que SGML, sem perder o poder de representação de dados. Um formato baseado em XML pode ser validado através dos padrões XML-Schema e de DTD (*Document Type Definition*). O primeiro é baseado em XML e o segundo é uma gramática para a definição da estrutura de um documento. Um documento baseado em um formato XML pode ser convertido para outro formato XML através da linguagem de marcação XSLT (*XSL Transformations*). A linguagem XSLT pertence à família de linguagens XSL (*eXtensible Stylesheet Language*).

Diversos padrões de anotação internacionais foram propostos para a anotação de córpus. Em (IDE, 2006) são levantados requisitos para a construção de uma infra-estrutura de anotação de córpus robusta e flexível. Alguns padrões e esforços de unificação de padrões são:

- CES (*Corpus Encoding Standard*) (IDE, 1998): um padrão de anotação baseado em SGML projetado para pesquisas em aplicação e engenharia de linguagem. O padrão foi criado pelo grupo EAGLES.
- CDIF (*Corpus Document Interchange Format*) (BURNAGE; DUNLOP, 1992): padrão de anotação SGML utilizado na compilação do córpus do projeto BNC.
- *Much.more*: padrão de anotação em vários níveis lingüísticos baseado em XML e utilizado no córpus *Muchmore* de textos do domínio da medicina.

- TIGER-XML: um formato XML de interface para o qual é possível converter outros padrões de anotação. Textos anotados nesse formato podem ser processados através da linguagem TIGER.
- ISLE: conjunto de padrões para uma infra-estrutura internacional para anotação em projetos de córpus.
- OLIF (*Open Lexicon Interchange Format*): padrão baseado em XML para as anotações lexicográfica e terminológica em córpus.
- SALT (*Standards-based Access to Lexicographical and Terminological multilingual resources*): um conjunto de padrões para anotações lexicográfica e terminológica.

Notadamente, os padrões baseados em XML têm se destacado graças à flexibilidade proporcionada pela linguagem. Devido ao grande número de padrões de anotação, esforços têm sido feitos para a sua unificação, incluindo a criação de um padrão ISO/TC (IDE; ROMARY, 2004) .

Os padrões TEI (TEI CONSORTIUM, 2006) e XCES (XML CES) (IDE; BONHOMME; ROMARY, 2000) são particularmente importantes, pois são utilizados em grandes córpus. O TEI é usado no córpus francês FRANTEXT (com 114 milhões de palavras) o padrão XCES é usado no ANC. Além disso, ambos são internacionalmente aceitos, o que facilita a criação de córpus reusáveis por diferentes grupos de pesquisas e tratáveis por diversas ferramentas de manipulação de córpus. Um comparativo entre os dois padrões realizado durante a construção de um córpus para a língua sueca falada pode ser encontrado em (GRÖNQVIST, 2003).

2.6.1 O padrão TEI

O TEI surgiu em 1987, durante uma conferência patrocinada pela ACH (*Association for Computers and the Humanities*) na instituição americana NEH (*National Endowment for the Humanities*) e seu primeiro rascunho foi lançado em 1990. A partir daí, diversas versões foram disponibilizadas incluindo as versões P1, P2, P3, P4 e P5. Existe também a versão simplificada de nome *Lite*, na qual as possibilidades de anotação são mais reduzidas, permitindo a criação de ferramentas computacionais mais simples. O padrão possui muitas semelhanças com o CDIF e pode ser utilizado não só para a codificação de córpus, mas também a criação de dicionários, enciclopédias e outros recursos lingüísticos.

O padrão permite as codificações de caracteres *Unicode* e ISO 646 e é compatível tanto com XML, quanto com SGML. Apesar disso, é mais comum o seu uso com codificação *Unicode* e anotação XML. É possível codificar uma vasta gama de metadados a respeito do córpus. Entretanto, poucos são obrigatórios, ficando a critério do usuário decidir efetivamente quais serão utilizados na compilação do córpus. O formato de um documento TEI é descrito por um conjunto de DTDs, no qual a maior parte dos elementos são opcionais. Os elementos são classificados em cinco níveis: (a) requerido, (b) mandatório quando aplicável, (c) recomendado, (d) recomendado quando aplicável e (e) opcional. O nível requerido define elementos que sempre deverão estar presentes na anotação, ao passo que no nível mandatório quando aplicável, os elementos são requeridos em determinadas condições. Elementos definidos nos níveis recomendado e recomendado quando aplicável são opcionais, embora seu uso seja importante. Por fim, elementos opcionais podem estar presentes ou não, de acordo com a preferência do pesquisador.

Diversos tipos de textos são permitidos. Por exemplo, existem esquemas de anotação diferenciados para prosa, verso, drama, conversação, entre outros. O conjunto de etiquetas XML é bem completo, incluindo etiquetas para indicar contrações de palavras, abreviaturas, notas do rodapé, problemas de legibilidade no texto original, entre outros. Além disso, é possível preservar a formatação do texto original com as etiquetas adequadas. Um arquivo TEI é dividido em duas partes: cabeçalho e corpo. No cabeçalho estão contidas informações como o autor do texto e a editora. O corpo contém o texto propriamente dito e informações extras como bibliografia e apêndices. A Figura 2.2 ilustra um cabeçalho TEI mínimo.

```
<teiHeader>
 <fileDesc>
  <titleStmt>
   <title>Thomas Paine: Common sense, a
    machine-readable transcript</title>
   <respStmt>
    <resp>compiled by</resp>
    <name>Jon K Adams</name>
   </respStmt>
  </titleStmt>
  <publicationStmt>
   <distributor>Oxford Text Archive</distributor>
  </publicationStmt>
  <sourceDesc>
   <bibl>The complete writings of Thomas Paine, collected and edited
         by Phillip S. Foner (New York, Citadel Press, 1945)</bibl>
  </sourceDesc>
 </fileDesc>
</teiHeader>
```

Figura 2.2: Exemplo de cabeçalho TEI (TEI CONSORTIUM, 2006)

2.6.2 O padrão XCES

O padrão XCES é uma evolução do CES. Durante a elaboração do CES, foi levado em conta as especificações do projeto TEI para a codificação e intercâmbio de textos eletrônicos. Dessa forma, é possível migrar textos entre os padrões TEI para XCES de forma relativamente simples, através da linguagem de transformação XLST. A principal diferença entre o XCES e o CES reside no fato de o primeiro empregar codificação em XML, enquanto que o segundo utiliza codificação em SGML.

O XCES é capaz de codificar os mesmos metadados codificados no padrão TEI (exemplo: informações morfossintáticas). De forma análoga ao TEI, poucos campos são obrigatórios, ficando a cargo do utilizador decidir quais metadados serão utilizados na compilação do córpus. Além disso, O XCES tem como vantagem o uso (opcional) de anotação *stand-off*, na qual o texto e as marcações em XML são armazenados em arquivos separados. Essa anotação torna os textos mais limpos, uma vez que seus metadados são armazenados em arquivos à parte. Atualmente, existe pouca documentação sobre o XCES disponível na *Web*. É possível encontrar a especificação completa do padrão, mas guias de uso são raros. Provavelmente, isso se deve ao fato de o XCES ainda estar em estágio inicial de desenvolvimento (a versão atual é a 0.2). Entretanto, os *schemas* XCES são publicamente disponíveis e parte da documentação do CES pode ser consultada, devido à similaridade entre

os dois padrões. Um exemplo de cabeçalho XCES é mostrado na Figura 2.3.

```
<xcesHeader
xlink:href="exampleHeader.xml"/>
<cesDoc version="1.0">
 <cesHeader version="1.0">
  <fileDesc>
   <titleStmt>
    <h.title>English Sample</h.title>
   </titleStmt>
   <publicationStmt>
    <distributor>ANC</distributor>
    <pubAddress>Vassar</pubAddress>
    <availability>Free</availability>
    <pubDate>2002</pubDate>
   </publicationStmt>
   <sourceDesc>
    <biblStruct>
     <monogr>
      <h.title>English Sample</h.title>
      <imprint/>
     </monogr>
    </biblStruct>
   </sourceDesc>
  </fileDesc>
 </cesHeader>
<!-- (...) -->
</cesDoc>
```

Figura 2.3: Exemplo de cabeçalho XCES
(VASSAR COLLEGE, 2006)

Um exemplo de projeto que utiliza XCES é o PLN-BR, que utiliza anotação *stand-off* para os níveis lógico (por exemplo: capítulos e parágrafos) e de segmentos (sentenças). Adicionalmente, há anotação tradicional, na qual o texto original é mesclado com os arquivos de anotação. Um exemplo de arquivo anotado pertencente ao projeto PLN-BR em *stand-off* pode ser encontrado no Anexo A.

2.7 Uso de córpus

Atkins, Clear e Ostler (1992) levantaram alguns campos de pesquisa nos quais os córpus são de grande valia. Um córpus é capaz de atender a profissionais com diferentes perfis, entre eles:

- Especialistas em linguagem: incluindo lexicógrafos e terminólogos em pesquisas para a elaboração de dicionários, tradutores em pesquisas empíricas sobre equivalência de palavras em diferentes línguas, e lingüistas em geral (teóricos ou aplicados).
- Especialistas em conteúdo: incluindo historiadores interessados em análises de córpus

diacrônicos, críticos literários em pesquisas sobre estilometria e sociólogos em pesquisas sobre as opiniões da sociedade e seus subgrupos acerca de um dado tema.

- Especialistas em mídia: incluindo profissionais que atuam no processamento automático de textos em áreas como Recuperação de Informações (RI), sumarização automática, tradução de máquina e projeto e implementação de sistemas de processamento de córpus.

McEnery e Wilson (1996) indicam as seguintes pesquisas relacionadas à Lingüística de Córpus: compilação de córpus, desenvolvimento de ferramentas, descrição da língua e aplicação de córpus (por exemplo: ensino, tradução e reconhecimento de voz). Para o estudo da língua, os córpus podem ser utilizados para as sub-áreas: estudo da fala, lexicografia, gramática, semântica, estudos pragmáticos, sociolingüística, estilometria (estudo dos estilos), ensino da língua, estudo dos dialetos, psicolingüística, estudos culturais e psicologia social. Os autores também citam aplicações de córpus para o Processamento de Língua Natural, nas sub-áreas de etiquetação, análise sintática, correção gramatical, análise do discurso (anáforas), análise retórica, tradução automática, sumarização automática e extração automática de terminologia.

3 Sistemas de processamento de córpus

3.1 Considerações iniciais

Este capítulo aborda os tipos de ferramentas para lingüística de córpus (Seção 3.2) e os principais processadores de córpus existentes (Seção 3.3). Uma ênfase é dada nos processadores de código aberto, pois como seu código fonte é disponibilizado e sua distribuição é livre, torna-se mais fácil adaptá-los para suprir as necessidades de projetos de córpus. Adicionalmente, um comparativo entre cinco processadores de córpus (GATE, *Philologic*, *Tenka*, *Unitex* e *Xaira*) é apresentado (Seção 3.4).

3.2 Tipos de ferramentas de trabalho com córpus

Um córpus pode demandar um conjunto de ferramentas diversificado, tanto para sua criação quanto para sua manipulação. Neste trabalho, as ferramentas são agrupadas em quatro grandes categorias, de acordo com a etapa de construção do córpus na qual são usadas: compilação de textos, anotação, acesso a córpus e extração de conhecimento. As duas primeiras apóiam às tarefas descritas nas Seções 2.5 e 2.6, respectivamente. As duas últimas estão relacionadas às tarefas descritas na Seção 2.7.

Geralmente, uma ferramenta pode ser classificada em duas ou mais das categorias apresentadas a seguir. Neste trabalho, essas ferramentas são denominadas de sistemas de processamento de córpus (ou processadores de córpus). Esses processadores, por sua vez, podem ser combinados, formando um ambiente de processamento de córpus. A seguir, os principais tipos de ferramentas serão listados.

Ferramentas para **compilação** de textos:

- Digitalizadores (*scanning software*): digitalizam textos impressos, convertendo-os para o formato digital (mais especificamente, para o formato de imagem). Geralmente, as ferramentas de digitalização são vendidas juntamente com equipamento para digitalização (*scanners*). Um exemplo de digitalizador é o *Microsoft Photo Editor* distribuído juntamente com o *MS-Office 2002*.
- Reconhecedores de caracteres ópticos (*OCR software*): convertem documentos em formato de imagem para o formato texto (com ou sem formatação) a partir da técnica de OCR. Esse tipo de ferramenta é sujeita a erros de conversão e, portanto, é

necessária uma verificação manual para a correção de erros. Um exemplo de reconhecedor óptico é o *Abby*.

- Buscadores *Web*: permitem a coleta de textos disponíveis via *Web*. Os textos são recuperados por palavras chaves. Exemplos de buscadores são o *Google* e o *Yahoo*. Adicionalmente, existem ferramentas que pré-processam o resultado da busca como o *WebCorp* e o *KWiCFinder*.
- Navegadores *offline*: armazenam sítios *Web* inteiramente no computador do usuário. Como todos os textos são coletados, o usuário pode analisá-los posteriormente para a escolha dos textos que farão parte do córpus. Um exemplo de ferramenta desse tipo é o *HTTrack*.
- Mineradores *Web* (*Web crawlers*): são ferramentas que varrem a *Web*, acessando diversos sítios. Essas ferramentas podem armazenar textos lidos na *Web* a partir critérios de armazenamento definidos pelo usuário. Um exemplo de minerador é o *HTDig*.
- Conversores de formato: são úteis para converter textos em diversos formatos coletados via *Web* (por exemplo, HTML, DOC, PDF, PS) para texto sem formatação (TXT). Um exemplo de conversor é o XPDF, capaz de converter o formato PDF para texto.

As ferramentas de anotação podem ser subdividias em duas categorias: de anotação manual e de anotação automática. As ferramentas de **anotação manual** de textos são:

- Editores de texto comuns: permitem a edição dos textos e a inserção de etiquetas. Entretanto, oferecem poucos recursos para o tratamento de etiquetas. Um exemplo é o editor *Emacs*, que diferencia as etiquetas do restante do texto através de mudanças nas cores dos elementos (destaque de sintaxe ou *syntax highlighting*).
- Editores para anotação XML/SGML: possuem tratamento próprio para etiquetas XML. O editor *Xanthipe* (RAYSON, 2002) é um exemplo de editor avançado e chegou a ser usado para etiquetação sintática no BNC. Esse editor é capaz de realizar alterações em série nos textos. O editor *Open XML Editor* é capaz de checar erros de sintaxe em documentos XML e analisar a integridade do documento a partir de DTDs.

Ferramentas para **anotação automática** de textos:

- Segmentadores: dividem o texto em sentenças e parágrafos de forma automatizada, inserindo as etiquetas adequadas. A princípio, o processo de segmentação pode parecer uma simples busca pelos sinais de pontuação. Entretanto, a presença de abreviaturas no texto com o caractere ponto é um fator que deve ser considerado para evitar erros de etiquetação durante o processo. Para resolver esse problema, é necessário que a ferramenta conheça *a priori* as abreviaturas do texto. Um exemplo desta ferramenta é o *Senter*, desenvolvido no NILC.
- Etiquetadores morfossintáticos e sintáticos: utilizados durante a etapa de compilação do córpus, os etiquetadores inserem automaticamente informações morfossintáticas e sintáticas com um alto grau de precisão. A precisão dos etiquetadores morfossintáticos chega a 98%, enquanto que a precisão de etiquetadores sintáticos chega a 91% (para o idioma inglês). Um exemplo de software pertencente a essa categoria é o etiquetador Palavras (BICK, 2000), baseado em regras construídas manualmente. O Palavras etiqueta textos em Português e sua precisão é superior a 97% (no nível morfossintático). Há três etiquetadores morfossintáticos treinados por regras automáticas disponibilizados no NILC: *Tree Tagger*, MXPOST e *TBL Tagger*. Os etiquetadores foram treinados a partir do córpus MacMorpho e possuem uma precisão superior a 96%. O resultado do treinamento está disponibilizado publicamente[12].
- Lematizadores: encontram e anotam o lema (ou forma canônica) das palavras presentes no córpus. A lematização é um processo útil para estudo de palavras com muitas flexões, como os verbos. Ferramentas dessa categoria também devem tratar ambigüidade. Por exemplo, a forma flexionada "foi" pode ter como lema o verbo "ser"' ou o verbo "ir". Essas ferramentas são particularmente úteis para tarefas lexicográficas. Um exemplo é o lematizador de verbos do Português *LX Lemmatizer*.
- Anotadores de co-referência: são ferramentas capazes de identificar expressões que se referem a um mesmo elemento dentro de um texto. Como exemplo, é possível definir o elemento a que um pronome se refere. Anotadores de co-referência são úteis para

12 http://www.nilc.icmc.usp.br/nilc/tools/nilctaggers.html

análise do discurso e sumarização. O sumarizador automático *RHeSumaRST* (SENO; RINO, 2005) identifica co-referências automaticamente a partir de heurísticas.

- Anotadores diversos: aplicados a outros níveis lingüísticos não descritos acima. Um exemplo é o *RSTTool* que atua no nível retórico e o segmentador de estruturas esquemáticas de resumos em português, também desenvolvido no NILC e disponível no ambiente *SciPo* (*Scientific Portuguese*).

Ferramentas para **acesso** a córpus:

- Concordanceadores: são os principais tipos de ferramentas, pois permitem o estudo do conteúdo do córpus através de buscas sofisticadas. Além disso, essas ferramentas são capazes de alinhar diversos resultados e exibi-los simultaneamente. Os concordanceadores podem ser agrupados em três tipos: (a) KWIC (*KeyWord In Context*) no qual as ocorrências buscadas são exibidas juntamente com seu contexto (blocos de texto a direita e a esquerda de cada ocorrência buscada), (b) KWAC (*KeyWord And Context*) no qual as ocorrências e seus contextos são exibidos separadamente (de forma a destacar as ocorrências) e (c) KWOC (*KeyWord Out of Context*) no qual as ocorrências são exibidas separadamente de seus contextos e exibidas novamente dentro dos contextos. A maior parte dos processadores de córpus possui concordanceadores, entre eles, o *Unitex*. Um concordanceador via *Web* de resultados de textos *Web* é disponibilizado pelo processador *WebCorp*.
- Buscadores textuais: realizam buscas no córpus de forma semelhante ao concordanceador, mas exibem apenas um resultado por vez. Alguns processadores de córpus não possuem buscadores textuais, pois estes são substituídos pelos concordanceadores. Exemplos de buscadores textuais podem ser encontrados em editores de texto simples como o *Notepad*, distribuído juntamente com *MS-Windows*.
- Buscadores de metadados bibliográficos: realizam buscas nos metadados que descrevem os textos. São úteis tanto para a obtenção de informações de bibliografia utilizadas para análise de balanceamento e representatividade do córpus quanto para a formação de subcórpus para pesquisas específicas. Por exemplo, é possível formar um subcórpus com todas as obras de um autor ou todas as obras de um período. O processador de córpus *Philologic* possui um buscador de dados de cabeçalho. O

projeto LW também possui e permite 3 tipos de buscas desde simples a sofisticadas.

- Contadores de freqüência: contam o número de palavras total em um córpus ou subcórpus, além de mostrar as palavras presentes no córpus e suas freqüências. As ferramentas devem ignorar pontuação, números e outros tipos de símbolos, pois não representam palavras. Da mesma forma, as etiquetas, caso presentes, devem ser ignoradas. Os contadores de palavras mais utilizadas contam formas (*tokens*) como é o caso do contador do *MS-Word* que conta palavras ortográficas (separadas por brancos). O contador do projeto LW, em especial, conta alguns padrões de unidades lexicais complexas.
- Geradores de *n*-gramas: são ferramentas capazes de reconhecer e levantar *n*-gramas em um texto. *N*-gramas consistem em seqüências de *n* palavras e são usados em PLN para extração de estatísticas sobre o texto e em Recuperação de Informações (RE) para a execução de buscas. Um exemplo é a ferramenta NSP (*Ngram Statistics Package*).
- Buscadores de colocações: colocações de uma dada palavra são sintagmas de uso comum na língua nos quais a palavra é empregada (FIRTH, 1935 apud SILVA, 2003, p. 16) . Um exemplo para a palavra "arma" é o enunciado "armas de destruição em massa". Buscadores de colocações são capazes de encontrar e exibir as colocações através de técnicas de análise estatística. O *Philologic* é capaz de gerar colocações.
- Buscadores orientados por glossários: permitem a pesquisa das palavras de um ou mais glossários em um córpus. O uso de glossários simplifica expressões de busca, por exemplo, através de um glossários de verbos é possível buscar por todos os verbos de um texto. O concordanceador do *Unitex* é capaz de realizar buscas orientadas por glossários.

Ferramentas de **extração de conhecimento**:

- Sumarizadores: realizam resumo automático de um texto. O uso do córpus é útil tanto para a construção do sumarizador, como para sua avaliação. Exemplos de sumarizadores incluem as ferramentas EXPLORA (Exploração de Métodos Diversos para a Sumarização Automática) e *GistSumm* (*Gist Summarizer*), desenvolvidos no NILC.

- Tradutores automáticos: esse tipo de ferramenta pode ser usado em córpus paralelos para aprendizado de regras de tradução automática e para avaliação automática de qualidade da tradução. A avaliação também pode ser feita (manualmente) com base em córpus monolíngües. No NILC é desenvolvido um tradutor automático entre Português e UNL (*Universal Networking Language*).
- Ferramentas gerais de recuperação de informação: são ferramentas para recuperação de documentos com base em buscas por palavras chaves. É possível avaliar a qualidade de uma busca a partir das medidas *precisão* (a proporção de entre documentos relevantes recuperados na busca e o total de documentos recuperados) e *cobertura* (a proporção entre documentos relevantes recuperados na busca e o total de documentos relevantes). Por exemplo, Aires (2005) apresenta uma forma de classificar os resultados de busca em textos do português segundo sete necessidades específicas dos usuários. A busca também inclui gênero e tipo textual.

O domínio de linguagens de programação é um apoio importante para o processamento de córpus, pois permite a construção de ferramentas de forma eficiente e rápida quando há ausência de ferramentas para suprir as necessidades do projeto de córpus. Por exemplo, a linguagem *Perl* permite a construção de contadores de freqüência a partir de estruturas de programação relativamente simples.

Outros autores agrupam as ferramentas de modos diferentes. Por exemplo, para Atkins, Clear e Ostler (1992) as ferramentas de processamento de córpus pertencem a dois grupos: ferramentas básicas e ferramentas avançadas. As ferramentas básicas incluem os tipos de ferramentas mais gerais e indispensáveis para a pesquisa de processamento de córpus como contadores de freqüência e concordanceadores. Ferramentas avançadas permitem a realização de pesquisas mais sofisticadas, além de tornarem as pesquisas usais mais eficientes. Exemplos desse tipo de ferramenta são lematizadores e etiquetadores morfossintáticos.

Rayson (2002) apresenta um estudo sobre diversas ferramentas de processamento de córpus e as agrupa em três conjuntos de acordo com suas funções: desenvolvimento de córpus, edição de córpus e extração de informação.

Ferramentas para **desenvolvimento de córpus** estão relacionadas à etapa de coleta e anotação de textos. Segundo o autor, geralmente são softwares de prateleira (*off-the-shelf software*), distribuídos comercialmente. Esses softwares incluem digitalizadores (*scanners*) e

reconhecedores ópticos para textos impressos e mineradores *Web* para textos digitais. As ferramentas desse grupo podem atender a três propósitos diferentes: codificação do texto, anotação e codificação da anotação. A codificação está basicamente relacionada à definição dos caracteres utilizados no córpus e ao modo como o texto será representado em formato digital. A anotação envolve o uso de padrões discutidos na Seção 2.6. A codificação de anotação define a representação computacional das etiquetas.

Ferramentas de **edição de córpus** são aplicadas após a coleta dos textos e têm como objetivos a correção de texto, a desambigüização (ou remoção de ambigüidade) e a conversão de formato. A correção é aplicada em erros gerados durante a coleta do texto como, por exemplo, erros de digitação, falhas de reconhecimento óptico, erros do etiquetador automático, entre outros. A desambigüização é necessária quando etiquetadores automáticos são utilizados, pois estes inserem múltiplas etiquetas quando encontram ambigüidades na análise do texto (no nível lingüístico em que se aplicam). A conversão de formato pode ser utilizada em textos que já possuam algum tipo de anotação como, por exemplo, textos em HTML obtidos da *Web*. Nesse caso, as etiquetas podem ser convertidas para o padrão de etiquetação utilizado no córpus. As ferramentas de edição variam de editores de textos convencionais como *Notepad* ou *Emacs* a editores mais sofisticados como o *Xanthipe*.

Ferramentas de **extração de informação** são utilizadas na etapa de uso do córpus e incluem contadores de freqüência, concordanceadores, programas de recuperação de informação, etc. Os programas de recuperação de informação são um subtipo das ferramentas de extração de informação.

A Seção 3.2.1 detalha os glossários computacionais, um recurso fornecido por algumas ferramentas, útil para pesquisas lexicográficas.

3.2.1 Glossários computacionais

Dicionários computacionais (também chamados de glossários ou de léxicos computacionais) consistem de uma lista de palavras e informações opcionais de diversos tipos (fonológicas, morfológicas, sintáticas, semânticas, etc). Um glossário pode ser utilizado em um córpus para diversas finalidades como, por exemplo, disponibilizar buscas por flexões de uma lexia a partir de sua forma canônica, identificar neologismos e identificar palavras que caíram em desuso. Em Muniz (2004), é apresentada uma classificação dos glossários, segundo seu grau de complexidade, em quatro tipos: dicionários legíveis por máquina (*Machine*

Readable Dictionary - MRD), dicionários tratáveis por máquina (*Machine Tractable Dictionary* - MTD), base de dados lexicais e bases de conhecimento lexicais.

MRDs consistem em glossários legíveis por humanos e por máquinas, enquanto que MTDs não podem ser lidos por humanos, mas sua estrutura interna possibilita o processamento por máquina de forma rápida e eficiente (WILKS et al., 1988). Já em uma base de dados lexicais, o glossário é tratado como uma complexa rede de relações (sintagmáticas, semânticas, paradigmáticas, etc). Por fim, uma base de conhecimento lexical se diferencia das bases de dados lexicais, pois a primeira é dinâmica, sendo continuamente alterada, ao passo que a segunda é estática. Além disso, as bases de conhecimento lexical possuem um padrão de representação próprio conhecido como LRL (*Lexical Representation Language*). Dois exemplos de glossários são os dicionários de francês do laboratório LADL (*Laboratoire d'Automatique Documentaire et Linguistique*) e o glossários multilingüe do projeto ISLE (*International Standards for Language Engineering*) (ATKINS et al., 2002).

Os glossários do laboratório LADL se baseiam no formalismo DELA (*Dictionnarie Electronique du LADL*). O formalismo define dois tipos principais de MTD: glossários de forma canônica (DELAS) e os glossários de formas flexionadas (DELAF). Além disso, existem duas variantes para dicionários de palavras compostas: DELAC para formas canônicas e DELACF para formas flexionadas. Muniz (2004) construiu um glossário DELA para o Português do Brasil contendo lexias simples e compostas a partir de uma versão do glossário do ReGra. A Tabela 3.1 compara o dicionário DELAF e DELACF criado:

Tabela 3.1: Exemplos de entradas no formato DELA

DELAF	**DELACF**
abacado,.N:ms	ab-rogamento,.N+XN:ms
abacados,abacado.N:mp	ab-rogamentos,ab-rogamento.N+XN:mp
abacalhoa,abacalhoar.V:P3s:Y2s	ab-rogação,.N+XN:fs
abacalhoada,abacalhoar.V:K	ab-rogações,ab-rogação.N+XN:fp
abacalhoadas,abacalhoar.V:K	ab-rogáveis,ab-rogável.A+XA:mp:fp
abacalhoado,abacalhoar.V:K	ab-rogável,.A+XA:ms:fs
abacalhoados,abacalhoar.V:K	abaixa-luz,.N+VN:ms
abacalhoai,abacalhoar.V:Y2p	abaixa-luzes,abaixa-luz.N+VN:mp

A entrada “abacados,abacado.N:mp” define a forma canônica “abacado” para a flexão “abacados”, sua categoria morfológica (classe substantivo denotada por “N”) e sua flexão (masculino plural, denotada por “mp”). Símbolos reservados podem ser representados como

parte de uma entrada se forem antecedidos pelo símbolo "\", como para a entrada "E\=mc2,.FORMULA". Mais informação sobre o formato DELA pode ser encontrada em (PAUMIER, 2006). Informações sobre os códigos gramaticais e flexionais utilizados no glossário do Português do Brasil podem ser encontradas em (MUNIZ, 2004).

3.3 Processadores de córpus analisados

Esta seção apresenta cinco ferramentas livres para processamento de córpus e uma breve introdução sobre algumas ferramentas proprietárias (gratuitas ou comerciais). Existem outros comparativos entre ferramentas para processamento de córpus, por exemplo, o estudo de Rayson (2002). Rayson compara 9 ferramentas de processamento de córpus segundo 8 critérios de avaliação: (1) licença de uso (comercial, gratuita ou aberta), (2) sistema operacional, (3) reconhecimento de anotação, (4) indexação de texto, (5) níveis de anotação, (6) contadores de freqüência, (7) comparação de listas de freqüência e (8) presença de concordanceadores. Entretanto, 6 ferramentas mostradas são de código proprietário e uma parte das ferramentas está desatualizada, pois são para o antigo sistema operacional MS-DOS ou para antigas variações do *Unix*. Santos e Ranchhod (1999) fazem uma comparação entre o *Intex* e o *IMS Workbench*.

3.3.1 GATE

GATE (*General Architecture for Text Engineering*) (Cunningham *et al.* 2007) é uma ferramenta para engenharia da linguagem desenvolvida e mantida pela Universidade de *Sheffield*, capaz de fornecer uma infra-estrutura robusta para desenvolvimento e distribuição de softwares para PLN. A ferramenta, desenvolvida em *Java*, teve sua primeira versão lançada em 1995 e tem sido usada em grandes organizações para pesquisas variadas. O GATE é composto por uma arquitetura, uma biblioteca e um ambiente para desenvolvimento e teste. A instalação padrão fornece um conjunto de recursos chamado de CREOLE (*Collection of REusable Objects for Language Engineering*). Existem diferentes três tipos de componentes: recursos de linguagem (glossários, documentos, córpus, entre outros), recursos de processamento (analisadores sintáticos, etiquetadores, entre outros) e recursos visuais (visualização gráfica e edição de componentes). Uma vantagem do GATE é o número de recursos disponíveis, principalmente para a Língua Inglesa. Uma desvantagem é curva de aprendizado da ferramenta é relativamente alta, em partes, devido ao grande número de

recursos. A Figura 3.1 mostra a interface da ferramenta. A esquerda é possível observar os recursos fornecidos organizados por categoria.

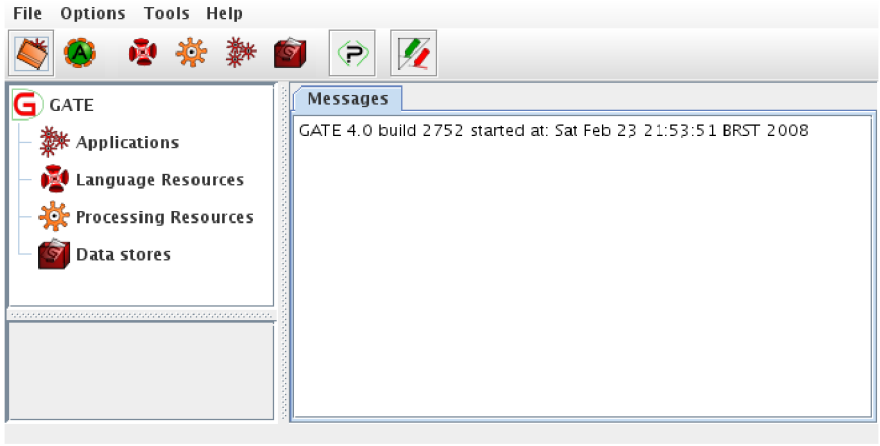

Figura 3.1: Tela inicial do GATE

3.3.2 Philologic

Philologic (UNIVERSITY OF CHICAGO, 2006) é uma ferramenta *Web* para buscas, recuperação e análise de córpus desenvolvida por Leonid Andreev e pesquisadores da universidade de *Chicago* como uma das metas do projeto ARTFL (*American and French Research on the Treasury of the French Language*) (WOLFF; ANDREEV; OLSEN, 1999). A ferramenta foi originalmente desenvolvida para gerenciar textos em francês, mas devido à codificação *Unicode*, diversos idiomas são permitidos. Além disso, graças ao mecanismo de indexação e o uso do padrão TEI, os usos da ferramenta não se limitam apenas ao processamento de córpus e incluem o gerenciamento de enciclopédias, de dicionários e até mesmo de sistemas multimídias (contendo sons e vídeos). Além de uso do padrão TEI, a ferramenta pode ser adaptada para processamento de outros padrões de anotação (como o XCES), através de seus arquivos de configuração.

Os recursos oferecidos pelo *Philologic* incluem: um concordanceador, um gerador de colocações, um contador de freqüências e um buscador de dados de cabeçalho. O buscador de dados de cabeçalho é capaz de listar os documentos do córpus e formar subcórpus. Para processamento de córpus históricos, a ferramenta possui um sistema para busca de variantes de grafia baseado no utilitário *Agrep*. As ferramentas do *Philologic* são acessíveis via ambiente *Web*.

A busca é dividida em 5 estágios: a) Definição de um subcórpus; b) Expansão de

palavras (tratamento de expressões regulares); c) Busca por palavras indexadas; d) Extração de texto; e) Resolução de *hyperlinks* (hiper-ligações) e formatação (em HTML). O item (a) é opcional. As informações recuperadas podem apontar para outros textos dentro do córpus atual, outros córpus ou mesmo para imagens e sons. Estruturas em textos tais como palavras, sentenças, parágrafos e capítulos podem ser localizadas e identificadas durante as buscas. Com esse recurso é possível, por exemplo, pesquisar duas ou mais palavras e recuperar apenas os resultados em que as palavras ocorrem na mesma sentença. O resultado de uma busca no concordanceador é exibido na Figura 3.2, na qual a palavra de busca é "*of*".

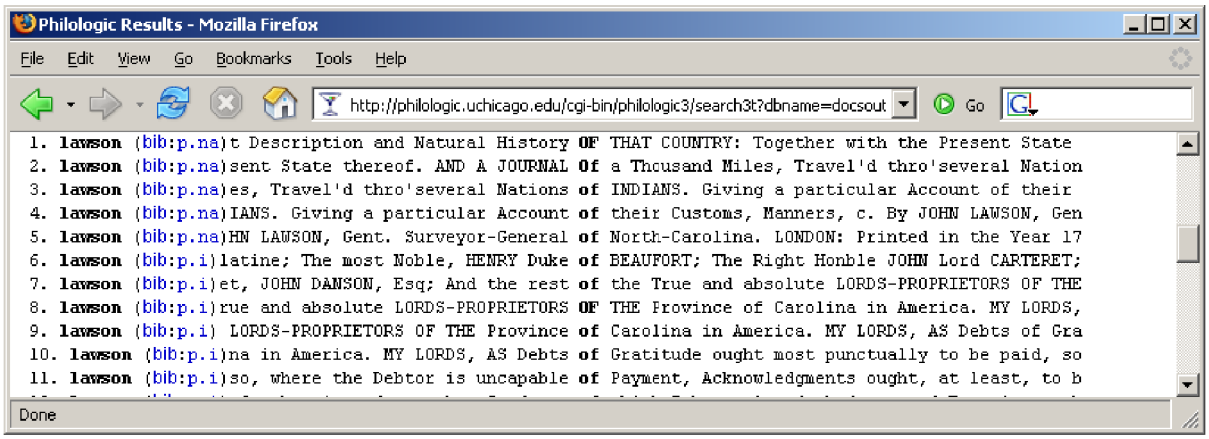

Figura 3.2: Concordanceador *Philologic*

A definição de subcórpus é feita, geralmente, a partir dos metadados autor, título da obra ou data de publicação. Entretanto, é possível especificar outros metadados para a busca ajustando os parâmetros de configuração da ferramenta. A criação de subcórpus pode ser feita de forma recursiva, o que permite ao usuário refinar gradualmente os resultados da busca.

O processador de córpus permite o uso de caracteres coringas (*wildcards*) para tratamento de caracteres acentuados. Como exemplo, uma busca por "*calderOn*" pode retornar as palavras "*calderon*" e "*calderón*". Letras em maiúsculas correspondem a qualquer variação acentuada do caractere em questão. Símbolos com diacríticos podem ser pesquisados por meio de expressões como "c,", "e^" e "e\" correspondentes respectivamente a "ç", "ê" e "è". O uso de tais expressões facilita a busca em sistemas sem auxílio à digitação de acentos.

Outras buscas avançadas podem ser efetuadas. Alguns exemplos de buscas são: (a) localizar todas as palavras iniciadas por "faz" ("fazendo", "fazem", "fazerem", entre outras); (b) determinar quão freqüente é uso de uma palavra na obra de um dado autor; (c) obter uma lista das palavras usadas com freqüência juntamente com a palavra "tradicional" (colocações

da palavra, como em “família tradicional”) e (d) ordenar os resultados de uma busca de acordo com a posição da palavra buscada na sentença.

O Philologic tem como ponto forte o ambiente Web e a montagem de subcórpus oferecendo ao utilizador a possibilidade de efetuar buscas em apenas um subconjunto do córpus. A desvantagem da ferramenta consiste no uso parcial ao padrão XML TEI, pois apenas a versão *TEI-Lite* é processada (apesar de poder ser personalizado para processar outros tipos de anotação). Além disso, a inserção de novos textos precisa ser feita manualmente pelo administrador do servidor de córpus, pois a interface *Web* não oferece esse recurso.

3.3.3 Tenka Text

Tenka Text, também conhecido como *Corsis* é uma ferramenta para análise de córpus escrita em linguagem C#. O projeto é uma alternativa livre ao processador de córpus *WordSmith Tools*, e é disponibilizado para 5 sistemas operacionais. A ferramenta oferece uma biblioteca e uma interface com duas ferramentas: lista de palavras e um concordanceador. O concordanceador permite expressões regulares e caracteres coringa de forma semelhante ao *WordSmith*. Uma vantagem dessa ferramenta é o fácil aprendizado por usuários que já estão acostumados ao *WordSmith*, devido a similaridade entre as duas. Uma desvantagem é o fato de o ciclo de desenvolvimento ainda estar em seu início, e parte dos recursos da ferramenta ainda estarem inacessíveis. Entretanto, esse problema pode ser resolvido em versões futuras. A Figura 3.3 mostra o concordanceador da ferramenta.

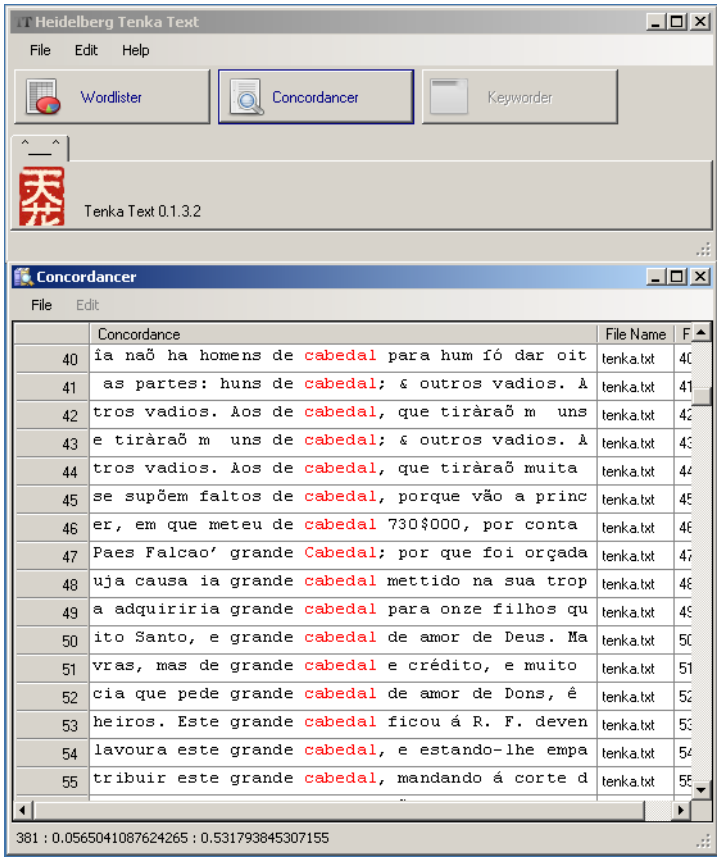

Figura 3.3: Concordanceador da ferramenta Tenka

3.3.4 Unitex

Unitex (PAUMIER, 2006) é um sistema de processamento de córpus baseado na teoria dos autômatos e consiste em um conjunto de ferramentas desenvolvidas em linguagem C e uma interface gráfica desenvolvida em linguagem Java. O sistema foi criado por Sébastien Paumier no Instituto *Gaspard-Monge* da Universidade de *Marne-la-Vallée* na França como uma implementação livre do software *Intex*. A versão 1.0 da ferramenta foi disponibilizada publicamente em 2002. Uma interface *Web* foi desenvolvida pelo grupo desenvolvedor da *GlossaNet* (FAIRON, 1999).

Entre os recursos oferecidos pelo *Unitex* encontram-se: (a) um concordanceador orientado por glossários, (b) um gerenciador de glossários DELA, (c) um utilitário para contrastar córpus com glossários, (d) um contador de freqüências, (e) um gerenciador de gramáticas e (f) um gerenciador de tabelas de léxico-gramática. Um córpus *Unitex* é codificado em UTF-16, geralmente sem nenhum tipo de anotação. Entretanto, a partir da versão 1.2, é permitido o uso de etiquetas simples (delimitadas por chaves) para classificação

gramatical. Os símbolos permitidos nas lexias são definidos pelo usuário, o que permite o processamento de formas (*tokens*) em diferentes idiomas. Por exemplo, no Português, é necessário incluir os símbolos "Ç" e "Á", entre outros.

A versão 2.0 (beta) da ferramenta provê dicionários para mais de 17 idiomas (incluindo o Português do Brasil). Além disso, novos idiomas podem ser adicionados facilmente pelo usuário. O *Unitex* permite o uso de vários dicionários simultaneamente. É possível classificar os dicionários DELA em mais prioritários e menos prioritários, definindo-se a ordem de pesquisa das palavras nos dicionários. A ferramenta também fornece recursos para a criação de dicionários DELA. É possível, por exemplo, comprimir um dicionário, verificar se contém erros de formatação e ordenar suas entradas. Além disso, um novo dicionário de formas flexionadas pode ser gerado a partir de um dicionário DELAS (ou DELACS).

As gramáticas são representadas por meio de autômatos de texto, um formalismo baseado em autômatos finitos. Os autômatos possuem diversos usos, incluindo flexão de palavras, segmentação, normalização do texto, remoção de formas ambíguas e construção de expressões de busca no concordanceador. Como exemplo de normalização (expansão de contrações), a palavra "daí" é convertida em "de aí". É importante notar que a normalização não pode ocorrer para palavras ambíguas tal como a palavra "desse" que pode significar tanto "de esse" quanto uma conjugação do verbo "dar". A remoção de ambigüidade fornece meios para o usuário identificar a classe gramatical das palavras de uma sentença. Um exemplo para a sentença "Nós chegamos a São Paulo" é mostrado na Figura 3.4. Diversos trabalhos aplicam autômatos finitos e suas variações para reconhecimento de gramáticas (JESUS; NUNES, 2000) e remoção de ambigüidade de texto (ROCHE, 1992). Além disso, os autômatos de textos podem ser gerados automaticamente a partir de tabelas de léxico-gramática. Essas tabelas contêm palavras de uma língua e uma representação compacta de suas propriedades.

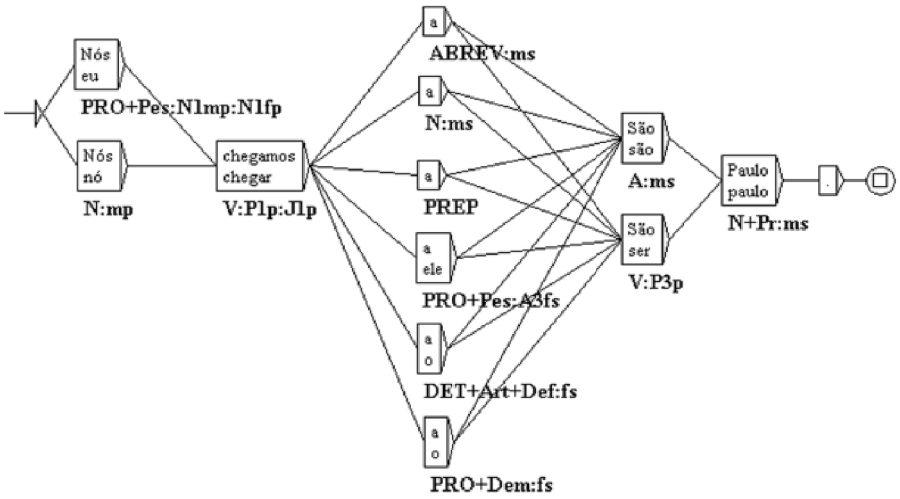

Figura 3.4: Autômato de texto para resolução de ambigüidade (MUNIZ, 2004)

Em (MUNIZ, 2004) foram construídas regras de resolução de ambigüidades, bibliotecas para acesso a glossários compactados e ferramentas para validar esses recursos, além do glossário discutido na Seção 3.2.1. As contribuições foram incorporadas na ferramenta *Unitex-PB* (MUNIZ; NUNES; LAPORTE, 2005) e no *Unitex* original.

A Figura 3.5 mostra o resultado de uma busca no concordanceador (à esquerda) e a aplicação de um glossário de Português do Brasil no córpus (à direita). As palavras foram agrupadas em três categorias: simples, compostas e desconhecidas. As classes gramaticais de cada palavra são exibidas. Para casos de homografias, são listadas todas as classes gramaticais possíveis.

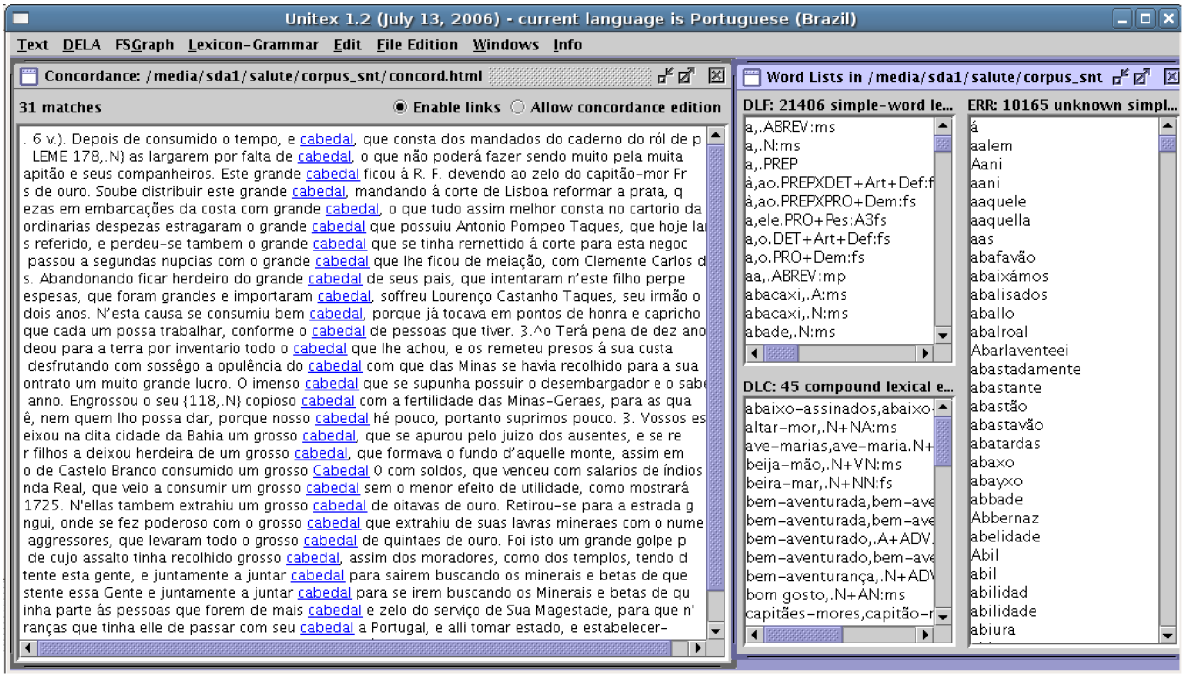

Figura 3.5: Interface *Unitex* com o concordanceador e a lista de palavras

A Tabela 3.2 contém alguns exemplos de expressões de busca no concordanceador *Unitex*.

Tabela 3.2: Expressões de busca *Unitex*

Expressão	**Busca**
book	A palavra “book”
red book	A palavra “red” sucedida por “book”
Mr+Dr	As abreviaturas “Mr” ou “Dr”
<MAJ>	Palavras inteiramente em maiúsculas
MR. <PRE>	Abreviatura “Mr.” sucedida por uma palavra com a primeira letra maiúscula
<NB>	Um conjunto de dígitos
<!DIC>	Palavras que não fazem parte do dicionário
<DET>.<N>	Artigo sucedido por nome
<be>	Entradas que contém “be” como forma canônica
<be.V>	Entradas que contém “be” como forma canônica e são verbos

A principal vantagem desse processador de córpus é o apoio a buscas no córpus a partir de informações definidas nos glossários. A desvantagem é o uso limitado de anotação. Outra limitação é a abertura de apenas um arquivo por vez, o que dificulta o gerenciamento do córpus.

3.3.5 Xaira

Xaira (*XML Aware Indexing and Retrieval Architecture*) (XIAO, 2005) é um mecanismo de busca XML utilizado para gerenciamento de córpus anotados, desenvolvido por Lou Burnard e Tony Dodd na Universidade de *Oxford* com financiamento da fundação *Andrew W Mellon* e do consórcio BNC. A ferramenta foi criada em 2004 com o objetivo de tratar XML e oferecer os recursos semelhantes aos do software SARA (*SGML Aware Retrieval Application*) (ASTON; BURNARD, 2001).

Entre os principais recursos oferecidos por essa ferramenta destacam-se: o concordanceador, o contador de freqüências e o indexador para acelerar as buscas. O concordanceador pode realizar buscas complexas, incluindo buscas em elementos XML.

É possível processar diversos padrões tais como TEI e XCES. A arquitetura *Xaira* é cliente-servidor, o que permite que um córpus seja disponibilizado remotamente. Além disso, é possível que o acesso seja efetuado por outras aplicações através de *Web Services* devido ao uso do protocolo SOAP (*Simple Object Access Protocol*). A ferramenta pode adicionar aos arquivos uma marcação XML básica caso não possuam nenhuma. As buscadas no servidor são otimizadas com o uso de indexação. O cliente fornece aos usuários diversas opções de busca. Uma terceira ferramenta (*Xaira Index Toolkit*) é utilizada para criar o arquivo de índice. A arquitetura *Xaira* é ilustrada na Figura 3.6.

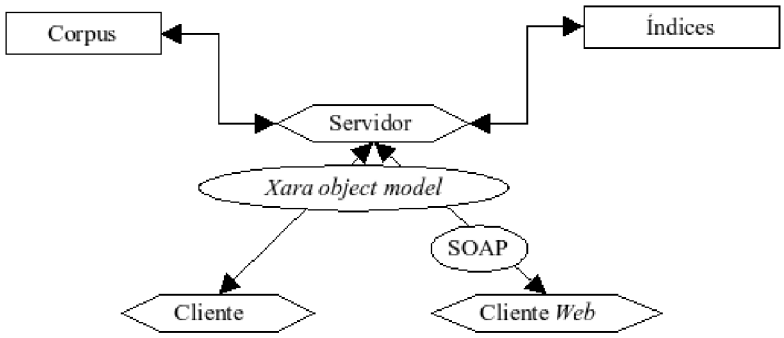

Figura 3.6: Arquitetura *Xaira*

O mecanismo de indexação utilizado permite que o tempo das buscas em bases de córpus grandes sofra uma redução significativa. O processo é dividido em três etapas. Na primeira etapa, é definido um arquivo de parâmetros e um arquivo de cabeçalho é criado. Na segunda etapa, são definidos alguns parâmetros sobre a indexação. A terceira etapa consiste na indexação propriamente dita. Além disso, o indexador fornece recursos extras como utilitários

para checar a formatação dos arquivos XML, adição ou remoção de diferentes idiomas e classificação do texto por diferentes taxonomias.

O cliente é composto por um concordanceador e por diálogos de busca. O usuário pode efetuar diferentes tipos de buscas, entre elas: busca por palavras simples, sentenças, expressões regulares e buscas em elementos XML. As buscas também podem ser feitas através de *XML-Based CQL* (*Corpus Query Language*), um mecanismo poderoso para buscas em córpus anotados em XML. Devido à proximidade entre a anotação do córpus BNC e o TEI, o uso de córpus em TEI dentro da ferramenta é facilitado. O Xaira também fornece um assistente que simplifica a tarefa de criação de pesquisas complexas e permite que as buscas sejam salvas em disco para serem reutilizadas. Duas opções de visualização dos resultados são disponibilizadas. Na primeira, os resultados são exibidos em XML enquanto que, na segunda, são exibidos em texto puro. Algumas possíveis pesquisas que podem ser efetuadas no processador de córpus *Xaira* são: (a) Descobrir a palavra mais freqüente em um córpus; (b) Procurar pelas ocorrências da palavra "ele" e "ela" em uma mesma frase e exibir uma amostra de 20 ocorrências; (c) Procurar sentenças iniciando com uma conjunção; (d) Mostrar todas as formas flexionadas da forma canônica "contar"; (e) Procurar sentenças começando com "quando" e terminando com um sinal de interrogação; (f) Verificar a freqüência da palavra "natureza" em diferentes tipos de córpus. A Figura 3.7 ilustra o concordanceador da ferramenta.

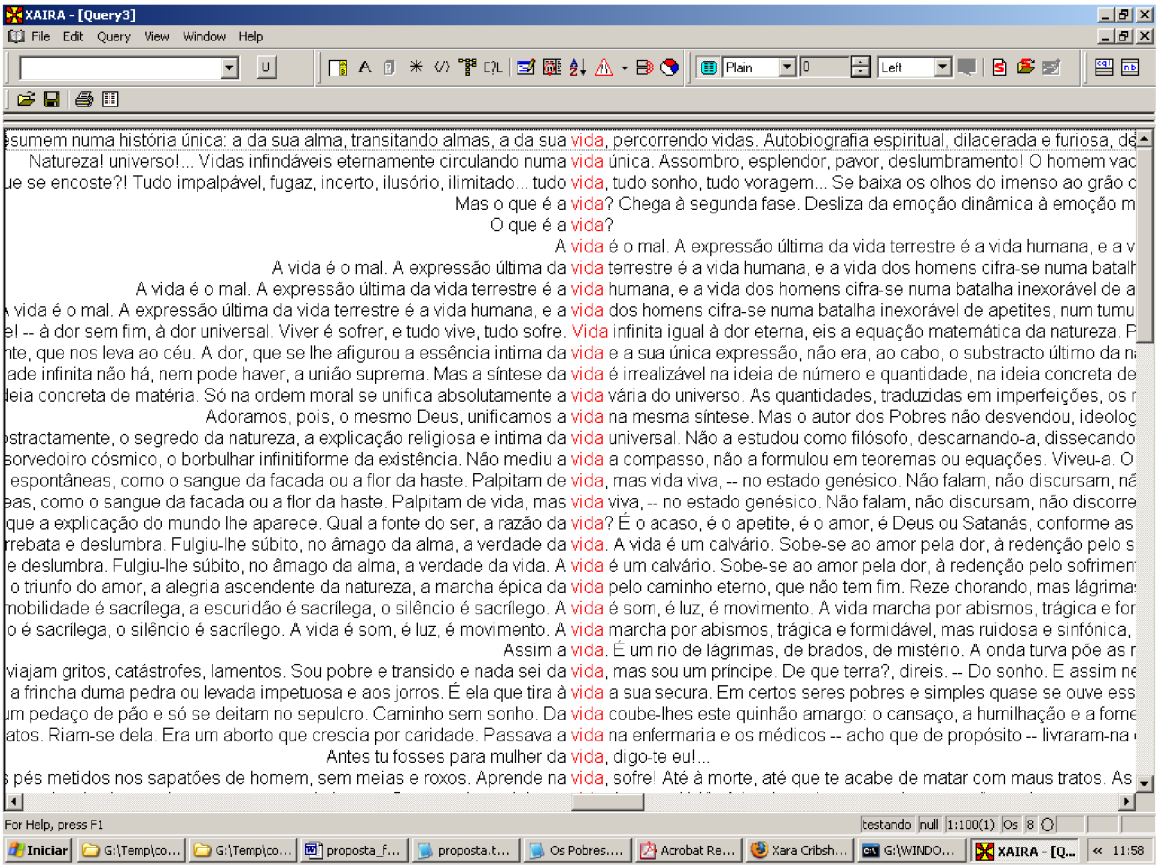

Figura 3.7: Concordanceador *Xaira*

É possível aplicar filtragens em buscas com um grande volume de resultados para criação de subcórpus. As técnicas para filtragem incluem: amostragem, ordenação e particionamento. Na amostragem, o subcórpus é gerado automaticamente. No caso de ordenação e particionamento, o usuário pode ordenar e selecionar os elementos que deseja filtrar.

A principal vantagem desse processador de córpus é o poderoso mecanismo de busca, capaz de processar elementos XML de forma simples e prática. A desvantagem se encontra na sua interface de difícil uso para iniciantes.

3.3.6 Outros processadores de córpus

A seguir, são referidos outros processadores de córpus.

- *Emdros*: mecanismo de busca em textos anotados.
- *IMS Corpus Workbench*: um conjunto de ferramentas para recuperação de textos em grandes córpus (CHRIST, 1994). Essa ferramenta é usada em projetos como o AC/DC

e o CNC.

- *Intex* (SILBERZTEIN, 1994): foi criado no laboratório LADL e tem como implementação livre o processador de córpus *Unitex*, de forma que os recursos oferecidos por ambos são bem semelhantes. O *Intex* e o *IMS Workbench* são ambos baseados na tecnologia de autômatos finitos. Santos e Ranchhod (1999) levantam as similaridades das ferramentas, O *Intex* se mostrou ideal para a criação de recursos lingüísticos, enquanto que o *IMS Workbench* forneceu boas ferramentas para realizar pesquisas no córpus.
- VIEW (*Variation in English Words and Phrases*): permite acesso público ao córpus do projeto BNC, com concordanceador e buscas sofisticadas. Também é utilizado no Córpus do Português e no Córpus do Espanhol (com 100 milhões de palavras).
- *Wordsmith Tools*: software para análise léxica desenvolvido pela Universidade de *Oxford*. Tem sido amplamente utilizado por lingüistas e difundido por sua facilidade de uso.

3.4 Comparativo entre os processadores de córpus

Os processadores de córpus foram avaliados pelas métricas definidas na ISO 9126 (UNIVERSITÉ DE GENÈVE, 2006). Seis métricas são definidas para a avaliação de qualidade de software, além de diversas outras derivadas a partir das métricas principais:

- Funcionalidade: consiste na análise das funções desempenhadas pelo o software em questão. Em muitas avaliações, essa é a métrica mais importante.
- Confiabilidade: analisa a capacidade do software de funcionar como esperado em todas as tarefas solicitadas pelo usuário. Também está relacionada à tolerância a falhas e recuperação de erros. Falhas de projeto e implementação podem causar erros na execução do software, comprometendo sua confiabilidade.
- Usabilidade: analisa a facilidade de usar e de aprender o software. Um software com uma boa usabilidade deve fornecer uma interface simples, amigável e intuitiva ao usuário.
- Eficiência: analisa o desempenho do software e o uso de recursos do sistema. Um programa é dito escalável quando apresenta boa eficiência mesmo para volumes grandes de dados.

- Manutenibilidade: avalia a facilidade em modificar e expandir as funcionalidades do software. Também análise a facilidade em entender, alterar e testar o software. Um bom projeto de software possibilita a criação de ferramentas com alta manutenibilidade. Outro fator importante é documentação do código.
- Portabilidade: avalia o número de plataformas de software e *hardware* sobre as quais o software pode operar em condições normais de funcionamento. Um software portável pode ser executado em diversas plataformas, de forma independente das configurações do usuário.

A funcionalidade foi avaliada a partir de 8 critérios (mostrados na Tabela 3.3). Um concordanceador e um contador de freqüências estão presentes em quase todas as ferramentas analisadas. As funcionalidades para buscas orientadas a glossários, processamento de texto anotado e geração de subcórpus foram consideradas por facilitar a criação de buscas elaboradas. A geração de colocações e o tratamento de codificação de caracteres também foram levados em conta.

A análise de usabilidade foi baseada em três métricas: facilidade no uso do concordanceador, presença de documentação e opções de idioma para a interface. O concordanceador foi escolhido por ser utilizado com freqüência em diversos tipos de tarefas e em particular para pesquisas lexicográficas. Nesse caso, optou-se por uma análise objetiva, baseando-se no número de cliques necessários para acessar concordâncias. O concordanceador do *GATE* não é ativado por padrão. Uma vez ativado, seu uso é fácil, mas o processo de ativação (não avaliado aqui) é relativamente complexo. A presença de documentação foi avaliada em uma escala subjetiva, em três níveis: (1) pouca/nenhuma, (2) média e (3) completa. Na análise da interface, foi constatado que as ferramentas estão disponibilizadas apenas em inglês, o que dificulta seu uso a pesquisadores com pouco ou nenhum conhecimento no idioma, embora seja o mais comum, pois o Inglês é a língua franca da ciência.

A eficiência foi analisada através de dois testes: (1) o tempo levado para as ferramentas pré-processarem um córpus e (2) o tempo levado para realizar uma busca e exibi-la no concordanceador. Os testes foram realizados em um computador com uma CPU de 1.5 GHz e 512 MB de memória RAM. Além disso, os requisitos mínimos de *hardware* para a execução das ferramentas foram avaliadas, pois não existe muita informação disponível sobre eles. O tempo de pré-processamento da ferramenta *Tenka* foi considerado como zero, pois a

ferramenta não pré-processa o texto. Como nenhum tipo de indexação é realizado, as buscas tendem-se a tornar-se mais lentas. Os testes foram realizados no córpus contemporâneo PLN-BR *Gold*, contendo 1.024 textos e totalizando 338.441 palavras. Optou-se por um córpus contemporâneo por não demandar configuração adicionais nas ferramentas. As anotações foram removidas antes dos testes, pois nem todas as ferramentas podem processar textos anotados. O cliente e o servidor *Xaira* foram executados no mesmo sistema ao invés de em seu modo de funcionamento distribuído. Da mesma forma, o servidor *Philologic* e o navegador cliente também foram executados no mesmo sistema. O *Philologic* possui uma vantagem no modo cliente-servidor, pois o cliente pode ser um navegador modesto em um sistema com poucos recursos.

A manutenibilidade foi avaliada na perspectiva do usuário ao invés dos desenvolvedores originais das ferramentas. Nesse sentido, o próprio usuário pode se tornar um desenvolvedor, algo possível em softwares livres ou abertos. Nesse caso, é importante observar a licença de distribuição de cada ferramenta, pois traz limitações e restrições de uso. A maioria das ferramentas é distribuída sobre a licença GNU GPL (*GNU General Public License*). No caso do *Philogic*, que pode ser classificado na categoria de software como serviço, é usada a licença *Afero GPL*, mais adequada a esse tipo de software.

Na portabilidade, foram avaliados os sistemas operacionais permitidos por cada ferramenta. O *Philologic* foi considerado o mais portável, pois a Web está disponível em praticamente todas as plataformas. GATE e o *Unitex* também obtiveram uma boa avaliação, devido ao fato da linguagem *Java* funcionar em diversas plataformas. O mesmo se aplica ao *Tenka*, desenvolvido em *C#* (apesar de não ter funcionado em um teste no ambiente *Linux*, diferentemente do indicado pelo desenvolvedor). O *Xaira*, desenvolvido em C++, pode rodar em *Windows* e *Linux* (sem interface no caso do *Linux*). A Tabela 3.3 traz a avaliação comparativa entre as ferramentas. Observa-se que cada ferramenta apresentada possui seus prós e contras.

Tabela 3.3: Comparativo entre as ferramentas

Métrica	**Critério**	**GATE (build 2752)**	**Philologic 3.1**	**Unitex 2.0 beta**	**Tenka 0.1.3.2**	**Xaira 1.23**
funcionalidade	concordanceador	sim	sim	sim	sim	sim
funcionalidade	contador de freqüência	não	sim	sim	sim	sim
funcionalidade	busca orientada a glossário	sim	sim	não	sim	sim

Métrica	**Critério**	**GATE (build 2752)**	**Philologic 3.1**	**Unitex 2.0 beta**	**Tenka 0.1.3.2**	**Xaira 1.23**
funcionalidade	processamento de anotação	sim (XCES)	sim (TEI-Lite)	Parcial (gramatical)	Parcial (gramatical)	sim (TEI ou similar)
funcionalidade	criação de subcórpus	não	sim	não	sim	sim
funcionalidade	colocações ou *n*-gramas	sim	sim	não	não	sim
funcionalidade	codificação de caracteres	UTF-8	UTF-8	UTF-16	UTF, ISO, etc	UTF-8/16
usabilidade	cliques para concordanceador	3	1	5	6	6
usabilidade	nível de documentação	3	3	3	1	2
usabilidade	idiomas da interface	Inglês	Inglês	Inglês	Inglês	Inglês
eficiência	tempo de pré-processamento (segs.)	663	61,5	19,5	0	36,9
eficiência	tempo do concordanceador	212	1,5	8	13,5	0,7
manutenibilidade	licença	GNU LGPL	Affero GPL	GNU GPL	GNU GPL	GNU GPL
Portabilidade	sistema operacional	diversos (java)	diversos (web)	diversos (java)	diversos (C#)	Windows e Linux

A avaliação foi feita em dois momentos, inicialmente com as ferramentas *Philologic*, *Unitex* e *Xaira* e novamente com a inclusão do *Tenka* e do GATE. A segunda avaliação foi feita em parceria com o pesquisador Filipi Silveira que também trabalha com avaliação de ferramentas. A versão beta do *Unitex* foi avaliada por trazer mais recursos que a versão anterior e apresentar boa confiabilidade. O GATE foi avaliado com os recursos padrões, sem a instalação de módulos de terceiros (o único recurso alterado foi a adaptação de seu concordanceador).

Outros comparativos podem ser encontrados em (RAYSON, 2002) (envolvendo 9 ferramentas avaliadas segundo 12 critérios), (SCHULZE et. al, 1994) (mais de 30 ferramentas diferentes avaliadas por diversos critérios) e (UNIVERSITÉ DE GENÈVE, 2006) (tratando ferramentas para auxílio a escrita). O comparativo de Schulze et. al está um pouco defasado em virtude da data do trabalho (1994), pois muitas ferramentas têm sido desenvolvidas desde então.

4 Processamento de córpus históricos para tarefas lexicográficas: problemas e soluções

4.1 Considerações iniciais

Neste capítulo são discutidos os problemas e as possíveis soluções para o processamento do córpus DHPB. As soluções propostas podem ser aplicadas também a outros córpus históricos, sejam eles de língua Portuguesa ou não. Dentre os problemas que podem ser encontrados na compilação de um córpus histórico, é possível citar: a utilização de caracteres que caíram em desuso (Seção 4.2), a grande quantidade de abreviaturas (muitas delas ambíguas) (Seção 4.3), as diversas variações de grafia para uma dada palavra (Seção 4.4), o problema das junções ou contrações (Seção 4.5) e a existência de poucos trabalhos sobre tipologia de córpus para textos históricos (Seção 4.6). A anotação de gêneros em textos históricos segundo uma tipologia permite avaliar o balanceamento e a representatividade do córpus. Adicionalmente, a falta de um ambiente livre e integrado de processamento de córpus e redação de verbetes do dicionário torna a tarefa mais morosa, pois estes possuem particularidades como é o caso das variações de grafia (Seção 4.7).

4.2 Codificação de caracteres para textos históricos

O uso de *Unicode* é particularmente importante em córpus históricos, pois é comum encontrar caracteres não permitidos pelos padrões de codificação usuais. No projeto DHPB, foram identificados três tipos de caracteres que caíram em desuso: ligaduras, consoantes acentuadas e símbolos gerais em Latim. As ligaduras são encontradas no córpus principalmente em expressões em Latim. Um exemplo de ligadura é o símbolo “æ” (união de “a” e “e”). Um exemplo de consoante acentuada é “m̃” em “com̃ercio”. A Tabela 4.1 mostra exemplos de palavras (e abreviaturas) que possuem símbolos que caíram em desuso.

Tabela 4.1: Símbolos encontrados no córpus DHPB

Símbolo	Descrição	Unicode	Freqüência	Exemplo
^	Combining circumflex accent	0302	2	quarŷ (*)
~	Combining tilde	0303	24.568	com̃ercio
¯	Combining macron	0304	596	cacaō
¨	Combining dieresis	0308	48	muÿ
̉	Combining hook above	0309	1.804	sỏmente
̓	Combining comma above	0313	371	tinhaỏ
´	Combining acute accent	0301	2	quaeś
̆	Combining breve	0306	2	apanhĕ
Æ	Latin capital letter AE	00C6	41	Æthyopia (**)
æ	Latin small letter ae	00E6	1.378	græti (**)
œ	Latin small ligature oe	0153	116	Cœteris (**)
§	section sign	00A7	1.131	§ (parágrafo)
ⅎ	turned capital f	2132	4	ⅎixit (**)
ſ	Latin small letter long s	017F	928	Deſcobrio (***)
ƒ	Latin small letter f with hook	0192	149.909	ƒeito (***)
ɐ	Latin small letter turned a	0250	4	passadɐ
&	Ampersand	0026	20.649	&c. (etc.)
@	commercial at	0040	192	@nrique

(*) nome indígena
(**) nomes em latin
(***) freqüência sujeita a revisão.

Os símbolos 017F e 0192 foram usados indistintamente em alguns textos devido a sua semelhança visual, o que gerou freqüências incorretas para esses dois símbolos. Uma revisão futura no córpus deverá resolver esse problema. Estima-se que o símbolo 017F (com som de "S") seja bem mais freqüente que o 0192 (com som de "F").

Símbolos sem representação na codificação de caracteres adotada devem ser substituídos por outros símbolos ou por etiquetas. O padrão TEI possui a etiqueta "<symbol>" para esses casos. Flexor (1991) encontrou documentos com sobreposições entre as letras "A" e "N", indicando a abreviatura de "Antônio". Esses símbolos não podem ser representados através de *Unicode* e precisam ser convertidos ou anotados.

4.3 Tratamento de abreviaturas

Rydberg-Cox (2003) levantou alguns problemas comuns em textos históricos, entre

eles: ausência de hifenização, junções de palavras (exemplo: “éamor”), símbolos tipográficos incomuns para a grafia de palavras e a alta freqüência de abreviaturas. No córpus do projeto DHPB ocorrem os mesmos problemas, e as abreviaturas são empregadas em boa parte dos textos. O uso de abreviaturas é comum em manuscritos antigos e também nos primeiros materiais impressos. O processamento de abreviaturas é particularmente difícil, pois estas são ambíguas e podem ter um grande número de significados. A ambigüidade para as abreviaturas “A” e “B°” é ilustrada na Tabela 4.2.

Tabela 4.2: Ambigüidade de abreviaturas em córpus históricos

Expansões da abreviatura A	**Expansões da abreviatura B°**
alteza, alvará, Amaro, Ana, anima, ano, anos, Antônio, arroba, arrobas, assembléia, assinado, atual, auto, autor, autos, autuado, autue	Bartolomeu, bastardo, bairro, beco, bento, Bernardo, bispo, Botelho, bueno

Embora existam técnicas para expansão automática de abreviaturas para as línguas contemporâneas (TERADA; TOKUNAGA; TANAKA, 2004) e especialmente para o domínio médico (PAKHOMOV, 2002; YU; HRIPCSAK; FRIEDMAN, 2002; SCHWARTZ; HEARST, 2003), há pouca pesquisa sobre o assunto para tratamento de abreviaturas em textos históricos. A presença de abreviaturas limita o poder das ferramentas gerais de extração e recuperação de informação. Além disso, o desempenho de ferramentas como etiquetadores automáticos, concordanceadores e reconhecedores de entidades nomeadas também é limitado e o processo de indexação eletrônica de documentos torna-se mais dispendioso. Caso as abreviaturas não sejam expandidas, algumas pesquisas sobre o córpus podem ser prejudicadas. Em particular, para estudos lexicográficos, é importante encontrar todos os significados e seus contextos em que uma determinada lexia ocorre no córpus, independentemente se essa lexia ocorre de forma abreviada ou não. Um problema extra que deve ser considerado durante a expansão de abreviaturas em textos históricos é a grafia utilizada na expansão, pois é comum encontrar-se palavras com mais de uma grafia em textos históricos devido à inexistência de um sistema ortográfico unificado em períodos anteriores da língua. Uma decisão de projeto poderia ser a utilização da grafia mais freqüente no século em que sua abreviatura ocorre ou a forma mais freqüente utilizada pelo autor do texto.

Se, por um lado, a expansão manual de abreviaturas é uma tarefa demorada e cara, por outro lado, a expansão automática é dificultada devido à presença de ambigüidade. Uma

possível abordagem é o uso de glossários de abreviaturas, com os quais um pesquisador pode formular expressões de busca facilmente para encontrar uma determinada lexia e todas as suas formas abreviadas. Além disso, o glossário também serve de apoio ao pesquisador, quando este se depara com abreviaturas desconhecidas presentes no contexto de buscas realizadas no córpus. Essa abordagem é computacionalmente mais barata, entretanto, traz algumas desvantagens. Por exemplo, uma busca por "bispo" pode remeter a contextos com a palavra "bento" ou "Bernardo", uma vez que as três possuem uma forma abreviada em comum ("Bº"). Para textos nos quais as abreviaturas não foram expandidas, a presença do glossário é um apoio importante, pois uma única palavra pode possuir um grande número de abreviaturas diferentes, como mostra a Tabela 4.3.

Tabela 4.3: Diferentes abreviaturas da lexia composta "Rio de Janeiro"

Abreviaturas
Rio de Jan.ro, Rio de Janro, Rio de Janr.o, Rio de Jan.o, Rio de Jnro, Rio de Janro, Rio de janro

4.4 Detecção automática de variação de grafias

Os textos do córpus DHPB foram escritos em uma época em que não havia um sistema ortográfico unificado para o Português. Devido a isso, pode-se encontrar uma palavra escrita com diferentes grafias no córpus. As variações acontecem até mesmo dentro de um único texto e dificultam as buscas para a tarefa lexicográfica em que é importante recuperar todas as variações de uma lexia. Além disso, os resultados da busca por um padrão tornam-se limitados quando não se conhece todas as variações de grafia da lexia desejada. Uma possível solução para o problema é a construção de glossários contendo as variações de grafia das palavras no córpus. O processo deve ser feito preferencialmente de forma automática, pois a construção manual do glossário é um processo caro para córpus com milhões de palavras.

Hirohashi (2004) apresenta três diferentes abordagens para detecção automática de variações de grafias: agrupamento por distância de edição, análise fonética e regras de normalização aprendidas automaticamente. A técnica de agrupamento por distância de edição se baseia em operações de inserção, remoção e troca de símbolos em palavras. Por exemplo, a grafia "caza" pode ser transformada em "casa" com uma operação de troca de símbolos. Como o número de operações é pequeno, as duas grafias são agrupadas. Está técnica é propensa a erros, por exemplo, a grafia "grade" pode ser agrupada juntamente com "grande".

A ferramenta *Philologic* realiza agrupamento por edição através do utilitário AGREP. A análise fonética consiste em agrupar grafias com a mesma pronuncia como "muito" e "muyto". Essa técnica é computacionalmente cara, e a ausência de um sistema ortográfico unificado em textos históricos dificulta a criação de regras de pronúncia a partir de uma grafia. O uso de regras de normalização é uma abordagem mais simples e menos propensa a erros, na qual são aplicadas regras sobre cada grafia. Por exemplo, é possível formular uma regra que substitui "ph" por "f". Essa regra aplicada à grafia "pharmácia" resulta em "farmácia", gerando um agrupamento entre as duas formas.

Hirohashi desenvolveu um etiquetador morfossintático com informações de variantes levantadas a partir de regras de normalização. As variantes de grafia são anotadas com etiquetas apropriadas e não são perdidas durante o processo de normalização. A ferramenta VARD (*VARiant Detector*) (RAYSON; ARCHER; SMITH, 2005; ARCHER et. al, 2006) foi desenvolvida para detectar e normalizar variantes de grafia de textos em inglês. A detecção de variantes é feita a partir do algoritmo SoundEx e algoritmos de distância de edição, além de um glossário de 45.805 variantes, regras contextuais e heurísticas diversas. A ferramenta RSNSR *Rule-Based Search in Text Data Bases with Non-standard Orthography*) (ARCHER et. al, 2006) foi desenvolvida para textos em alemão e utiliza um mecanismo de busca baseado em regras *fuzzy*. As regras foram obtidas de fontes diversas como analise estatística, materiais históricos e informações lingüísticas.

Menegatti (2002) destaca diversas práticas comuns em textos anteriores ao século XVIII, entre elas: consoantes dobradas, inconsistência no uso de acentuação gráfica e troca entre símbolos. As consoantes dobradas constituem uma prática comum do Latim para representação de vogais longas (como em "anno") e sua prática foi mantida nos primeiros textos em Português. A inconsistência no uso de acentuação gráfica pode ser observada comparando-se diferentes documentos. É possível encontrar documentos nos quais o uso de acentuação gráfica é praticamente inexistente, ou que usam um padrão de acentuação parecido com o atual, ou mesmo que utilizam a acentuação para marcação tônica. A troca de símbolos é alta especificamente para os símbolos "e" e "i" e para os símbolos "o" e "u" (quando a sílaba em questão não é tônica). Os símbolos "j" e "v" também podem ser trocados, por "i" e "u", respectivamente. Essas informações são úteis na construção de regras de normalização.

Giust et. al (2007) propõem uma abordagem de uso de regras de transformação criadas manualmente que podem ser acrescidas ao conjunto inicial por pesquisas nos relatórios do

sistema Siaconf (Sistema de Apoio à Contagem de Freqüência em Córpus), criado para detecção de variação de grafias. As regras de transformação serão discutidas na Seção 5.3.3. A proposta foi baseada no trabalho de Hirohashi (2004) e de Menegatti (2002).

4.5 Junções de palavras

Além da ausência de um sistema ortográfico unificado, também foi observado o uso de junções no córpus DHPB. Os dois principais problemas gerados pelas junções são as distorções na contagem de freqüência e a necessidade de expressões de busca mais sofisticadas no concordanceador, que sejam capazes de recuperar uma palavra dentro de uma junção. A solução mais adequada nesse caso é a separação das junções. O padrão TEI permite a anotação de junções de forma que a versão separada seja inserida no córpus e a versão contraída seja mantida. Para tal, podem ser usadas as etiquetas "<choice>", "<sic>" e "<corr>", como em "<choice><sic> éamor </sic><corr> é amor </corr></choice>".

Foi observado que as junções ocorrem entre preposições e substantivos com relativa freqüência, como em "acargo" e "depernambuco". Entretanto, muitos outros casos acontecem, envolvendo artigos ("ocapitão"), pronomes ("seusfilhos"), apenas substantivos ("FranciscoCoelhoBitancur"), e até casos mais complexos ("seriamaisconveniente"). Em parte dos exemplos, são encontrados casos em que as palavras podem ser diferenciadas por maiúsculas, entretanto não há uma regra, o que dificulta a extração automática das junções. Duas técnicas para o levantamento de junções no córpus incluem: (a) análise manual das formas simples únicas do córpus (*types*) e (b) extração automática através da análise de formas simples do córpus.

Na técnica de extração automática, cada forma simples é dividida em duas através de separação silábica. Por exemplo, a forma "éamor" pode ser dividida em "éa mor" e "é amor". Cada par de formas é então procurado na lista de formas simples e caso as duas formas do par sejam encontradas, uma candidata a junção é criada. Por fim, as candidatas são verificadas manualmente, e as junções são escolhidas. O processo pode ser generalizado para junções de três ou mais palavras. Entretanto, quanto maior o número de palavras por junção, mais caro computacionalmente é o processo de extração.

Após a fase de levantamento das junções, é possível separá-las no córpus. O processo pode ser feito automaticamente de forma simples através de busca e substituição por padrões ou por expressões regulares de substituição.

4.6 Extração automática de metadados

Metadados têm um papel importante na compilação de um córpus, pois descrevem diversas informações acerca dos textos e podem ser utilizados como entradas para as ferramentas de busca. Em particular, quando os metadados domínio e gênero textual não são conhecidos *a priori*, é possível obtê-los da leitura e interpretação dos textos e de outros parâmetros situacionais. Contudo, a tarefa é custosa quando feita manualmente, pois a quantidade de textos que compõe um córpus é, geralmente, grande e alguns textos podem ser extensos. Uma alternativa para extração desses metadados é o uso de técnicas de Inteligência Artificial como a Classificação Automática (SEBASTIANI, 2002). Por exemplo, um classificador automático pode ser construído para classificar um texto em um dentre diferentes gêneros, ou entre diferentes domínios. A Classificação de Textos é uma importante sub-área da Classificação Automática. Aires (2005) apresenta uma série de trabalhos envolvendo classificação automática para classificação de textos em gêneros discursivos. Embora existam diversos trabalhos sobre o assunto, não temos conhecimentos sobre pesquisas para a classificação automática de textos por gêneros e domínios em córpus históricos.

4.6.1 Definição de domínio e gênero

Na área de lingüística aplicada existem diferentes definições para o conceito de gênero, mas a maioria delas está relacionada ao propósito comunicativo dos textos. Além do propósito, Biber (1994) cita outros parâmetros situacionais: (a) características comunicativas dos participantes, (b) relação entre o falante e ouvinte, (c) contexto de comunicação, (d) canal de comunicação, (e) circunstâncias de produção e compreensão do texto, (f) avaliação pessoal do falante e do ouvinte e (g) tópico (ou assunto) do texto. Kauffmann (2005) discute detalhadamente algumas das definições. Para Swales (apud SILVA, 2005, p. 1) um gênero pode ser definido como um conjunto de textos com o mesmo propósito comunicativo:

> "Um gênero compreende uma classe de eventos comunicativos, cujos membros compartilham os mesmos propósitos comunicativos. Tais propósitos são reconhecidos pelos membros especialistas da comunidade discursiva de origem e, portanto, constituem o conjunto de razões (*rationale*) para o gênero. Essas razões moldam a estrutura esquemática do discurso e influenciam e impõem limites à escolha de conteúdo e de estilo."

Para Marcuschi e Xavier (2005) gêneros referem-se a textos com funções comunicativas

bem definidas e sua composição abrange diversos fatores como estilo e composição:

> "(...) realizações lingüísticas concretas definidas por propriedades sociocomunicativas; constituem textos empiricamente realizados cumprindo funções comunicativas; sua nomeação abrange um conjunto aberto e praticamente ilimitado de designações concretas determinadas pelo canal, estilo, conteúdo, composição e função."

Exemplos de gêneros são: telefonema, sermão, carta, bula de remédio, entre outros. Marcuschi (2002) também define o conceito de domínio:

> "(...) uma esfera ou instância de produção discursiva ou de atividade humana. Esses domínios não são textos nem discursos, mas propiciam o surgimento de discursos bastante específicos (...)"

Exemplos de domínios incluem: jurídico, jornalístico, religioso, científico, entre outros. O autor também define o conceito de tipo textual como sendo "(...) uma seqüência teoricamente definida pela natureza lingüística de sua composição (...)". Exemplos de tipos seriam as categorias: narração, argumentação, exposição, descrição, injunção, pareceres técnicos e escrituras.

O projeto LW (ALUÍSIO et al., 2003b), usa a definição de gêneros Swales. Nove gêneros foram criados: científico, de referência, informativo, jurídico, prosa, poesia, drama, instrucional e técnico-administrativo. Além disso, foi criado o super-gênero literário que engloba prosa, poesia e drama. Os textos também foram classificados de acordo com 39 tipos textuais e um genérico chamado "Outros". Exemplos de tipos de texto empregados incluem apostila, declaração, monografia, carta e notícia.

No projeto DHPB, optou-se pela definição de gêneros e domínios de (MARCUSCHI; XAVIER, 2005). Atualmente, 99 gêneros foram definidos para o projeto DHPB (como decreto, contrato, livro texto, sermão, etc) por Jacqueline Souza, aluna de mestrado do Programa de Pós-Graduação em Lingüística (PPGL) da Universidade Federal de São Carlos (UFSCar). Adicionalmente, a definição dos domínios que serão estudados já está concluída e é composta por 9 domínios: religioso, jurídico, científico, informativo, referencial, instrucional, técnico administrativo e/ou oficial, literário e pessoal. Cada domínio está dividido em subdomínios. Os subdomínios são então divididos nos 99 gêneros citados acima, que por sua vez, se dividem em subgêneros. Os domínios, subdomínios, gêneros e subgêneros estão listados no Anexo B.

4.6.2 Técnicas de classificação automática de textos

Muitos sistemas de classificação automática são baseados em aprendizado indutivo. O aprendizado indutivo (MONARD et al., 1997) tem como objetivo reproduzir a capacidade humana de generalizar informações a partir de um subconjunto pequeno de fatos (ou de dados) e se subdivide em duas técnicas: (a) aprendizado supervisionado, no qual o sistema aprende a classificar itens de um grande conjunto a partir de um subconjunto pré-classificado (generalizando a informação de classificação) e (b) aprendizado não supervisionado, no qual o sistema agrupa itens com características semelhantes. O aprendizado supervisionado é particularmente útil para extração de metadados, pois permite a construção de classificadores automáticos para um conjunto prévio de categorias.

É possível associar os valores que um metadado pode assumir às classes utilizadas por um classificador. Por exemplo, para o metadado "data de edição" é possível criar as classes "século XVI", "século XVII" e "século XVIII". Um classificador pode ser treinado com um pequeno conjunto de textos datados desses três séculos e então ser utilizado para identificar novos textos sem datação conhecida criados durante esses três séculos. Esse exemplo poderia ser aplicado como trabalho futuro ao projeto DHPB, no qual 86 textos foram compilados sem informação de datação. A maior parte das pesquisas feitas para classificação automática de textos se baseia na análise de traços lingüísticos (ou características lingüísticas) presentes nos textos.

Em geral, os classificadores não são capazes de processar textos em sua forma original. É preciso converter um texto para uma representação matemática (ou modelo matemático) de seus traços. Esse modelo, por sua vez, é analisado por um classificador. Por exemplo, o pronome de tratamento "você" pode ser um bom traço lingüístico para discriminar cartas e resumos de artigo. É mais provável que um texto em que o traço assuma o valor 3 (ou seja, o pronome aparece três vezes) seja uma carta do que um resumo de artigo científico, pois o pronome é mais comum em textos informais. Um classificador pode analisar diversos traços lingüísticos. Para isso, o classificador tem como entrada vetor de traços da forma $(t_1, t_2, ..., t_n)$, em que o elemento t_i representa o valor assumido pelo traço i para algum documento. Um vetor é associado a cada documento do córpus.

Em geral, os classificadores não são totalmente precisos, mas boa parte deles possui uma precisão alta. Para o caso particular de classificação de textos históricos, um pequeno percentual de textos classificados incorretamente é aceitável quando a classificação manual

com precisão de 100% torna-se muito custosa. A precisão de um classificador depende do subconjunto de dados utilizado em seu treinamento, do seu algoritmo de classificação e, principalmente, da forma com que os dados são modelados matematicamente. Para classificação de textos, isso quer dizer que os traços devem ser escolhidos criteriosamente. Quanto melhor a escolha de traços lingüísticos, melhor será o desempenho do classificador.

Em particular, para a criação de um classificador de gêneros e domínios, é necessária a definição de quais gêneros e domínios serão analisados. É preciso então adotar uma definição formal sobre esses conceitos.

No padrão TEI, apresentado na Seção 2.6.1, não há etiquetas explicitas para a definição de gêneros e domínios. Entretanto, há etiquetas para a criação e uso de taxonomias. A etiqueta "<taxonomy>", usada dentro da estrutura "<classDecl>", define uma nova taxonomia. As taxonomias são definidas na seção *encoding description* (descrição da codificação) do cabeçalho TEI. A classificação do texto segundo taxonomias é feita na seção *text profile* (perfil do texto) do cabeçalho TEI com a etiqueta "<catRef>". Adicionalmente, a etiqueta "<keywords>" pode ser utilizada para classificar um texto por assuntos. A Figura 4.1 mostra exemplos de assuntos para um texto do projeto PLN-BR.

```
<catRef target="genero.8 genero.8.18 genero.8.18.10  distribuicao.12 tipotextual.35 " />
<keywords>
  <keyTerm>VIOLÊNCIA</keyTerm>
  <keyTerm>AGRESSÃO</keyTerm>
  <keyTerm>ATAQUE</keyTerm>
  <keyTerm>JOGADOR</keyTerm>
  <keyTerm>FUTEBOL</keyTerm>
  <keyTerm>CORINTHIANS</keyTerm>
  <keyTerm>CLUBE</keyTerm>
  <keyTerm>GAVIÕES DA FIEL</keyTerm>
  <keyTerm>TORCIDA ORGANIZADA</keyTerm>
</keywords>
```

Figura 4.1: Exemplo de assuntos para um texto do PLN-BR

Existem pesquisas sobre análise de gêneros que empregam tanto técnicas de aprendizado supervisionado, quanto técnicas de aprendizado não supervisionado. O trabalho pioneiro na análise automática de gêneros textuais foi proposto por Biber (1988, 1993b, 1994, 1995) e baseado na técnica matemática Análise Multidimensional (aprendizado não supervisionado). Biber estudou os gêneros a partir de cinco dimensões:

1. Produção com interação *versus* produção informacional.

2. Discurso narrativo *versus* não-narrativo.
3. Referências explícitas *versus* referências dependentes do contexto.
4. Expressão explícita de argumentação.
5. Estilo impessoal *versus* não-impessoal.

Biber constatou que textos pertencentes ao mesmo registro (terminologia utilizada pelo autor com o mesmo sentido de gênero) tendem a se agrupar em regiões próximas no espaço *n*-dimensional formado pelas cinco dimensões definidas. Apesar de se tratar de uma técnica de aprendizado não supervisionado, é possível usá-la também para classificação automática. Para isso, o autor usa a técnica de análise de discriminante (LEWICKI; HILL, 2008), na qual o espaço *n*-dimensional é inicialmente povoado com um subconjunto de textos pré-classificados (de forma semelhante ao aprendizado supervisionado). Novos textos (sem classe conhecida *a priori*) são adicionados ao espaço e recebem como rótulo a classe dos textos pré-classificados mais próximos.

O cálculo da posição dos textos em cada uma das dimensões é feito de acordo com regras de pontuação. Essas regras se baseiam na presença ou ausência de traços lingüísticos nos textos. Por exemplo, na dimensão 1 os verbos pessoais, pronomes de primeira pessoa, palavras de tamanho grande e substantivos são pontuados respectivamente com os valores +0,96, +0,74, -0,58 e -0,80. Para esse caso, os valores positivos são mais comuns em textos com produção com interação, enquanto que os valores negativos são mais freqüentes em textos com produção informacional.

Em (KAUFFMANN, 2005) é feito um trabalho empregando a análise fatorial, uma técnica baseada na Análise Multidimensional, de forma semelhante à metodologia proposta por Biber. A diferença em relação ao trabalho de Biber reside no fato de o córpus de estudo estar em Língua Portuguesa ao invés de Língua Inglesa. O córpus é constituído por textos referentes a uma amostra de uma semana construída do jornal "A Folha de São Paulo". Uma semana construída consiste em sete edições do jornal de diferentes períodos do ano e de diferentes dias semanais.

Foram utilizados 15 gêneros diferentes aplicáveis a textos jornalísticos, alguns exemplos são: artigo, carta, chamada, coluna de notas, comentário e crítica. Os gêneros foram distribuídos entre duas dimensões: a dimensão argumentativo *versus* informativo e a dimensão expositivo *versus* narrativo. Um dos experimentos foi realizado com 14 traços

(variáveis, na terminologia da análise multidimensional) utilizados para calcular o valor das dimensões. Alguns exemplos de traços são o número de substantivos no texto e a presença de verbos no pretérito do indicativo.

Em (AIRES, 2005), técnicas de classificação automática de textos a partir de aprendizado supervisionado são empregadas como auxílio para a execução de buscas em sistemas de recuperação de informação. O sistema proposto é capaz de classificar textos em língua Portuguesa segundo gêneros, tipos de texto, sete necessidades gerais de busca na *Web* e necessidades de busca personalizadas. A classificação de gêneros e tipos textuais foi a mesma utilizada no Projeto LW. Parte do córpus de treinamento também foi o mesmo do projeto LW. Entretanto, esse córpus foi posteriormente ampliado por Aires com textos da *Web*. O trabalho faz um estudo aprofundado sobre diversos traços lingüísticos e diversos algoritmos de aprendizado de máquina são testados em um córpus de treinamento. Os experimentos realizados permitiram encontrar os traços mais importantes para a classificação de textos. Exemplos de traços utilizados para classificação dos textos são o tamanho médio das palavras, o tamanho do texto em caracteres e a ocorrência de expressões como “acho”, “acredito que” e “parece que”. O número de traços utilizadas nos experimentos variou de 46 a 135, de acordo com o conjunto de textos utilizados.

Além de mostrar os traços mais relevantes, o trabalho também apresentou uma avaliação dos algoritmos mais precisos para a classificação de textos. No total, 44 algoritmos do ambiente *Weka* (GARNER, 1995) foram analisados, sendo que os algoritmos J48, SMO e LMT obtiveram um desempenho satisfatório. A taxa de acerto do classificador chegou a índices de 95% para gêneros (com o mesmo número de textos para cada gênero) e 75% para tipos.

É possível observar que os traços lingüísticos são convertidos para vetores matemáticos tanto no trabalho de Aires, quanto no trabalho de Kauffmann. Mas existe uma diferença importante além da estratégia de aprendizado utilizada nos dois trabalhos. No trabalho de Aires, cada coordenada do vetor corresponde diretamente a um traço lingüístico. Já no trabalho de Kauffmann, diversos traços são combinados para a criação de uma única coordenada, o que resultou na utilização de apenas duas dimensões. A vantagem dessa estratégia é a redução de dimensionalidade, obtida graças à análise multidimensional, gerando vetores mais fáceis de se manipular matematicamente. Por outro lado, a técnica reduz o nível de informação semântica dos dados, o que não acontece no trabalho de Aires. O uso de

classificação por gênero e domínio no córpus DHPB será deixar como trabalho futuro.

4.7 Auxílio à redação de verbetes

Sistemas de apoio a pesquisas lexicográficas e terminológicas provêem auxílio para tarefas diretamente relacionadas à terminologia e à lexicografia, como por exemplo, extração automática de termos e gerenciamento de bases de dados lexicográficas. Em (UNIVERSITÄT LEIPZIG, 2007) são descritas diversas ferramentas de apoio a pesquisas lexicográficas e terminológicas, a maioria voltada para terminologia. Haddad (1999) constatou que sistemas de apoio a tarefas lexicográficas e terminológicas de prateleira são pouco utilizados na indústria de tradução canadense. Na maior parte dos casos, softwares específicos são desenvolvidos para cada projeto, o que sugere que ferramentas de prateleira não são amplamente difundidas para pesquisas lexicográficas e terminológicas em geral. É importante notar que softwares comerciais desse tipo são, em geral, caros.

O processamento de base de dados lexicográficas e terminológicas é fornecido pela maioria das ferramentas de apoio à lexicografia e à terminologia, e é importante para pesquisas envolvendo a criação de dicionários (de uso geral da língua, de tradução ou de terminologia), pois a velocidade de acesso às bases de dados propicia um ganho de produtividade ao redator durante a tarefa de redação de verbetes. Um exemplo de sistema com essa funcionalidade é a ferramenta *System Quirk* (AHMAD, 1994). O sistema é dividido em módulos e conta com o módulo *Browser/Refiner*, responsável pelo gerenciamento de bases de dados terminológicas. Para o Português, existe o Corpógrafo e, atualmente, está sendo desenvolvido no NILC o ambiente E-termos (ALMEIDA; OLIVEIRA; ALUÍSIO, 2006).

A partir do estudo de caso envolvendo o projeto DHPB, é possível projetar a estrutura de uma base de dados lexicográfica para dicionários históricos. Para uma dada lexia, a base de dados precisa armazenar diferentes acepções (ou definições) e armazenar as diferentes variações de grafia da palavra em questão. Cada acepção, por sua vez, é acompanhada por abonações (excertos do córpus nos quais a palavra aparece). Adicionalmente, é preciso associar cada abonação a seu texto de origem. Para isso, é necessário um bom sistema de referência (utilizando, por exemplo, a data de produção do texto e a página da citação), capaz de identificar unicamente cada texto do córpus. Por exemplo, é possível criar referências bibliográficas em formato ABNT (Associação Brasileira de Normas Técnicas). Outro ponto importante, é que um mesmo texto pode ser utilizado em mais de uma abonação.

5 Uma metodologia para a criação de recursos e ferramentas para tarefas lexicográficas em córpus de Português Histórico

5.1 Considerações iniciais

Este capítulo contém a metodologia utilizada para a compilação e uso do córpus DHPB, além de descrever o processo de redação de verbetes. A metodologia apresentada aqui foi baseada nas decisões do grupo de pesquisadores em reuniões do projeto. As reuniões foram organizadas com periodicidade média de seis meses. Inicialmente, as reuniões eram focadas no dicionário histórico. Entretanto, percebeu-se que o córpus gerado também é uma contribuição importante do projeto. As decisões para a compilação do córpus são mostradas na Seção 5.2. Um dos problemas encontrados inicialmente foi a alta freqüência fenômenos como abreviaturas, junções de palavras e variantes de grafia no córpus (detalhados na Seção 5.3). Para o início da fase de redação de verbetes, era ideal contar com um ambiente *Web* que possuísse as funcionalidades do *Philologic* e do *Unitex*. Entretanto, com a urgência em oferecer o córpus e com a contemporaneidade do projeto de mestrado com o projeto DHPB (seu motivador), optou-se por oferecer o córpus para os dois pré-processadores (diferentes versões do córpus foram lançadas a medida que a fase de compilação avançava). A Seção 5.4 descreve a adaptação das duas ferramentas para atender as necessidades do projeto DHPB. Por fim, a Seção 5.5 descreve informações sobre o processo de redação de verbetes. As ferramentas e os glossários apresentados neste capítulo estão disponíveis publicamente[13].

5.2 Pré-processamento do córpus

As tarefas de pré-processamento do córpus DHPB são a limpeza e a anotação dos textos digitalizados. Os textos são digitalizados em formato DOC e convertidos para um formato XML simplificado. O formato XML, por sua vez, permite a geração de córpus em XML-TEI ou texto com informações catalográficas, usados nos processadores de córpus *Philologic* e *Unitex*, respectivamente. No *Unitex*, usou-se apenas o título da obra e nome do autor (informações extraídas do cabeçalho das páginas). Para a conversão de DOC para XML foi

13 http://www.nilc.icmc.usp.br/~arnaldo/dhpb/

desenvolvida a ferramenta Protew (PROcessador de TExtos históricos em *Word*). Entretanto, a ferramenta apresentou um baixo desempenho no tempo de conversão dos arquivos. Para solucionar esse problema, o processo foi dividido em duas partes. Inicialmente, o arquivo DOC é convertido para texto puro (TXT) e sua ficha catalográfica é extraída. A seguir o arquivo TXT é convertido para um formato de XML simplificado. A vantagem do uso de XML simplificado é a possibilidade de geração de versões do córpus para diferentes processadores de córpus.

Duas novas ferramentas foram derivadas do Protew original com foco na melhoria do desempenho: o Protew-lite para conversão do DOC para TXT (semelhante ao Protew original, porém com menos recursos) e o Protej (PROcessador de TExtos históricos em Java), com o papel de processar o TXT gerado pelo Protew-lite e a ficha catalográfica extraída. O processo é ilustrado na Figura 5.1. Os retângulos representam as ferramentas e folhas representam os diferentes formatos dos textos.

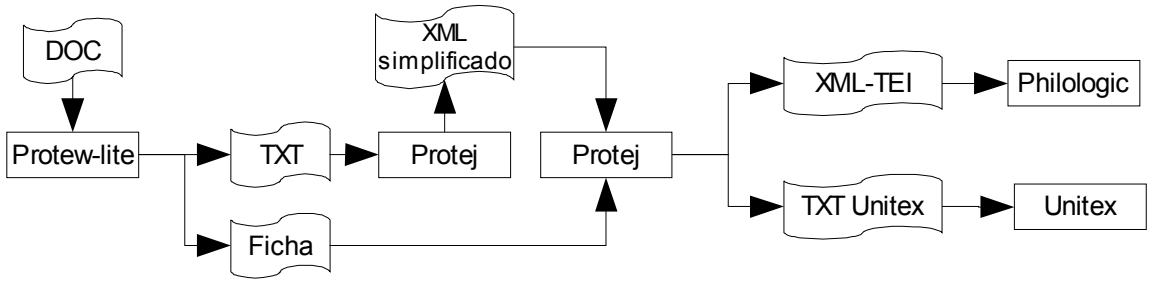

Figura 5.1: Pré-processamento do córpus DHPB

Para o desenvolvimento das ferramentas Protew, Protew-lite e Protej foi realizada a análise de requisitos, uma técnica clássica da engenharia de software, usada para levantar característica que um software deve atender. Os requisitos levantados foram:

- As ferramentas devem remover informações estruturais irrelevantes do texto como numeração de linhas ou de parágrafos.
- As ferramentas devem etiquetar informações estruturais relevantes para a tarefa lexicográfica como números de página.
- As ferramentas devem extrair metadados da ficha catalográfica automaticamente.
- As ferramentas devem ser capazes de gerar diferentes versões do córpus para diferentes processadores.
- As ferramentas devem extrair abreviaturas automaticamente dos textos e processar

bases de dados de abreviaturas como a base descrita em (FLEXOR, 1991).

As ferramentas Protew-lite e Protej foram apresentadas aos participantes no II Encontro do Projeto DHPB em Julho de 2006 na Universidade Estadual Paulista Júlio de Mesquita Filho (UNESP), campus de Araraquara.

5.2.1 Conversão para texto puro

Durante a conversão automática dos arquivos DOC para o formato texto, foram observados alguns problemas:

- Os metadados contidos no cabeçalho e do rodapé dos textos estão mesclados com o texto. Exemplos de metadados comuns na região do cabeçalho ou do rodapé incluem o título da obra, o nome do autor, o número da página e notas de rodapé. Uma vez mesclados, torna-se difícil extrair metadados interessantes como a numeração de página de forma automática. Além disso, os metadados causam distorção na contagem de freqüências do córpus.
- A tabela contendo a ficha catalográfica é convertida para texto e os metadados da ficha são mesclados com o texto. Esse processo também pode gerar distorção na contagem de freqüências, e dificulta a geração automática do cabeçalho TEI, baseada nas informações da ficha catalográfica.
- A formatação de sobrescrito é removida. Essa formatação é útil para a identificação de abreviaturas e deve preservada através de anotação adequada ao processamento do córpus.

A ferramenta Protew-lite trata os casos acima, executando algumas operações em série para um conjunto de arquivos e convertendo-os para texto puro, automaticamente. Os processamentos são realizados diretamente no *MS Word*, graças ao uso da Linguagem VBA (*Visual Basic for Applications*), utilizada no desenvolvimento do Protew-lite. Essa linguagem é incluída nos aplicativos do *MS Office*. As principais tarefas realizadas pelo Protew-lite são:

- Segmentação de páginas e de rodapés. As divisões entre as páginas são identificadas para separar cabeçalhos (a primeira linha da página) e rodapés (a(s) última(s) linha(s)) do restante do texto. Nos casos em que existem notas de rodapés, a

segmentação é facilitada, pois o marcador "---" (seqüência de traços) é inserido manualmente no texto durante o processo de OCR.

- Tratamento de sobrescrito: a única formatação do texto mantida no córpus DHPB é a presença de símbolos em sobrescrito, pois são importantes para o estudo de abreviaturas. As ocorrências de sobrescrito são processadas e denotadas pelo símbolo "^" (circunflexo) antes da etapa de remoção de formatação dos textos. Por exemplo, a abreviatura "jan^{ro}" é convertida para "jan^ro". Inicialmente, optou-se pelo uso desse símbolo ao invés da etiqueta "<sup>" (recomendada pelo padrão TEI), pois dessa forma os textos podem ser processados por ferramentas sem tratamento de anotação como o *Unitex*. Além disso, é possível converter com facilidade os símbolos de circunflexo para etiquetas "<sup>" se se desejar utilizar o padrão TEI. Essa etapa é ilustrada na Tabela 5.1, para quatro abreviaturas apresentadas em negrito.

Tabela 5.1: Exemplo de tratamento de sobrescrito

Antes da remoção de sobrescrito
(...) apartida de belem como vosa alteza sabe foy **seg^{a}** feira ix demarço. e sabado xiij do dito mes amtre as biij e ix oras nos achamos amtre as canareas mais perto da gram canarea e aly amdamos todo aquele dia em calma avista delas obra de tres ou quatro legoas. e domingo xxij do dito mes aas x oras pouco mais ou menos ouuemos vista dasjlhas do cabo verde. s. dajlha de sã njcolaao **seg.^{o}** dito de **p°** escolar piloto. e anoute segujmte **aaseg^{da}** feira lhe (...)
Após a remoção de sobrescrito
(...) apartida de belem como vosa alteza sabe foy **seg^a** feira ix demarço. e sabado xiij do dito mes amtre as biij e ix oras nos achamos amtre as canareas mais perto da gram canarea e aly amdamos todo aquele dia em calma avista delas obra de tres ou quatro legoas. e domingo xxij do dito mes aas x oras pouco mais ou menos ouuemos vista dasjlhas do cabo verde. s. dajlha de sã njcolaao **seg.^o** dito de **p^o** escolar piloto. e anoute segujmte **aaseg^da** feira lhe (...)

- Remoção de formatação da ficha catalográfica e conversão para formato texto: a remoção de formatação é uma operação simples que inclui a remoção de formatação como negrito, itálico, tamanhos de fonte diferentes, etc. As imagens são removidas e as tabelas são convertidas para um formato de texto. A presença de imagens em textos históricos é rara. A presença de tabelas é mais comum e as informações presentes em tabelas complexas (nas quais o número de células varia de coluna para coluna) podem tornar-se desconexas quando estas são convertidas para formato texto. Após a

remoção de formatação, a ficha catalográfica é temporariamente removida e o texto é então armazenado em formato TXT.

- Extração da ficha catalográfica. A extração da ficha é feita em uma etapa após a conversão de DOC para TXT. Em uma segunda etapa, as fichas dos textos são extraídas e armazenadas em arquivos separados.

A Figura 5.2 mostra a interface da ferramenta Protew-lite. É possível observar, por exemplo, a aplicação das operações de tratamento de sobrescrito e remoção de formatação em um texto (além de um tratamento operacional para páginas e rodapés). Além da remoção de sobrescrito, também são tratados símbolos *Unicode* que representam a formatação sobrescrito como os símbolos "ª" (00AA) e "º" (00B0).

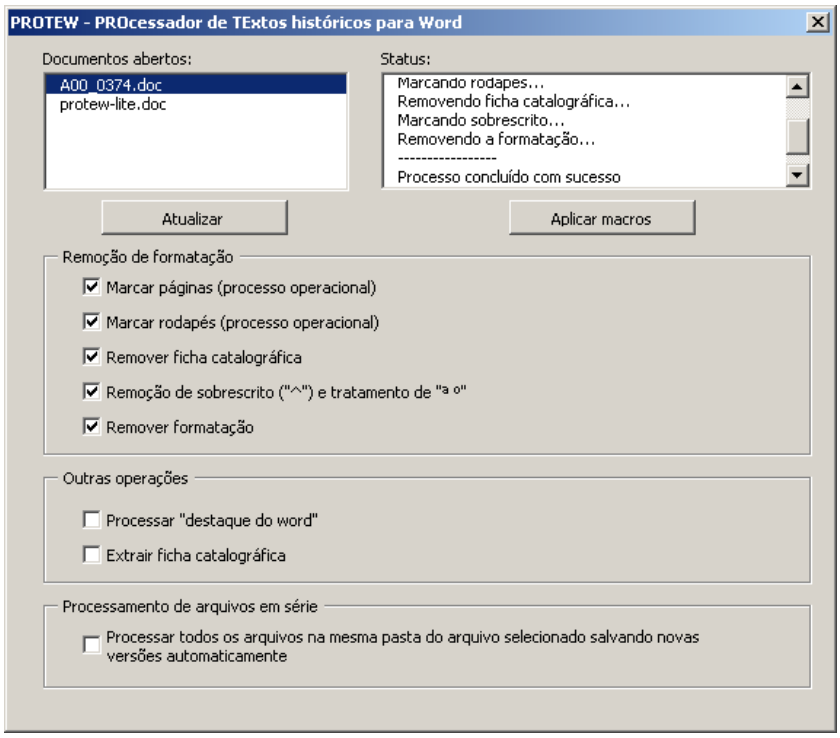

Figura 5.2: A ferramenta Protew-lite

5.2.2 Conversão para XML simplificado

A conversão é feita com o auxílio da ferramenta Protej. A parte mais custosa do processamento é feita nessa etapa, pois a linguagem *Java* mostrou-se consideravelmente mais rápida que a VBA durante o processamento dos textos. As tarefas realizadas pela ferramenta incluem 8 etapas elencadas abaixo:

1. Etiquetação da ficha catalográfica: a ficha extraída pela ferramenta Protej é analisada e convertida para o formato TEI (na forma de cabeçalho). A parte superior da Figura 1.5 mostra o resultado da conversão da ficha catalográfica. Os metadados bibliográficos extraídos do texto são autor, título, data de criação e data de publicação.
2. Tratamento de expressões numéricas com "I": é comum encontrar a letra "I" em expressões numéricas (como "2II" ao invés de "211"). Isso acontece em partes devido ao estilo no qual alguns textos eram escritos e em partes devido a problemas de digitalização. Em ambos os casos, a presença do símbolo acaba distorcendo algumas estatísticas sobre o uso de números no córpus. Para evitar esse problema, o símbolo é convertido automaticamente para "1". A Tabela 5.2 mostra um texto tratado com cinco expressões numéricas convertidas.

Tabela 5.2: Exemplo de tratamento de expressões numéricas

Antes do tratamento de expressões numéricas
(...) com a revolta e **II5** fiquarão muitos feridos e mui maltractados, dos quaes depois diserão que morrera hum ou dous; e ainda hé muito de espantar como não arebentavão os filhos. O numero dos que se então bautizarão foy novecentos menos oyto, e casais em ley de graça, sendo o primeiro **I20** [**II7r**] bautismo solemne que naquella Aldea se fizera, e foy a **I2** de Outubro de **I56I** (...)
Após o tratamento de expressões numéricas
(...) com a revolta e **115** fiquarão muitos feridos e mui maltractados, dos quaes depois diserão que morrera hum ou dous; e ainda hé muito de espantar como não arebentavão os filhos. O numero dos que se então bautizarão foy novecentos menos oyto, e casais em ley de graça, sendo o primeiro **120** [**117r**] bautismo solemne que naquella Aldea se fizera, e foy a **12** de Outubro de **1561** (...)

3. Remoção de hifenização denotada por sinal de igual: em alguns documentos, o hífen foi utilizado exclusivamente em lexias compostas e a hifenização foi feita com o uso de sinais de igual ("="). Nesses casos, foi possível remover a hifenização automaticamente. A remoção de hifenização é ilustrada na Tabela 5.3, para três casos de uso do símbolo "=".

Tabela 5.3: Exemplo de remoção de hifenização denotada pelo sinal "="

Antes da remoção de hifenização
(...) hostilidadez, E mortes que sem se IHe dar Cauza tem **ex=** **ecutado** no Rio da Madeira o Gentio da Nasçaô **Mu=** **ra**, impedindo o Comersio dos Moradores naquelle Rio, E pondo em temor, E consternaçaô as Missoenz **esta=** **blecidas** nelle. Ordeno ao Dout.or Ouvidor geral desta (...)
Após a remoção de hifenização
(...) hostilidadez, E mortes que sem se IHe dar Cauza tem **executado** no Rio da Madeira o Gentio da Nasçaô **Mura**, impedindo o Comersio dos Moradores naquelle Rio,E pondo em temor, E consternaçaô as Missoenz **establecidas** nelle. Ordeno ao Dout.or Ouvidor geral desta (...)

4. Etiquetação de numeração de página: a etiquetação de páginas é um processo relativamente simples, dado que na maior parte dos documentos, o número da página é sempre a primeira ocorrência numérica dentro da página (numeração superior) ou sempre a última (numeração inferior). A ferramenta Protej etiqueta automaticamente a numeração (e informações extras sobre a página) com o uso de chaves, como é mostrado Tabela 5.4. As chaves permitem que as informações estejam presentes em córpus usados com o *Unitex*. Como trabalho futuro, as chaves deverão ser convertidas para as etiquetas TEI correspondentes para permitir seu intercâmbio em formatos amplamente usados.

Tabela 5.4: Exemplo de etiquetação de numeração de páginas

Antes da etiquetação de numeração de página
(...) **44. -PERNAMBUCO 4 DE JUNHO DE 1552 323** mym, tam falto de virtudes, Nosso Senhor tanto faz nesta terra? Estavão sperando, com a sperança que o Padre lhes deu, por hum Padre que fosse letrado e pregador, porque esta fama de letrado faz muyto ao proposito. (...)
Após a etiquetação de numeração de página
(...) **{44. -PERNAMBUCO 4 DE JUNHO DE 1552 323,.N}** mym, tam falto de virtudes, Nosso Senhor tanto faz nesta terra? Estavão sperando, com a sperança que o Padre lhes deu, por hum Padre que fosse letrado e pregador, porque esta fama de letrado faz muyto ao proposito. (...)

5. Conversão de notas de rodapé: antes que a etiquetação de notas seja feita, as notas de um documento são normalizadas, ou seja, convertidas para um formato padrão. O processo é necessário, pois o formato de exibição de notas varia muito de documento para documento. As notas podem ser numeradas por algarismos arábicos (1, 2, 3, ...),

letras do alfabeto (a, b, c, ...) ou símbolos diversos (*, **, ***, ...). Todas as estratégias de numeração são convertidas para algarismos arábicos. Além disso, a numeração pode aparecer no formato sobrescrito ou entre parênteses. Nesse caso, todo o tipo de numeração de notas é convertido para o formato de sobrescrito. Outro processamento importante é a conversão de notas que ocupam múltiplas páginas. Uma nota pode ser dividida em mais de uma página quando seu conteúdo é muito extenso. Nesse caso, a nota deve ser removida e reinserida em uma única página. Uma heurística simples é aplicada: os rodapés das páginas seguintes são analisados, e rodapés que não iniciam por número (o esperado para notas) são tratados como continuação da nota da página anterior. A Tabela 5.5 ilustra o processo de conversão para notas alfabéticas, “a” e “b”.

Tabela 5.5: Exemplo de conversão de notas

Antes da conversão de notas
(...) São os contratos dos príncipes leis, e suas condições com tanta eficácia que os mesmos príncipes contraentes não podem encontrar nem modificar o que neles prometeram e estipularam,**^a** e neles nada pode inovar-se.**^b** E quando não é lícita qualquer alteração ao príncipe (...) _____ **a** Roland. a Valle, consil. 25, n. 35; Amaya, in leg. fin,. cod. de ann. et trib., libro 10, n. 2, 3, e 4. **b**.Partium alterutra renuente, Pelaes de maior. Hispan. P. 1. q. 28, n. 14 (...)
Após a conversão de notas
(...) São os contratos dos príncipes leis, e suas condições com tanta eficácia que os mesmos príncipes contraentes não podem encontrar nem modificar o que neles prometeram e estipularam,**^1** e neles nada pode inovar-se.**^2** E quando não é lícita qualquer alteração ao príncipe (...) _____ **1** Roland. a Valle, consil. 25, n. 35; Amaya, in leg. fin,. cod. de ann. et trib., libro 10, n. 2, 3, e 4. **2**.Partium alterutra renuente, Pelaes de maior. Hispan. P. 1. q. 28, n. 14 (...)

6. Etiquetação de notas de rodapé: durante a etiquetação de notas de rodapé, as notas são removidas do rodapé (no fim da página ou no fim do documento) e inseridas no texto já em formato TEI. As notas podem referenciar três estruturas de texto diferentes: palavras, linhas e parágrafos. Para cada nota que referencia palavra, há uma palavra na mesma página sucedida pelo número da nota. Nesse caso, a inserção da nota é simplificada. Para notas que referenciam linhas, geralmente a linha referenciada pela nota não está numerada, pois em boa parte dos textos as linhas são

numeradas por múltiplos de 5. Por exemplo, é possível para um determinado documento a existência de uma nota que referencie a linha 8 e que esse mesmo documento possua apenas as linhas 5 e 10 numeradas. Uma possível estratégia para obter precisamente a posição da linha 8 é analisar o número de quebras de linha entre as linhas 5 e 10. Contudo, essa estratégia não é confiável, pois nem sempre a ferramenta de digitalização insere quebras de linha corretamente. Nesse caso, é calculada a posição aproximada da linha 8 a partir da média de palavras entre as linhas 5 e 10. Notas que referenciam parágrafos são mais raras e têm sido tratadas manualmente. É possível que um mesmo documento possua notas de palavras e de linhas. Nesse caso, algumas heurísticas simples foram utilizadas para diferenciar um tipo de nota do outro. A etiquetação é mostrada nas tabelas 5.6 e 5.7.

Tabela 5.6: Exemplo de etiquetação de notas que referenciam palavras

Antes do processamento de notas que referenciam palavras
huma maneira de igreja**^3**, junto da (...) os nossos determinamos de hos confessarna nao. 2. Ho primeiro domingo que dissemos missa foy a 4.^a dominga da Quadragessima **^4**. Disse eu missa cedo (...) _____ **3 Esta «maneira de igreja» ou (...)** **4 3I de Março de I549.**
Após o processamento de notas que referenciam palavras
huma maneira de igreja**<note place="foot"n="3"> Esta «maneira de igreja» ou ermida foi a primeira origem da Igreja da Graça. LEITE I 20; II 312 ; CALMON, História da Fundação da Bahia 101; VAN DER V AT, Princípios 296. </note>**, junto da (...) os nossos determinamos de hos confessarna nao. 2. Ho primeiro domingo que dissemos missa foy a 4.^a dominga da Quadragessima **<note> place="foot“ n="4"> 31 de Março de 1549. </note>** . Disse eu missa cedo (...)

Tabela 5.7: Exemplo de etiquetação de notas que referenciam linhas

Antes do processamento de notas que referenciam linhas	
(...)	da pose vyrem: Como, no anno do nacimento
5	de Nosso Senhor Jesu Christo de mil e quynhentos
	e sessenta 1 anos, haos dose dias do mes d'Aguosto do dyto (...)
10	campo e borda do matto, Fernão Jorge juiz hordinario
	dyta vila e campo, ante my apareceu ho Irmão (...)
_____	
6	**Impresso setenta em vez de sessenta. (...)**
Após o processamento de notas que referenciam linhas	
(...)	da pose vyrem: Como, no anno do nacimento
5	de Nosso Senhor Jesu Christo de mil e quynhentos
	e sessenta 1 anos, haos dose dias do mes d'Aguosto do dyto
	<note place="foot"n="6" type="line">Impresso setenta (...) </note>
	(...)
10	campo e borda do matto, Fernão Jorge juiz hordinario
	dyta vila e campo, ante my apareceu ho Irmão (...)

7. Remoção de numeração de linhas ou de parágrafos: consiste na remoção de numeração de linhas ou de parágrafos. A numeração cria algumas distorções nas estatísticas geradas pelas ferramentas de processamento de córpus (exemplo: imprecisão na contagem do total de palavras). A remoção é feita automaticamente sobre números múltiplos de 5. Contudo, nem sempre é possível remover todos os números, pois em alguns casos há números repetidos (o número da linha e um numeral qualquer pertencente ao texto). É feita uma análise no texto para remoção manual da numeração que não pode ser removida automaticamente. A remoção de numeração de linhas é mostrada na Tabela 5.8.

Tabela 5.8: Exemplo de remoção de numeração de linhas

Antes da remoção da numeração de linhas	
	da pose vyrem: Como, no anno do nacimento
5	de Nosso Senhor Jesu 5 Christo de mil e quynhentos
	e sesenta 1 anos, haos dose dias do mes d'Aguosto do dyto (...)
	<note place="foot"n="6" type="line">Impresso setenta (...) </note>
10	campo e borda do matto, Fernão Jorge 2 juiz hordinario
	dyta vila e campo, ante my apareceu ho Irmão (...)
Após a remoção da numeração de linhas	
	da pose vyrem: Como, no anno do nacimento
	de Nosso Senhor Jesu 5 Christo de mil e quynhentos
	e sesenta 1 anos, haos dose dias do mes d'Aguosto do dyto (...)
	<note place="foot"n="6" type="line">Impresso setenta (...) </note>
	campo e borda do matto, Fernão Jorge 2 juiz hordinario
	dyta vila e campo, ante my apareceu ho Irmão (...)

8. Conversão de quebras de linha: trata-se de um processo operacional para simplificar o processamento das outras tarefas. As quebras de linha são convertidas para o padrão POSIX (símbolo *Unicode* 000C).

A maior parte dos processos descritos acima não pôde ser totalmente automatizada, pois ocorrem situações no texto que demandam análise humana. Por exemplo, em alguns textos há notas sem posição de inserção definida e há páginas sem numeração. Nesses casos, o processamento é semi-automático, ou seja, o usuário deve revisar o texto para corrigir os problemas levantados pelas ferramentas. A Figura 5.3 mostra a interface da ferramenta Protej. É possível observar a janela de visualização de texto (ao fundo) antes da aplicação da conversão de notas que referenciam palavras para XML. A janela de seleção (à frente) mostra algumas das tarefas para o processamento de notas.

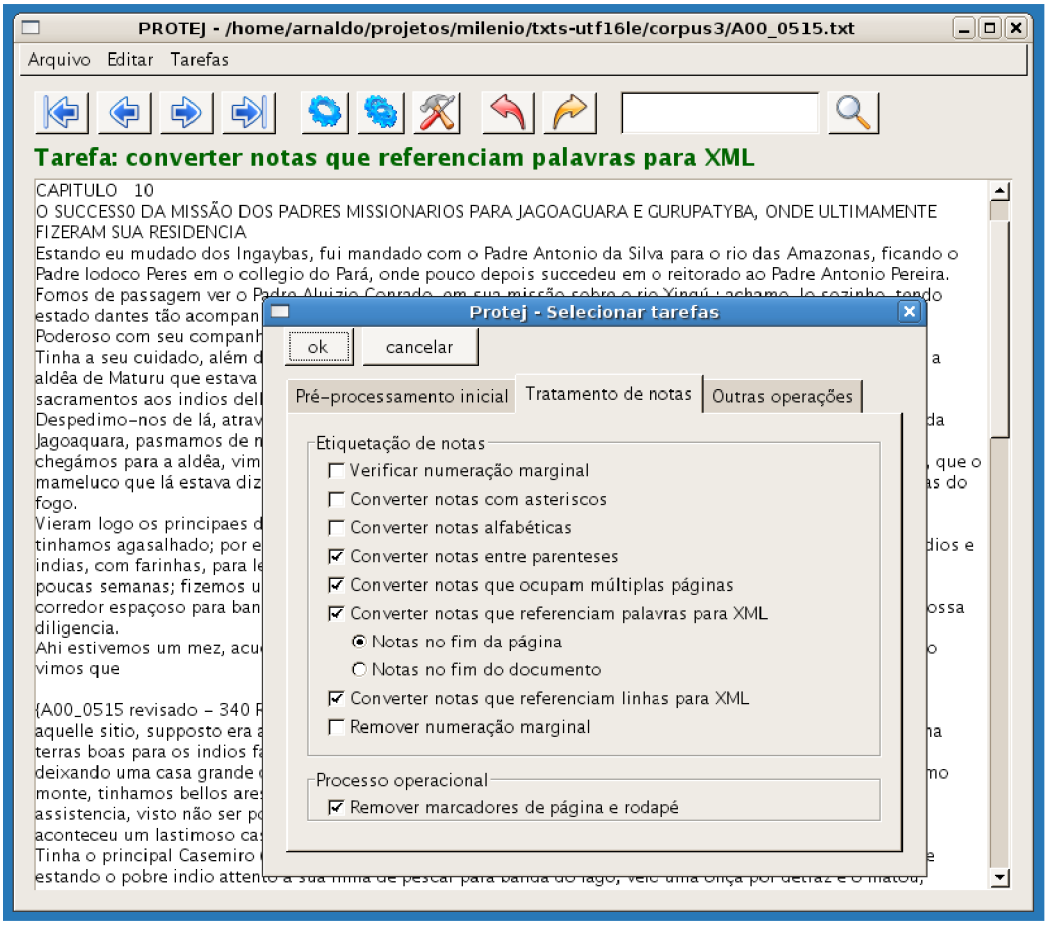

Figura 5.3: A ferramenta Protej

5.2.3 Geração das versões *Philologic* e *Unitex*

A geração do córpus usado no Unitex consiste na remoção das notas rodapé e de qualquer elemento XML presente no texto, pois o Unitex não é capaz de processar XML. A seguir, todos os textos são concatenados e um único arquivo é gerado, pois a ferramenta só é capaz de tratar um arquivo por vez. Esse arquivo passa por outro pré-processamento na ferramenta Unitex, no qual são gerados os índices para as pesquisas e as freqüências das palavras são calculadas.

Na geração do córpus usado no *Philologic*, as notas de rodapé também são removidas. Apesar do *Philologic* reconhecer notas em formato TEI, essas notas são mostradas dentro do concordanceador, o que pode confundir os usuários durante as pesquisas para a criação do dicionário histórico. Em trabalhos futuros, as notas serão corretamente exibidas aos usuários. A seguir, os parágrafos são etiquetados de acordo com as recomendações TEI. Por fim, é inserido o cabeçalho TEI para cada texto. As informações do cabeçalho são extraídas dos arquivos textos gerados pelo Protew-lite.

A etiquetação de parágrafos é necessária para que a ferramenta *Philologic* possa processar adequadamente os textos. A principal heurística utilizada nessa tarefa é a análise de ponto no final de uma sentença seguido imediatamente por uma quebra de linha. Em alguns casos, existem diversas quebras de linha para um parágrafo devido ao formato visual do parágrafo no texto impresso original. Devido a isso, a heurística não é totalmente precisa, pois uma nova quebra de linha pode ser inserida exatamente após o ponto final de uma sentença interna ao parágrafo. Normalmente, parágrafos que ocupam duas páginas são divididos em dois parágrafos. Isso é feito com o objetivo de evitar que a etiqueta de página seja encapsulada dentro de uma etiqueta de parágrafo. O processo é ilustrado na Tabela 5.9.

Tabela 5.9: Exemplo de etiquetação de parágrafos

Antes da etiquetação de parágrafos
(...) 11. Ho governador Thomé de Sousa me pedio hum Padre pera ir com certa gente que V. A. manda a descobrir ouro. Eu lho prometi porque também nos releva descobri-lo pera ho tisouro de Jesu Christo Noso Senhor, e ser cousa de que tanto proveito resultará hà gloria do mesmo Senhor e bem a todo ho Reino e consolação a V. A. E porque ha-í muitas novas delle e parecem certas, parece-me que irão. (..)
Após a etiquetação de parágrafos
(...) 11. Ho governador Thomé de Sousa me pedio hum Padre pera ir com certa gente que V. A. manda a descobrir ouro.   Eu lho prometi porque também nos releva descobri-lo pera ho tisouro de Jesu Christo Noso Senhor, e ser cousa de que tanto proveito resultará hà gloria do mesmo Senhor e bem a todo ho Reino e consolação a V. A. E porque ha-í muitas novas delle e parecem certas, parece-me que irão. (...)

A Figura 5.4 mostra a saída da ferramenta Protej para a criação do córpus *Philologic*.

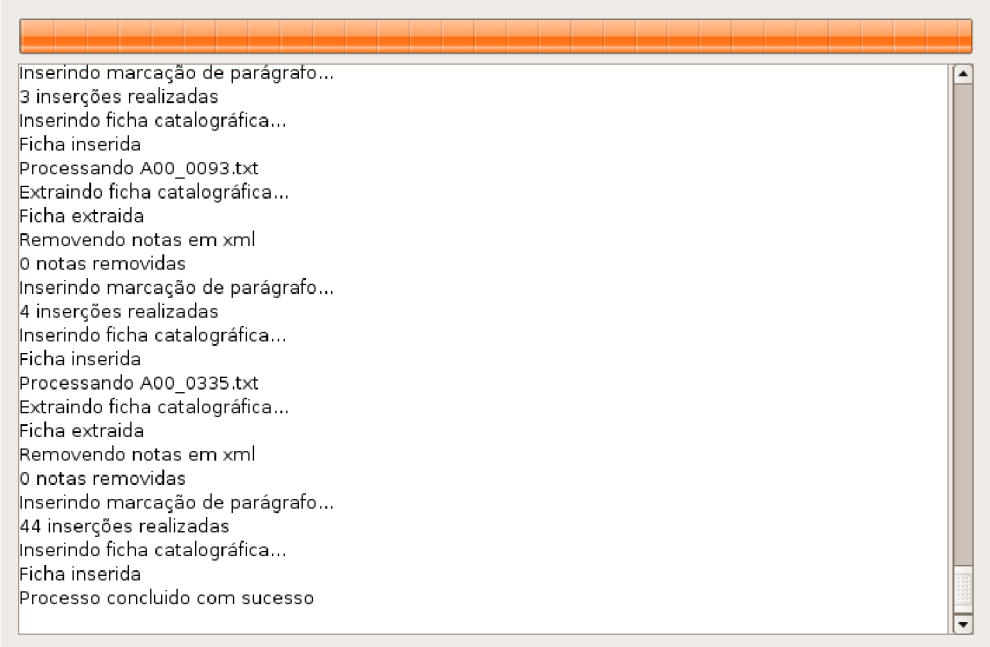

Figura 5.4: Geração do córpus para uso no *Philologic*

5.3 Geração de glossários

Quatro glossários foram desenvolvidos: (a) dois um glossário de abreviaturas (um deles com expansões) (Seção 5.3.1), (b) um glossário de junções de palavras (Seção 5.3.2) e (c) um glossário de variantes de grafia para auxiliar a busca por concordanceas e a contagem de freqüências (Seção 5.3.3). Os glossários de abreviaturas e de variantes seguem o formato

DELA, utilizado pelo *Unitex*.

5.3.1 Abreviaturas

Dois glossários de abreviaturas foram criados no escopo do projeto DHPB. O primeiro com abreviaturas extraídas de (FLEXOR, 1991) e listas de abreviaturas anexas a textos do córpus DHPB. Essa versão será referenciada por glossário de abreviaturas Flexor. A segunda versão contou com heurísticas para a extração de abreviaturas extraídas diretamente do córpus e será referenciada por glossário de abreviaturas do córpus.

Na construção do **glossário de abreviaturas Flexor**, as abreviaturas foram extraídas através da técnica de OCR pela bolsista do projeto Clarissa Galvão Bengtson. Adicionalmente, foram incluídas um conjunto de abreviaturas obtidas em anexo a livros de inventários e testamentos do córpus pela bolsista Livia Aluisi Cucatto. A ferramenta Protej foi utilizada para converter as abreviaturas para o formato DELA. Durante a conversão, o século de ocorrência é inserido como atributo semântico e cada abreviatura é classificada genericamente como substantivo masculino no singular (as informações morfossintáticas são corrigidas manualmente, posteriormente). O processo é ilustrado na Figura 5.10. Esse glossário já está sendo utilizado pelos pesquisadores do projeto.

Tabela 5.10: Exemplo de processamento de abreviaturas

Antes do processamento do glossário de abreviaturas
maced. - macedo (18) macenr^a - marcenaria (18) mach. - machado (10)
Após do processamento do glossário de abreviaturas
maced\.,macedo.N+ABREV+sec18:MS macenr^a,marcenaria.N+ABREV+sec18:MS mach\.,machado.N+ABREV+sec19:MS

Em particular, as letras A, B e C do dicionário de Flexor (1991) receberam informações morfossintáticas e semânticas (VALE et. al, 2008). As informações morfossintáticas facilitam buscas gramaticais no *Unitex*, enquanto que as informações semânticas inseridas são relacionadas a um conjunto de Entidades Nomeadas[14] (ENs) utilizado na avaliação conjunta da tarefa de reconhecimento de ENs (HAREM, 2008). Para tal, as ENs receberam a etiqueta semântica "ENT" e palavras comuns antes de ENs como pronomes de tratamento e alguns adjetivos foram etiquetadas com a etiqueta "INIT". A Tabela 5.11 contém exemplos de ENs e

14 Palavras que se referem a entidades (concretas ou abstratas) que possuam um nome próprio.

de palavras que antecedem ENs para cada as letras A, B e C. Os pontos das abreviaturas estão precedidos por barra invertida (“\”) devido ao uso do formato DELA.

Tabela 5.11: Abreviaturas de entidades nomeadas e de palavras que as precedem

a\.l\. prov^al,assembléia legislativa provincial.N+ENT+ABREV+sec19:ms
a il^ma e ex^ma pessoa de v\. ex^a g^e d^s m^s a^s,a ilustríssima e excelentíssima pessoa de vossa excelência guarde deus muitos anos.N+INIT+ABREV+sec18:ms
bert^meo,bartolomeu.N+ENT+ABREV+sec19:ms
bombr^o,bombeiro.N+INIT+ABREV+sec19:ms
c^a da st^a miser^a,casa da santa misericórdia.N+ENT+ABREV+sec19:fs
capã^m de granadr^os,capitão de granadeiros.N+INIT+ABREV+sec18:ms

As etiquetas ENT e INIT permitem a extração de novas ENs no córpus através de um processo iterativo. Por exemplo, a partir de uma busca no concordanceador da abreviatura “r^o” (rio), é possível obter a entidade nomeada “r^o de s. fran^co” (Rio de São Francisco). Da mesma forma, a partir da EN abreviada “Fran.^co” (Francisco), é possível obter a EN “Mosteiro de Sam Fran.^co” (Mosteiro de São Francisco). O processo iterativo pode ser automatizado através de grafos sintáticos e/ou expressões regulares do *Unitex*. Entretanto, ainda é necessária uma filtragem manual para remover palavras detectadas incorretamente como ENs. No córpus DHPB foi utilizado um processo semelhante ao usado na criação do repositório público de ENs chamado Repentino (REPositório para reconhecimento de ENtidades NOmeadas) para aumentar o dicionário de abreviaturas. Esse processo está sendo realizado por um bolsista do projeto.

No **glossário de abreviaturas do córpus** utilizaran-se três heurísticas simples para a extração de abreviaturas do córpus: (a) busca por palavras com marcador de sobrescrito, por exemplo “jan^ro” (janeiro), (b) busca por palavras com ponto interno, por exemplo: “jan.ro” (janeiro) e (c) palavras terminadas por consoantes (exceto “l”, “m”, “n”, “r”, “s” e “z”) e sucedidas ponto final, por exemplo “av.” (avenida). A heurística (b) causou uma série de erros durante os testes, então se optou por modificá-la para que apenas 4 caracteres fossem permitidos após o ponto.

Além das heurísticas acima, é possível formular outras, como por exemplo, (d) a presença de palavras sem vogais, como em “dr” (doutor), ou (e) palavras terminadas em ponto e sucedidas por palavras que se iniciam por letras minúsculas, como “auxar.” (auxiliar) em “auxar. de cozinha”. Espera-se que as heurísticas englobem grande parte das abreviaturas

presentes em um córpus. Entretanto, podem existir exceções não reconhecidas por nenhuma das heurísticas, como “sr. Afonso” (abreviatura com ponto não detectada, pois é sucedida por palavra com letra maiúscula), algo comum para pronomes de tratamento abreviados. Também é possível a existência de palavras detectadas que não são abreviaturas. Por exemplo, se no córpus uma sentença for iniciada por letra minúscula e a heurística (e) estiver em uso, então a última palavra da sentença anterior será tratada como abreviatura. O processo iterativo para levantamento de entidades nomeadas não foi aplicado para a criação do segundo glossário, pois o glossário é focado em abreviaturas simples.

5.3.2 Junções de palavras

Para a criação do glossário de junções, optou-se pela extração manual, já que esta é menos sujeita a erros que a extração automática, além de ser um processo relativamente rápido. Além disso, supõe-se que o número de erros de extração no processo automático seria grande, principalmente para palavras pequenas como “dado”, “cala” e “tudo”. As junções foram levantadas pela bolsista do projeto Vanessa Marquiafavel através da análise manual, em um total de 10.369 junções. A expansão será aplicada ao córpus em trabalhos futuros. As formas contraídas continuarão sendo mantidas e denotadas por etiquetas TEI, como é mostrado na Tabela 5.12. A Tabela 5.13 mostra o número de junções de acordo com o total de palavras por junção.

Tabela 5.12: Junções anotadas em TEI

<choice> <sic> asmesmas </sic> <corr> as mesmas </corr> </choice>
<choice> <sic> doestillo </sic> <corr> do estillo </corr> </choice>
<choice> <sic> serraniasque </sic> <corr> serranias que </corr> </choice>
<choice> <sic> sobpena </sic> <corr> sob pena </corr> </choice>

Tabela 5.13: Junções VS palavras por junção

Palavras por junção	**2**	**3**	**4**	**5 ou mais**
Total de Junções	9.561	737	60	11

5.3.3 Variantes de grafia

O agrupamento de variantes de grafia no córpus foi feito com a ajuda do Siaconf, desenvolvida pelo bolsista do projeto Rafael Giusti durante o segundo semestre de 2006

(GIUST et. al, 2007). A ferramenta foi construída em linguagem de programação *Perl*. A metodologia para detecção de variações de grafia se baseia na aplicação de regras de transformação. O sistema processa um córpus a partir de uma lista inicial de regras e gera três relatórios principais:

- Agrupamentos de grafias relacionadas.
- Estatísticas das regras aplicadas.
- Lista de palavras não processadas.

O agrupamento gerado é diferente das abordagens de normalização (HIROHASHI, 2004), pois não busca encontrar a variante de grafia contemporânea (ou normalizada), embora isso aconteça com certa freqüência. Por exemplo, as variantes “chaõ” e “chaão” são agrupadas sob a grafia “xam”, inexistente no Português contemporâneo. Na maioria das vezes, a grafia normalizada também fará parte do agrupamento, apesar de não ser possível identificá-la automaticamente como a versão normatizada do agrupamento. A Figura 5.5 mostra as variantes de “chão” detectadas automaticamente.

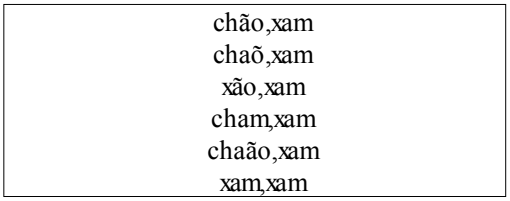

Figura 5.5: Variante de grafia de "chão"

Inicialmente, um conjunto de regras é definido e aplicado ao córpus. A partir da análise dos três relatórios, em especial, do relatório de palavras não processadas, é possível formular novas regras de transformação. Os relatórios também são úteis para verificar erros de detecção. O processo é iterativo, conforme mostrado na Figura 5.6.

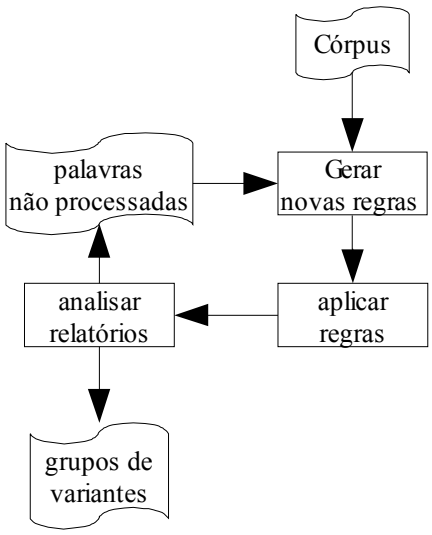

Figura 5.6: Processo de geração
de regras de transformação

Uma regra de transformação é dada na forma (E_1 E_2 S). E_1 e E_2 representam expressões regulares[15] e *S* é uma cadeia de substituição. E_1 determina o critério de cobertura da regra, isto é, as palavras W_i que sofrerão transformação. E_2 determina a subcadeia em W_i que será substituída por *S*. Por exemplo, a regra (e[ao], e, i) é aplicada da seguinte forma:

1. E_1 é verificada para todas as palavras do córpus e seleciona palavras que contenham "eo", ou "ea", como "m**eo**" e "ald**ea**".
2. E_2 determina a sub-cadeia a ser substituída, nesse caso a letra "e" em "m**e**o" e "ald**e**a".
3. *S* determina a subtiuição realizada, nesse caso, "e" por "ei", resultando em "m**ei**o" e "ald**ei**a".

A Tabela 5.14 ilustra possíveis aplicações de regras.

Tabela 5.14: Exemplos de regras de normalização (GIUST et. al, 2007)

Forma orignal	**Regras de transformação**	**Grafias geradas**
pedy	(y, y, i)	pedi
appellido	(pp, pp, p), (ll, ll, l)	apellido, apelido
estaõ	(aõ, aõ, ão), ([aã]o$, [aã]o, am)	estão, estam

O cifrão ("$") da expressão "[aã]o$" indica fim de cadeia. Ou seja, a expressão regular se aplica apenas a cadeias terminadas por "ao" ou "ão". É importante observar que nem

15 Um padrão de símbolos que denota uma ou mais cadeias de símbolos.

sempre a grafia gerada corresponde à forma normatizada da grafia original. Por exemplo, a grafia “estaõ” foi convertida para “estam”. Isso não gera nenhum tipo de problema durante o agrupamento, pois a grafia modernizada “estão” também será convertida para “estam”, de forma que ambas as grafias permaneçam agrupadas. É possível impedir a aplicação automática de regras para algumas cadeias específicas através de um glossário (no caso, o glossário do ReGra), o que é útil para impedir a criação de agrupamentos inconsistentes (como “contribui” e “contribuí”). Foram criadas 43 regras de transformação, agrupadas em seis categorias:

1. Regras para grafias que caíram em desuso.
    1. Substituição de “y” por “i”.
    2. Substituição de “ph” por “f”.
    3. Substituição de “ò” por “ó”.
    4. Substituição de “th” por “t”.
2. Regras para consoantes dobradas.
    1. Substituição de “ff” por “f” (análogo para “cc”, “bb”, “dd”, “gg”, “ll”, “mm”, “nn”, “pp”, “tt”, “uu”, “vv” e “zz”).
3. Regras para normas ortográficas.
    1. Substituição de “m” por “n” antes de consoantes diferentes de “p” ou “b”.
    2. Substituição de “n” por “m” antes das consoantes “p” ou “b”.
    3. Substituição de “aã” por “ã”.
    4. Substituição de “aõ” por “ão”.
    5. Substituição de “à” por “á”, exceto para palavras iniciadas por “à”.
    6. Substituição de “aes” por “ais”.
4. Regras baseadas em freqüência.
    1. Substituição de “chr” por “cr”.
    2. Substituição de “ch” por “x”.
    3. Substituição de “ee” por “é”.
    4. Substituição de “he” por “é”.
    5. Remoção de “p” em “pt”.
    6. Remoção de “c” em “ct”.
    7. Substituição de “v” no final da palavra por “u”.

8. Substituição do primeiro “i” em “issimo”, “issima”, “issimos”, “issimas” por “í”.
9. Substituição de “mn” por “m”.
10. Substituição de “oens” por “ões”.
11. Substituição de “s” por “z” em “oso”.
12. Substituição de “ão” por “am”.

5. Regras léxicas (para palavras específicas).
   1. Substituição de “o” por “u” em “deos” e “judeos”.
6. Regras automáticas (extraídas da metodologia usada no córpus Tycho-Brahe).
   1. Substituição de “a” por “á” em palavras terminando em “ágio”.
   2. Substituição de “z” por “s” em “preciz”.
   3. Substituição de “ss” por “ç” em palavras começadas por “serviss”.
   4. Substituição de “z” por “s” em “zente” (por exemplo, “presente”).
   5. Substituição de “c” por “ss” em “acem” (por exemplo, “tirassem”).

Além do glossário de grafias, as variantes também podem ser procuradas no córpus através do *Philologic*, que utilizada a ferramenta *Agrep*. Essa segunda fonte de pesquisa de variações de grafia também foi disponibilizada aos pesquisadores.

5.4 Acesso a córpus

Além do comparativo apresentado na Seção 3.4 (Comparativo entre as ferramentas), a escolha dos processadores de córpus utilizados também se baseou no atendimento de alguns requisitos de software levando em conta as necessidades do projeto DHPB. A seguir, os processadores selecionados foram então adaptados para atender as necessidades do projeto DHPB.

5.4.1 Levantamento de requisitos

Os processadores de córpus utilizados no DHPB devem atender aos seguintes requisitos de software:

- O sistema deve ter interface *Web*, permitindo acesso simultâneo para vários pesquisadores ao córpus e aos glossários a partir de qualquer computador com acesso a Internet.
- O sistema deve conter um concordanceador com buscas sofisticadas. As

concordâncias são de fundamental importância durante a redação de verbetes.

- O sistema deve permitir buscas orientadas a glossários, em especial, para a busca no glossário de variantes de grafia.
- O sistema deve apresentar um bom desempenho. Esse requisito é importante já que o córpus possui mais de 7 milhões de palavras.
- O sistema deve ser capaz de processar textos anotados em padrões internacionais de anotação como TEI ou XCES.
- O ambiente deve fornecer buscas bibliográficas e permitir a criação de subcórpus. A busca por data de criação dos textos é particularmente importante, pois assim é possível realizar o registro de datas no dicionário. Outros dados bibliográficos pertencentes à ficha catalográfica apresentada na Tabela 1.1 também devem ser considerados.

Parte dos requisitos foi levantada a partir do comparativo entre as ferramentas e parte foi levantada nos encontros dos pesquisadores do projeto DHPB. Observou-se que nenhuma das ferramentas atendeu completamente aos requisitos levantados. Então se optou pelo uso de duas delas: o *Philologic* e o *Unitex*. O *Philologic* atendeu a todos os requisitos, exceto o uso de glossários. O *Unitex* também teve uma boa avaliação, apesar de possuir interface baseada em janelas e de não aceitar córpus anotados em padrões internacionais. Como o *Unitex* não é capaz de trabalhar no modo cliente-servidor, o córpus foi distribuído para os pesquisadores do projeto, juntamente com a ferramenta. Isso causou alguns problemas de sincronização de versões dos usuários, pois a cada nova atualização do córpus, é necessária uma distribuição e instalação nos computadores dos usuários. Como o *Philologic* e o *Unitex* apresentaram vantagens em relação as demais ferramentas e complementaram um ao outro em termos de funcionalidade, optou-se pelo uso das duas ferramentas no projeto.

5.4.2 Adaptação das ferramentas *Philologic* e *Unitex*

O processador de córpus *Unitex* foi escolhido por sua capacidade de buscas orientadas a glossários e por ser útil a usuários sem acesso a Internet ou com conexões de baixa qualidade. O processador de córpus *Philologic* foi escolhido por sua interface *Web* simples e pelo uso do padrão TEI. Com o passar do tempo, o *Philologic* começou a ser mais utilizado, possivelmente devido a sua interface simples de acesso e ao funcionamento sem necessidade

de instalação.

A primeira mudança no *Unitex* foi em relação aos caracteres permitidos. Para o idioma Português, são permitidos apenas o alfabeto romano, as versões acentuadas das vogais e o caractere "ç". Números e sinais de pontuação são ignorados, pois não são utilizados em lexias. Entretanto, em documentos históricos é comum a presença de consoantes acentuadas e outros símbolos como o S-longo (em "diſcurſo"). Foi necessário incluir os símbolos mostrados na Tabela 4.1 ao alfabeto de trabalho do *Unitex*. Os acentos listados na tabela são chamados de acentos combinados e são capazes de se aglutinar a quaisquer símbolos, incluindo vogais, consoantes e até mesmo outros acentos. Adicionalmente, foi incluído o acento circunflexo tradicional ao alfabeto *Unitex*, para que as abreviaturas com sobrescrito pudessem ser corretamente processadas. Entretanto, foi constatado que abreviaturas com ponto interno estavam sendo divididas em duas palavras. Por exemplo, a abreviatura "jan.^ro" é dividida em "jan" e "^ro". Caso o ponto fosse inserido no alfabeto do *Unitex*, as sentenças não seriam corretamente segmentadas. Uma solução para o problema é a criação do glossário de abreviaturas. Para isso, o glossário deve conter todas as abreviaturas com ponto interno do córpus.

Além do glossário de abreviaturas, também foi incluído no *Unitex* o glossário de variações de grafia gerado automaticamente a partir da ferramenta Siaconf. Adicionalmente, o projeto DHPB usa o glossário de Português Contemporâneo do Brasil do *Unitex*, pois este é útil para buscas por flexões de verbos contemporâneos (MUNIZ, 2004). O subcórpus compilado, o alfabeto de trabalho com os novos símbolos e os três glossários (contemporâneo, de abreviaturas e de variações de grafias) foram agrupados em um idioma de trabalho do *Unitex* chamado de "Português Histórico". A versão foi alterada para incluir apenas as três variantes do Português (do Brasil, de Portugal e Histórico). Dessa forma, o processador teve seu tamanho reduzido significativamente, facilitando sua obtenção via Internet.

Juntamente com o *Unitex*, foi distribuído para os usuários o programa Dicionário, um programa desenvolvido por Marcelo Caetano Martins Muniz para permitir buscas nos glossários. Enquanto o *Unitex* permite buscas orientadas a glossários, o Dicionário permite buscas no glossário. Uma desvantagem do Dicionário é o fato de que as buscas são unidirecionais. Por exemplo, na criação do glossário é possível escolher se uma busca no glossário de abreviaturas será feita para abreviaturas e retornará expansões ou se será feita para expansões e retornará abreviaturas, mas ambas as buscas não serão permitidas. Por fim,

foi criado um programa instalador para o *Unitex*. O instalador foi criado a partir de pedidos dos usuários, pois a instalação manual do *Unitex* é relativamente difícil para iniciantes. O pacote de software criado com todas as mudanças do *Unitex* foi chamado de *Unitex-milênio*.

No caso do *Philologic*, poucas mudanças foram necessárias e optou-se apenas pela tradução parcial da interface do processador de córpus para o Português. A maior parte dos recursos do processador já estava funcionando após a instalação. Entretanto, para ativar os recursos de detecção de variantes de grafia, exibição de concordâncias em uma única linha (*kwic*) e uso de banco de dados para indexação de informações catalográficas foi necessário instalar algumas bibliotecas adicionais. O cabeçalho TEI reconhecido pelo *Philologic* difere do cabeçalho usado no projeto DHPB. Isso ocorre, pois o padrão TEI permite a especificação de metadados em diferentes seções do cabeçalho. Por exemplo, o nome do autor pode ser especificado dentro das seções "<sourcedesc>" e "<filedesc>". O *Philologic* foi então adaptado para aceitar o cabeçalho usado no córpus. Por fim, a interface gráfica foi parcialmente traduzida para o Português para facilitar o acesso ao processador. Uma apresentação sobre as ferramentas foi feita durante III Encontro do Projeto DHPB em janeiro de 2007.

5.5 Redação de verbetes

A redação de verbetes é feita atualmente (até fevereiro de 2008) em com o auxílio do *MS Word*. Quando um verbete é finalizado, este é submetido para o processo de revisão e então é armazenado em uma base centralizada de verbetes. Para evitar problemas de sincronização, uma vez que o verbete é centralizado, este não é distribuído aos pesquisadores. Dessa forma, um redator não tem acesso aos verbetes redigidos pelos demais redatores. Outro problema se relaciona a variações na forma e no conteúdo dos verbetes, dificultando sua padronização. O assunto da padronização foi sempre debatido nas reuniões do projeto.

Para solucionar os problemas do processo de redação em *MS Word*, foi desenvolvido um editor de verbetes chamado Procorph (Processador de Córpus de Português Histórico). O nome do editor relaciona-se ao processamento de córpus ao invés da redação de verbetes, pois pretende-se adicionar recursos de processadores de córpus em versões futuras, tornando-o um sistema mais abrangente. Exemplos de recursos que poderão ser adicionados são um gerenciador de glossários e um concordanceador (possivelmente, o concordanceador do *Philologic*). A versão atual do editor conta com um módulo para a redação de verbetes e um

módulo para buscas bibliográficas. Para um verbete, o usuário pode cadastrar informações morfossintáticas, variantes de grafia, acepções (ou definições) acompanhadas de abonações, notas, e sugestões para verbetes relacionados. As buscas bibliográficas são úteis para referenciar as abonações dos verbetes. Entre as vantagens do editor de verbetes em relação ao *MS Word*, é possível citar:

- A disponibilidade *Web*, que permite acesso de forma simplificada aos participantes do projeto. Como os dados ficam sempre centralizados no servidor, essa estratégia evita problemas de sincronização entre cópias diferentes do mesmo verbete.
- A padronização da forma dos verbetes, incluindo a estrutura do texto e a formatação visual. Em particular, as referências bibliográficas podem ser construídas automaticamente para uma dada abonação a partir do código de seu texto e de sua página de ocorrência.
- A opção de geração de versões. Por exemplo, é possível gerar uma versão completa e uma versão resumida do dicionário DHPB. Além disso, também é possível gerar uma versão eletrônica para permitir consultas via Internet.

Quatro níveis de acesso ao sistema são permitidos:

- Consulente: acessa e consulta a base.
- Redator: inclui verbetes na base, e altera os próprios verbetes. Um redator não tem não tem permissão para alterar o verbete de outro.
- Revisor: acesso completo a base de verbetes para a tarefa de revisão de verbetes.
- Administrador: acesso completo a base e ao cadastro de usuários.

O editor foi desenvolvido na linguagem PHP (*Hipertext Preprocessor*), muito utilizada no desenvolvimento de aplicações *Web*. As tecnologias *Web* usadas foram os padrões XHTML (*eXtensible HTML*), *JavaScript* e CSS (*Cascading Style Sheets*). Essas tecnologias têm sido amplamente utilizadas para a construção de Aplicações de Internet Ricas (*Rich Internet Applications* – RIA). A troca de dados entre o servidor *Web* e o cliente (navegador) foi feita principalmente através de transferência síncrona via *JavaScript*. A técnica de transferência assíncrona (conhecida como AJAX - *Asynchronous Javascript And XML*) não foi utilizada, pois aumentaria desnecessariamente o tempo de desenvolvimento do editor. A Figura 5.7

mostra a tela de listagem de verbetes e a Figura 5.8 mostra um trecho da edição do verbete "baía". O verbete está sendo mostrado apenas parcialmente devido a questões de espaço.

Figura 5.7: Procorph - tela de listagem de verbetes

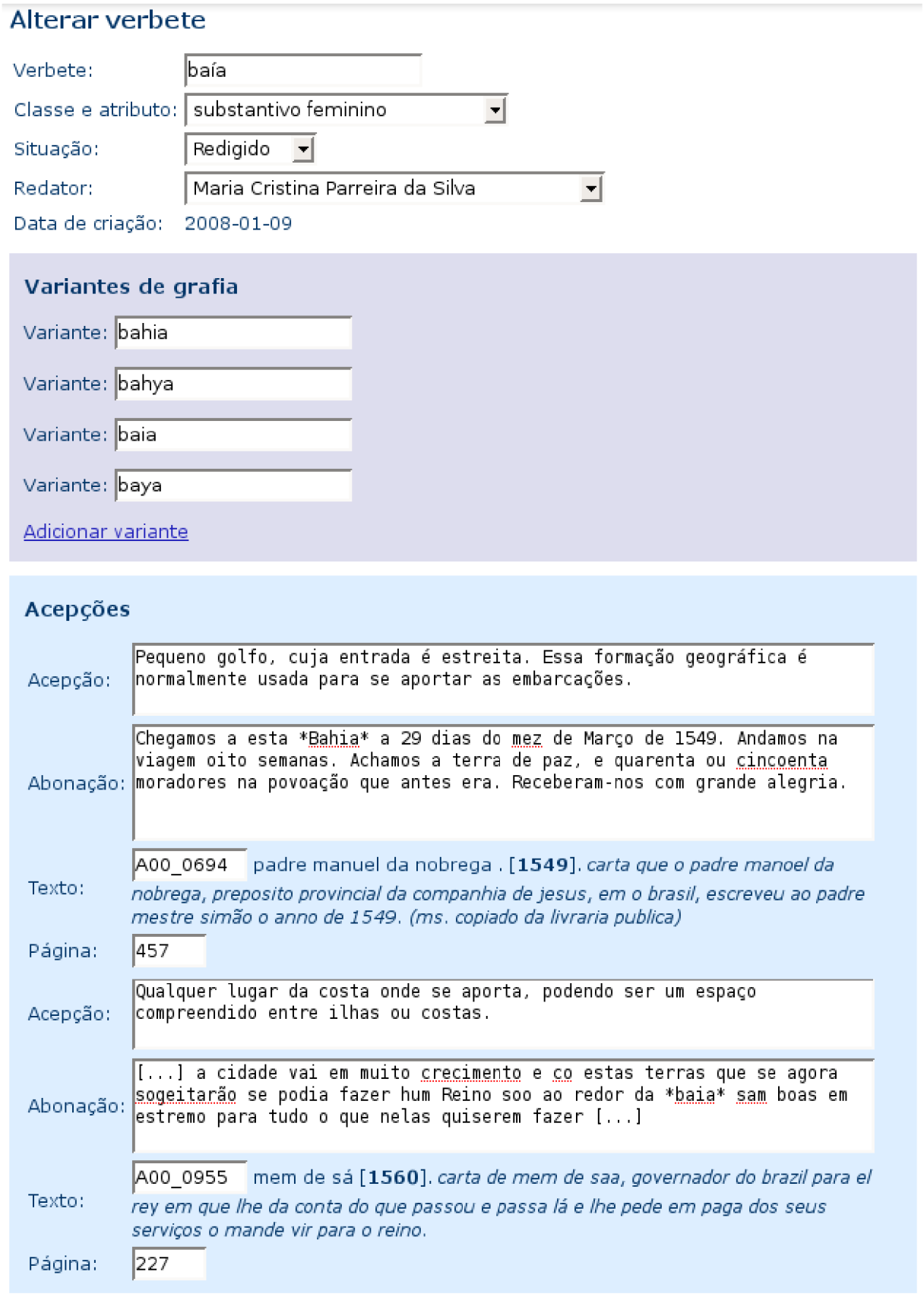

Figura 5.8: Procorph - tela de redação de verbetes (parcial)

6 Avaliação da metodologia proposta

6.1 Considerações iniciais

Este capítulo avalia a metodologia de trabalho descrita no capítulo 5 e apresenta os resultados obtidos. A Seção 6.2 contém estatísticas do córpus gerado com o uso das ferramentas Protew-lite e Protej. A Seção 6.3 apresenta estatísticas dos glossários de abreviaturas e variantes de grafia. A Seção 6.4 descreve a ISO 9126, usada para comparar e avaliar os processadores de córpus. Por fim, a Seção 6.5 discute a situação atual do editor verbetes e suas perspectivas futuras.

6.2 Pré-processamento do córpus

As ferramentas Protew-lite e Protej puderam ser avaliadas durante a construção do córpus DHPB. Para avaliar a robustez das ferramentas, é necessário validá-las em um número relativamente grande de textos que somados possuam um grande volume grande de palavras. A Tabela 6.1 contém informações sobre o tamanho do córpus.

Tabela 6.1: Estatísticas do córpus DHPB

Dados	**Valores**
Tokens	16.505.808
Types	368.850
Formas simples	7.492.473
Formas simples únicas	368.529
Sentenças	287.570
Textos	2.458
Tamanho em MegaBytes (UTF-16)	82,2

Nas formas simples (constituídas por letras do alfabeto de Português Histórico, criado neste trabalho) são contabilizadas apenas as palavras do córpus. Números, sinais de pontuação, espaços, formas simples e outros são contabilizados como *tokens*. Nas formas simples únicas o cálculo é feito sobre palavras simples de maneira análoga, mas a partir da análise de *types*. Esses dados foram calculados pelo *Unitex*. Como o *Unitex* não divide abreviaturas com ponto em duas formas simples, os dados são aproximados. O número de sentenças é uma estimativa feita com base no número de pontos no texto sucedidos por

palavras iniciando por letra maiúscula ou em fim de parágrafo. A Figura 6.1 mostra o gráfico percentual da distribuição do córpus por século.

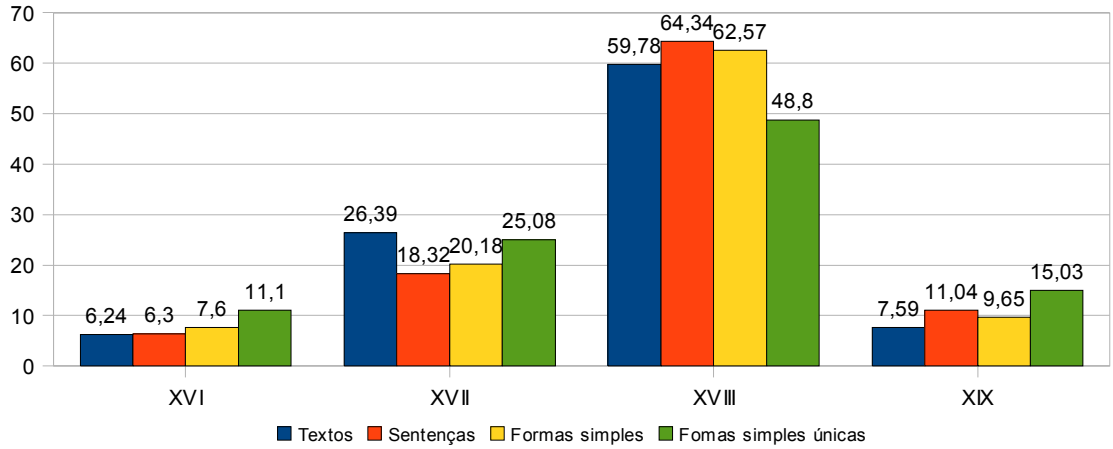

Figura 6.1: Distribuição do córpus por séculos

No córpus existem 86 textos sem século conhecido, esses textos não foram levados em conta nas estatísticas da Figura 6.1. O número de textos do século XVI é pequeno, pois naquela época ainda havia poucos brasileiros alfabetizados. Além disso, por serem mais antigos, é mais fácil que os documentos originais estejam perdidos ou danificados pela ação do tempo. O problema se repete, porém com menos intensidade, para o século XVII. O número de textos do século XVIII é o maior, superando até mesmo o número de textos do século XIX escolhidos para fazerem do córpus, pois o córpus contém textos apenas até 1808.

O fato de poucas mudanças terem sido necessárias nas ferramentas durante todo o processo de compilação do córpus sugere que as ferramentas são robustas, e capazes pré-processar diversos textos históricos com poucas adaptações. O córpus poderia ter sido gerado da forma manual, entretanto, devido ao fato de a tarefa ser manual, o tempo para a construção seria maior e demandaria um maior número de pessoas para o seu pré-processamento. Também é importante citar que pelo fato do procedimento manual ser uma tarefa repetitiva, os erros tornam-se mais freqüentes do que no procedimento automático.

6.3 Geração de glossários

Durante a construção dos glossários foi possível constatar que os fenômenos de abreviaturas, junções e variantes de grafia podem acontecer juntos, como é mostrado na Tabela 6.2. Casos mistos não foram considerados, pois seu tratamento é mais difícil e são

mais raros que as demais instâncias dos três fenômenos.

Tabela 6.2: Fenômenos combinados

Exemplo	Abreviatura	Junção	Variante
Sarg.^José (Sargento José)	X	X	
abaſe (a base)		X	X
supp^te (supostamente)	X		X
héalagadacomm^tos (é alagada com muitos)	X	X	X

6.3.1 Abreviaturas

Para a geração de estatísticas dos glossários, os pontos foram removidos para evitar que abreviaturas como “dr” e “dr.” fossem contadas duas vezes. Dados do glossário de abreviaturas Flexor são mostrados na Tabela 6.3. Observa-se que cerca de 18% das abreviaturas do glossário também ocorrem no córpus. Como uma abreviatura pode ter mais de uma expansão e uma expansão pode possuir diferentes abreviaturas, torna-se importante verificar o número de abreviaturas e de expansões. O número de abreviaturas foi maior que de formas expandidas, o que indica que é mais comum o caso em que uma mesma palavra é abreviada de inúmeras formas.

Tabela 6.3: Estatísticas do glossário de abreviaturas Flexor

Abreviaturas / expansões	Total
Abreviaturas simples e compostas	21.869
Expansões das abreviaturas simples e compostas	8.721
Abreviaturas simples	16.067
Expansões das abreviaturas simples	5.635
Abreviaturas Simples que ocorreram no córpus	3.040
Abreviaturas simples que ocorreram no córpus (%)	18,92%

O glossário de abreviaturas do córpus foi construído com as heurísticas apresentadas na Seção 5.3.1. Exemplos de abreviaturas para cada uma das heurísticas são mostrados na Tabela 6.4. A Tabela 6.5 contém o número de abreviaturas detectados por cada heurística e o total de abreviaturas do glossário.

Tabela 6.4: Exemplos de abreviaturas levantadas

Heurística	**Exemplos**
Presença de marcador de sobrescrito	ant.^o, cid.^e, p.^a, s.^to, mag.^e
Ponto interno seguido de até 4 símbolos	cid.e, embg.e, ex.mo, principalm.e, test.as
Consoante (exceto “l”, “m”, “n”, “r”, “s” e “z”) seguida de ponto	cap., reg., liv., v., vmc.

Tabela 6.5: Número de abreviaturas por heurística

Heurística	**Número de abreviaturas**
Presença de marcador de sobrescrito	4.290
Ponto interno seguido de até 4 símbolos	1.675
Consoante (exceto “l”, “m”, “n”, “r”, “s” e “z”) seguida de ponto	1.083
Total	7.045

Os glossários apresentaram 2.473 abreviaturas em comum, conforme mostrado no diagrama da Figura 6.2.

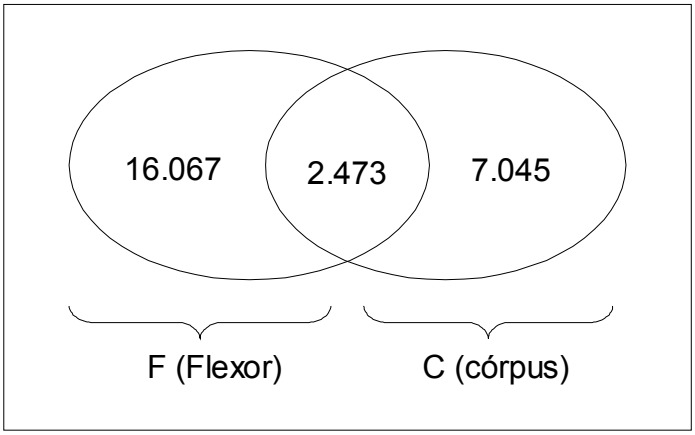

Figura 6.2: Comparativo entre os glossários de abreviaturas

As equações (1) e (2) mostram o percentual de abreviaturas de C em F e de F em C, respectivamente. O percentual de abreviaturas do glossário F que também estão em C é de 15,39% e o percentual de abreviaturas de C em F é de 35,10%. Como F contém muitas abreviaturas que não estão C, é possível concluir que o glossário de Flexor (1991) é bem abrangente. Entretanto, como C também contém muitas abreviaturas que não estão em F, outra conclusão é que o glossário de Flexor poderia ser melhorado com o uso de heurísticas como as apresentadas aqui.

$$\frac{|F \cap C|}{|F|} = 0,1539 \quad (1)$$

$$\frac{|F \cap C|}{|C|} = 0,351 \quad (2)$$

A Tabela 6.6 mostra a distribuição das abreviaturas do glossário F por século. São consideradas as abreviaturas únicas (*types*). Observa-se que somando-se as abreviaturas dos séculos, o total excede 100%, já que uma única abreviatura pode ocorrer em mais de um século. Como o glossário de Flexor também possui abreviaturas compostas, foi levantado o número médio de palavras por abreviatura, ilustrado na Tabela 6.7. Abreviaturas com ponto foram divididas em duas, o que gerou uma margem de erro no número de abreviaturas com uma e duas palavras. Estima-se que o número real de abreviaturas com duas palavras seja menor.

Tabela 6.6: Distribuição das abreviaturas por século

Abreviaturas	**XVI**	**XVII**	**XVIII**	**XIX**
Simples (%)	10,96	21,39	64,14	45,29
Simples que ocorrêm no córpus (%)	22,46	38,88	69,20	49,06

Tabela 6.7: Número de elementos por abreviaturas

Tamanho das abreviaturas	**1**	**2**	**3**	**4**	**5**	**6 ou mais elementos**
Total de Abreviaturas (%)	81,73	7,42	3,81	2,41	1,38	3,25

6.3.2 Variantes

Através das regras de transformação, foram encontradas 18.082 palavras com variações de grafias ou agrupamentos, num total de 41.170 variações através da regras de transformação. A Tabela 6.8 mostra exemplos de variantes detectadas para as palavras “apelido”, “mais”, “não” e “vila”.

Tabela 6.8: Variantes detectadas para as palavras “apelido”, “mais”, “não” e “vila”

apelido (90) appellido (48) apelido (30) appelido (7) apellido (5)	nam (37,100) não (33,684) naõ (2,652) nam (439) nao (325)"
mais (23053) mais (22,918) majs (67) maes (38) mays (30)	vila (5,218) villa (4,073) vila (1,113) vyla (13) vjlla (9) vylla (9) vjla (1)

As técnicas de regras de transformação (Siaconf) e de distância de edição (*Philologic*) foram avaliadas em conjunto através das medidas precisão e cobertura comparativa. A cobertura comparativa é uma medida usada em Recuperação de Informações quando a cobertura não é conhecida. É difícil calcular a cobertura no córpus, pois seria necessário conhecer *a priori* todas as possíveis variantes de grafia para cada palavra analisada. A precisão comparativa pode ser calculada da seguinte forma: (a) uma palavra é escolhida do córpus, (b) as variantes da palavra são levantadas através dos sistemas Siaconf e Philologic, (c) os erros de detecção são desconsiderados e são obtidos dois conjuntos *P,* (verdadeiros positivos do *Philologic*) e *S* (verdadeiros positivos do Siaconf), e (d) calcula-se as coberturas comparativas das ferramentas Philologic (C_P) e Siaconf (C_S) através das equações (3) e (4):

$$C_P = \frac{|P|}{|P \cup S|} \quad (3)$$

$$C_S = \frac{|S|}{|P \cup S|} \quad (4)$$

Um experimento realizado no córpus consistiu na escolha aleatória de 23 palavras (agravou, benditas, continuam, determinavam, enterro, fruta, galante, herdar, inquisidores, javali, kisleu, legião, mineravam, novela, oprimido, piloto, queimar, reinos, servir, tenente, usei, vieram, zelar) no relatório de variantes gerado pela ferramenta Siaconf. Cada palavra pertenceu a uma letra do alfabeto Português com a exceção da letra X e a inclusão da letra K. Novas variantes foram geradas através da busca por variantes no *Philologic*. A Tabela 6.9 mostra as médias das precisões e das coberturas comparativas para as 23 palavras. É possível observar uma alta precisão da ferramenta Siaconf e uma alta cobertura comparativa do

Philologic. Novas regras serão adicionadas ao Siaconf para aumentar a sua cobertura comparativa, sem grandes perdas em sua precisão.

Tabela 6.9: Precisão e cobertura comparativa para o experimento

Técnica	**Verdadeiros positivos**	**Falsos positivos**	**Precisão**	**Cobertura comparativa**
Regras de transformação (Siaconf)	36	0	100%	72%
Distância de edição (Philologic/Agrep)	41	196	21%	84%

6.4 Acesso a córpus

Esta seção contém detalhes adicionais sobre a metodologia utilizada no comparativo entre as ferramentas (*Gate*, *Philologic*, *Tenka*, *Unitex* e *Xaira*) mostrado na Seção 3.4 e sobre a ISO 9126. O grupo EAGLES estendeu a ISO 9126 com o objetivo de guiar avaliações em diversos softwares aplicados à engenharia da linguagem (UNIVERSITÉ DE GENÈVE, 2006). Atualmente, as extensões detalham o processo de avaliação para ferramentas de amparo à escrita (por exemplo, corretores ortográficos e revisores gramaticais) e ferramentas para amparo à tradução. Futuras versões das extensões incluirão detalhes para softwares de acesso a córpus e diversas outras categorias (gerenciamento de informação, tradução de máquina, geração de textos, entre outros). Optou-se pelo uso da ISO 9126 ao invés das extensões do grupo EAGLES, pois estas últimas estão mais focadas no amparo à escrita e tradução. Há outras ISOs também focadas em qualidade de software (9241, 12119, 14598), mas a ISO 9126 é a mais difundida e possui a versão brasileira NBR 13596.

A ISO 9126 permite que os desenvolvedores elaborem suas próprias métricas para avaliação de qualidade de software. A avaliação de cada métrica pode ser efetuada com base em critérios subjetivos ou objetivos. Fica a cargo do desenvolvedor decidir quais critérios serão utilizados. Além disso, a avaliação pode ser feita não apenas pelo desenvolvedor, mas também pelo gerente de software e pelo próprio usuário.

A principal métrica de avaliação das ferramentas apresentadas foi a funcionalidade. Além da funcionalidade, a avaliação de eficiência também foi considerada, uma vez que o córpus DHPB possui mais de 7 milhões de palavras. Com relação a usabilidade, foi constatado que ferramentas para ambiente *Web* geralmente são consideradas mais amigáveis, pois boa parte dos usuários está familiarizada com navegação *Web*. Ferramentas *Web* também levam vantagem no quesito portabilidade, graças ao fato da *Web* ser ubíqua nas mais diversas

plataformas de software e de *hardware*. Apesar de cada métrica ser dividida em submétricas, estas não foram utilizadas na análise. Além disso, a confiabilidade não foi avaliada devido a dificuldade em realizar testes para medi-la. Entretanto, as ferramentas se mostraram estáveis nos demais testes realizados.

6.5 Redação de verbetes

O editor de verbetes ainda não possui todas as funcionalidades propostas. Atualmente, estão disponíveis os módulos para gerenciamento de verbetes e para listagem de textos. Como trabalho futuro, será desenvolvido um módulo para acesso aos glossários, capaz de fazer buscas bidirecionais entre abreviaturas e suas expansões. Além disso, o módulo de redação de verbetes receberá melhorias para permitir o cadastro de sub-entradas e a inserção facilitada de símbolos *Unicode*. Sub-entradas são verbetes completos associados a um verbete principal (por exemplo o verbete "ouvidor geral" associado ao verbete "ouvidor") e geralmente consistem de lexias complexas. A Tabela 6.10 mostra cadeias de caracteres usadas para denotar símbolos *Unicode* indisponíveis em teclados brasileiros e para inserí-los de forma simples nos verbetes.

Tabela 6.10: Conversão de cadeias para *Unicode*

Original	**Convertido**
grati{ae}	gratiæ
{f}eito	ƒeito
c{oe}teris	cœteris
dis{s}cur{s}o	diſcurſo
{F}ixit	ixit
passad{a}	passadɐ
quar\^y	quarŷ
co\~mande	comm̃ande
caca\-o	cacaō
mu\"y	muÿ
s\comente	sȯmente
tinha\,o	tinhaȯ
\oAfonso	Åfonso

Uma apresentação do editor de verbetes foi feita no IV Encontro do Projeto DHPB em dezembro de 2007 para todos os lexicógrafos do projeto que redigirão verbetes. Durante a

apresentação, os usuários mostraram-se interessados na capacidade de formatação automática de verbetes do Procorph, pois o processo de formatação do texto no *MS Word* mostrou-se demorado por ser manual. Em Janeiro de 2008 o acesso ao editor de verbetes foi liberado para testes. Entre os 21 redatores cadastrados, 6 tem testado a ferramenta e 17 verbetes estão na base para exemplos e testes. Outro trabalho futuro consiste na avaliação da ferramenta através da ISO 9126.

7 Um ambiente para o processamento de córpus de Português Histórico para fins lexicográficos

7.1 Considerações iniciais

Neste capítulo é proposto um modelo de ambiente para processamento de córpus históricos. O modelo foi concebido a partir das experiências obtidas durante a participação no projeto DHPB. Um enfoque é dado a atividades lexicográficas, mas espera-se que o modelo possa ser utilizado para atender às necessidades de córpus históricos em Português com usos variados. O ambiente é constituído por módulos que provêem acesso a diferentes ferramentas de processamento de córpus. A vantagem do uso de módulos consiste na facilidade em adicionar novos recursos ao ambiente, substituir módulos com funcionamento inadequado e personalizar módulos para outros projetos de córpus. Os módulos podem ser agrupados em duas arquiteturas: arquitetura para processamento de córpus e criação de glossários (mostrada na Seção 7.2) e arquitetura para acesso a córpus, glossários e redação de verbetes (mostrada na Seção 7.3).

7.2 Arquitetura para compilação de córpus e criação de glossários

A arquitetura para compilação do córpus e criação de glossários é mostrada na Figura 7.1. Os quadrados representam módulos, as folhas de papel representam dados não estruturados (textos), os cilindros representam dados estruturados (bases de dados). As setas representam interação entre módulos ou escrita e leitura de dados.

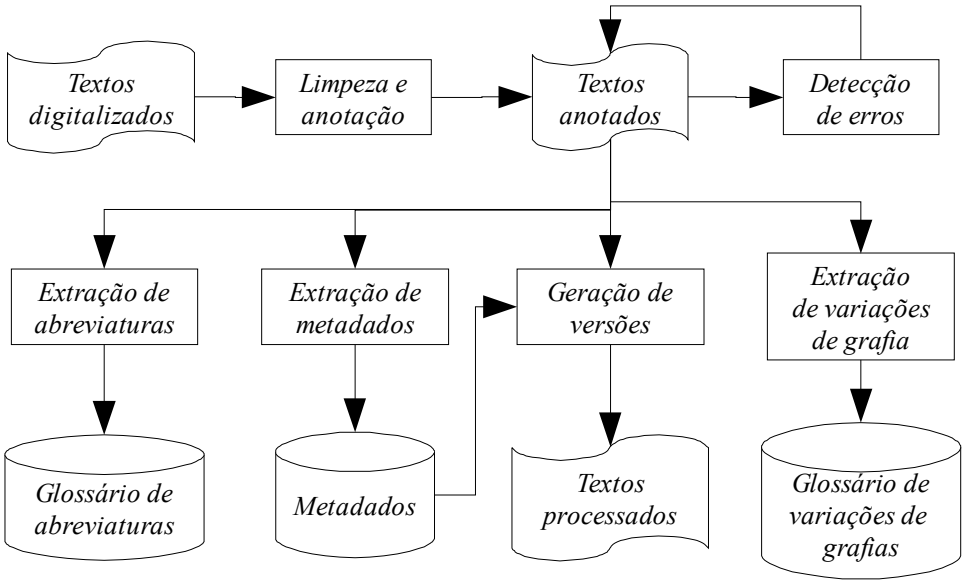

Figura 7.1: A arquitetura de módulos de compilação de córpus e criação de glossários

O **módulo de limpeza e anotação** é responsável pela remoção de metadados indesejados para no texto e anotação de metadados úteis. Para a tarefa lexicográfica, exemplos de metadados indesejados são notas de rodapé (pois podem ser abonadas incorretamente) e numeração de linhas (informações desnecessárias ao consulente). Esse módulo é necessário, pois, geralmente, os textos não estão em um formato adequado para uso em ferramentas de processamento de córpus após sua digitalização (ou transcrição, ou digitação). Por exemplo, pode ser necessária a remoção (limpeza) de informações estruturais como numeração de linhas ou de parágrafos, informações de cabeçalho ou rodapé das páginas. A partir da análise de padrões de informações estruturais é possível realizar diversas operações de limpeza automaticamente. Entretanto, algumas estruturas como numeração de páginas, nomes de capítulos e nomes de seções devem ser mantidas com anotação apropriada, pois fornecem informações úteis sobre o texto. A numeração de páginas, por exemplo, é utilizada no projeto DHPB nas referências das abonações. Nesse caso, é possível anotar tais informações automaticamente (de preferência, com padrões internacionalmente aceitos como TEI e XCES). É possível que nem todas as informações estruturais sejam automaticamente tratadas, uma vez que nem sempre se encontram no padrão esperado. Nesse caso, faz-se necessária uma revisão manual para limpeza e anotação de informações que não puderam ser tratadas automaticamente. No projeto DHPB a limpeza e a anotação são feitas pelas ferramentas

Protew-lite e Protej.

Após a limpeza dos textos, é possível ainda encontrar erros de digitalização (ou de digitação) no córpus. O **módulo de detecção de erros** pode encontrar automaticamente os tipos de erros mais freqüentes, a partir da análise de padrões. Por exemplo, uma busca por palavras contendo os números "1" e "0" pode revelar falhas de digitalização, uma vez que a presença desses números em palavras geralmente está associada a falhas no reconhecimento de "I", "L" ou "O". Outro exemplo de tarefa útil para levantamento de erros é a busca por símbolos pouco usuais, como o símbolo de *copyrigth* (©). Também é possível contrastar cada símbolo do córpus com um alfabeto de símbolos previamente definido, em busca de símbolos pouco usuais, inseridos no córpus devido a erros de OCR. A correção de erros pode ser aplicada na versão recém digitalizada dos textos ou na versão já limpa e anotada. Utilizar a versão limpa e anotada é aconselhável, pois toda correção na versão recém digitalizada implica em limpeza e anotação dos textos novamente (o que pode envolver trabalho manual). A ferramenta Siaconf busca por símbolos desconhecidos no córpus para a detecção de erros de OCR.

O **módulo de extração de abreviaturas** pode ser utilizado para a construção do glossário de abreviaturas a partir heurísticas simples. As abreviaturas também podem ser obtidas a partir de pesquisas sobre abreviaturas históricas e contemporâneas, como a realizada em (FLEXOR, 1991). São disponibilizadas mais de 21 mil abreviaturas para Português do Brasil usadas entre os séculos XVI e XIX que podem ser utilizadas na construção do glossário. No córpus DHPB, a ferramenta Protej é responsável pelo pré-processamento de abreviaturas e pela conversão destas para o formato DELA.

A extração de metadados é feita pelo **módulo de extração de metadados** de acordo com as técnicas mostradas na Seção 4.6 (Extração Automática de Metadados). Os metadados extraídos podem então ser incluídos nas diferentes versões do córpus geradas através do módulo de geração de versões.

Como cada processador de córpus permite padrões de anotação diferentes e em níveis estruturais e lingüísticos diferentes, é necessária a conversão dos textos anotados para diferentes formatos. Para isso, é utilizado o **módulo de geração de versões**. A conversão entre padrões pode ser feita através da linguagem de transformação XSLT ou através de programas construídos especificamente para a conversão de formato. Um caso simples de conversão é a remoção de toda a estrutura XML para uso em ferramentas que não permitam

anotação. A conversão de formato aumenta a reusabilidade do córpus, pois este pode ser utilizado para outros tipos de pesquisas e em conjunto outras ferramentas computacionais.

Por fim, a geração do glossário de variações de grafia é feita pelo **módulo de extração de variações de grafia** com base nas técnicas de agrupamento por distância de edição, análise fonética e/ou regras de transformação. No córpus DHPB, essa tarefa é realizada pela ferramenta Siaconf.

7.3 Arquitetura para acesso a córpus, glossários e redação de verbetes

A arquitetura para acesso a córpus, glossários e redação de verbetes é mostrada na Figura 7.2. O diagrama é análogo ao da Figura 7.1, exceto pela adição do círculo, que representa os usuários do ambiente. A arquitetura proposta é baseada em ambiente *Web* e tem como vantagem a centralização de dados, característica de sistemas do tipo cliente-servidor. Com um ambiente integrado, é possível garantir que todos os pesquisadores trabalhem sobre a mesma base de dados. A centralização evita inconsistência na base de dados, pois todos pesquisadores terão acesso a versão mais atualizada do córpus e dos glossários. A centralização também minimiza os custos de equipamentos necessários para o processamento de córpus, pois os usuários podem utilizar estações de trabalho com configurações modestas, uma vez que a maior parte do processamento é feita no servidor. Muitos usuários estão familiarizados com interfaces *Web*, o que permite que o aprendizado do ambiente seja rápido.

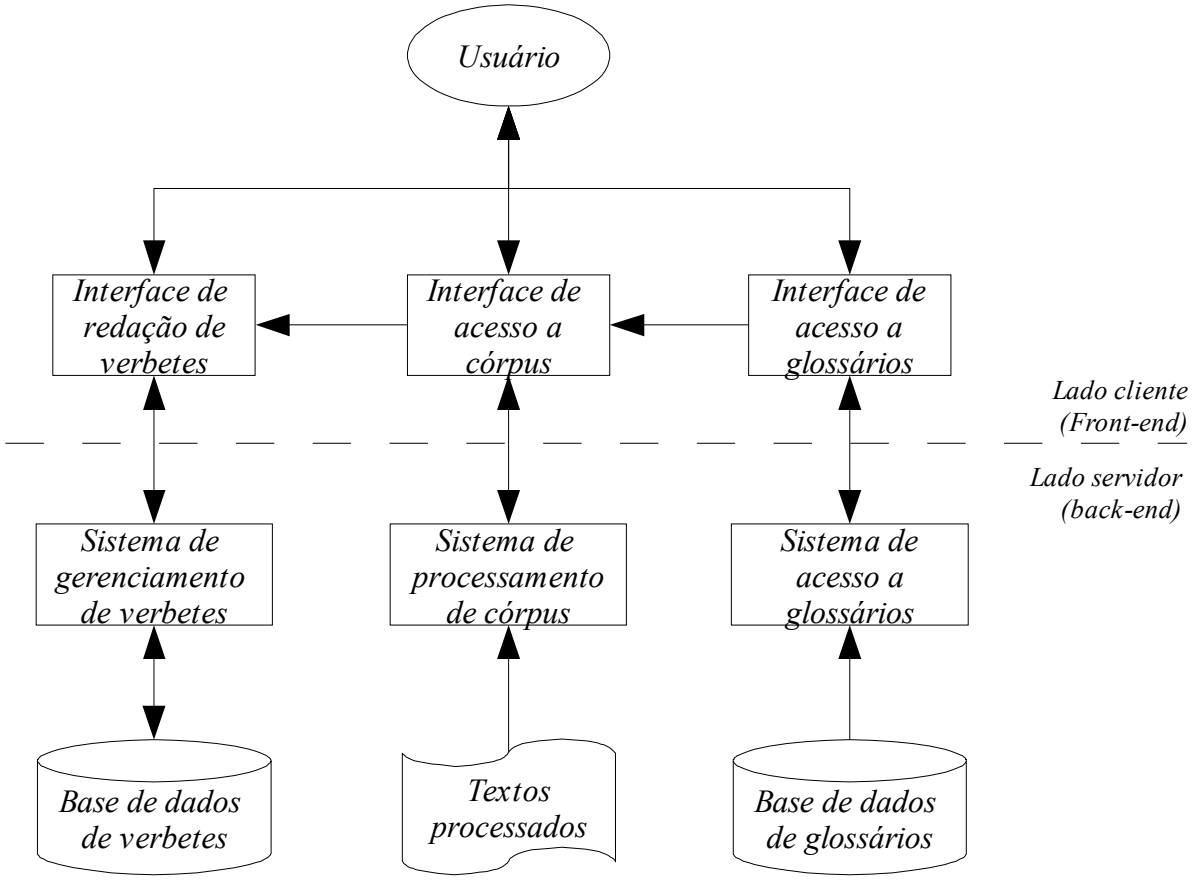

Figura 7.2: A arquitetura de módulos de acesso a córpus, glossários e redação de verbetes

A arquitetura de acesso à córpus fornece basicamente módulos agrupados em três categorias: acesso ao córpus, redação de verbetes e acesso aos glossários. Os **módulos de acesso ao córpus** fornecem algumas das funcionalidades das ferramentas para acesso a córpus, descritas na Seção 3.2 (por exemplo, concordâncias, buscas em dados de cabeçalhos e buscas orientadas a glossários). No projeto DHPB, o *Philologic* e o *Unitex* têm provido acesso ao córpus. Os **módulos de acessos a glossários** permitem buscas nos glossários de abreviaturas e de variações de grafias. Para pesquisas lexicográficas por verbos (como as do projeto DHPB), um glossário contemporâneo pode ser utilizado como filtro, permitindo a identificação de variações de grafia e detecção de palavras que caíram em desuso. Essa tarefa vem sendo realizada com a ajuda do *Unitex*. Os **módulos de redação de verbetes** são os mais específicos da arquitetura, pois se aplicam apenas a pesquisas lexicográficas (ou terminológicas). Esses módulos permitem ao usuário a inserção de verbetes na base de verbetes, além de suas acepções, abonações e referências ao córpus. Essa tarefa vem sendo realizada com a ajuda do editor Procorph. Um modelo de entrada de dicionário específico foi

definido para o projeto DHPB.

A arquitetura é dividida em duas partes: lado cliente e lado servidor. O lado cliente é composto por módulos implementados através de *scripts* (programas executados no navegador Web) responsáveis pela apresentação e formatação de dados (lexias, concordâncias e verbetes). O lado servidor é composto por módulos criados a partir de tecnologias heterogêneas. O uso de tecnologias distintas é mais vantajoso que a criação de um ambiente homogêneo, pois os módulos podem ser construídos a partir de diferentes sistemas já implementados, evitando assim o retrabalho. Além disso, cada tecnologia de construção de software possui vantagens para determinadas tarefas em detrimento de outras.

A estrutura é proposta de forma que os módulos de acesso a córpus sejam capazes de acessar sempre que necessário os módulos de acesso a glossário. Dessa forma, é possível expandir as buscas do usuário. Por exemplo, é possível buscar por todas as variações de grafia e todas as abreviaturas de uma determinada palavra. Outra possibilidade é a busca por todas as formas flexionadas de um verbo (utilizando-se o glossário de Português contemporâneo). Da mesma forma, os módulos de redação de verbetes podem utilizar os serviços providos pelos módulos de acesso a córpus, simplificando os processos de abonação e referência ao córpus.

8 Conclusões

8.1 Contribuições

Este trabalhou foi motivado pelas necessidades de tratamento de córpus históricos levantadas no decorrer do projeto DHPB. Quatro tarefas foram identificadas: (a) a compilação do córpus histórico do Português do Brasil, (b) a construção de glossários de apoio à tarefa lexicográfica, (c) o acesso ao córpus e (d) a redação de verbetes. Uma metodologia para resolver os problemas encontrados em cada tarefa foi proposta e implementada através de um ambiente para processamento de córpus históricas. O ambiente foi implementado através do desenvolvimento de ferramentas (Protew, Protej e Procorph) para amparar a compilação do córpus, a geração de glossários no formato DELA e a redação de verbetes. O ambiente também contou com duas ferramentas amplamente difundidas (*Philologic* e *Unitex*) para acesso a córpus, adaptadas para as necessidades do projeto DHPB. A metodologia proposta foi generalizada em um modelo para que possa ser aplicada em outros projetos que envolvam o processamento de córpus de Português Histórico e/ou a redação de verbetes em dicionários históricos.

As contribuições deste trabalho podem ser agrupadas em três categorias: a metodologia, os recursos (os glossários e o córpus como um todo) e as ferramentas desenvolvidas para processá-los e para redigir os verbetes do dicionário, objetivo principal do projeto DHPB, um projeto de relevância social e histórica para a sociedade. Espera-se que os recursos, as ferramentas e, em especial, a metodologia apresentados aqui possam ser úteis a outros projetos, aumentando o alcance deste trabalho. Dessa forma, optou-se pela disponibilização pública das ferramentas e os recursos. O córpus e o glossário de abreviaturas de Flexor, entretanto, não poderão ser disponibilizados publicamente devido a questões de direitos autorais. Uma contribuição adicional é o comparativo apresentado entre os processadores de córpus, que pode ser útil a pesquisadores da área de lingüística de córpus para amparar a escolha de ferramentas.

Foi observado que a construção do córpus e do dicionário DHPB é uma tarefa de grandes dimensões, que demandam o trabalho e a integração de diversos pesquisadores. É importante observar que as contribuições apresentadas neste projeto só foram possíveis graças a ajuda de inúmeros participantes do projeto DHPB, citados no decorrer do texto.

8.2 Resultados e limitações

Esta seção discute resultados obtidos e algumas limitações encontradas nas diferentes fases deste trabalho. O córpus gerado com o auxílio das ferramentas atingiu 7.5 milhões de palavras, um tamanho considerável por tratar-se de um córpus histórico e pela dificuldade no levantamento e digitalização de textos com essas características, principalmente no Brasil, já que poucos eram os alfabetizados aptos a escrever textos até meados do século XX. Considera-se que as ferramentas foram úteis na compilação desse córpus, apesar de haver a necessidade de utilização de mais etiquetas TEI, principalmente para o processamento de mais campos de cabeçalho. As ferramentas desenvolvidas para o pré-processamento necessitaram de pouca adaptação para o processamento de diferentes gêneros textuais do córpus, o que sugere que estas possam ser utilizadas por projetos com outros gêneros. Entretanto, as ferramentas estão muito focadas nas decisões de projeto do córpus DHPB, o que indica que um número razoável de mudanças pode ser necessário para o seu uso em outros projetos. Por exemplo, o uso particular dos campos da ficha catalográfica e a não expansão de abreviaturas.

A metodologia para a criação de glossários se mostrou eficiente, em especial para os glossários de abreviaturas e de variantes de grafia. Acredita-se que o uso de heurísticas é especialmente útil para a detecção de abreviaturas, mas observa-se que as heurísticas apresentadas aqui precisam ser expandidas para detectar um volume mais completo de abreviaturas do córpus. Além disso, os processos de pós-detecção de abreviaturas levantam duas questões que necessitam de mais estudo: (1) É possível detectar com precisão o conjunto de expansões para uma abreviatura extraída automaticamente do córpus? (2) De posse do conjunto de expansões de uma abreviatura, é possível expandir automaticamente as ocorrências da abreviatura no córpus com uma taxa de erros baixa? As regras de transformação se mostraram a forma mais flexível para a detecção de variantes no córpus. Entretanto, o número de grafias agrupadas ainda é relativamente pequeno, segundo relatos dos participantes do projeto. Esse problema pode ser sanado com a criação de mais regras de transformação. O uso de algoritmos de distância de edição também se mostrou útil para a detecção de variantes de grafia. Adicionalmente, há necessidade de mais estudo para a criação automática do glossário de junções.

As ferramentas *Philologic* e *Unitex* se mostraram robustas e adequadas para permitir aos pesquisadores do projeto o acesso ao córpus. Também foi possível constatar a flexibilidade desses dois softwares durante a fase de adaptação das ferramentas. Observa-se que o

Philologic está mais bem preparado para o trabalho com textos históricos, devido a detecção de variantes de grafia através de distância de edição e do processamento do padrão TEI. Entretanto, é desejável uma única ferramenta capaz de agregar os pontos fortes do *Unitex* e do *Philologic*, simplificando o acesso ao córpus do projeto DHPB.

O editor de verbetes se mostrou útil para centralizar e simplificar o processo de redação de verbetes, e acredita-se que seu uso proporcione um ganho de produtividade na tarefa de redação em relação ao MS Word, pois este é um processador de texto e, devido a isso, não é adequado a tarefa de redação de verbetes do dicionário. Entretanto, algumas mudanças são necessárias, como a inclusão do recurso de sub-entradas.

8.3 Trabalhos futuros

Esta seção apresenta algumas atividades que poderão vir a ser implementadas em trabalhos futuros. Em relação ao córpus, algumas melhorias podem ser feitas. Há erros devido ao processo de OCR, que ocorrem com certa freqüência, e seria interessante detectá-los automaticamente. Também existem alguns erros oriundos da fase de pré-processamento que precisam ser tratados, em especial, a existência de cabeçalhos e rodapés que não foram detectados automaticamente. Outra mudança importante para a padronização do córpus é o uso das etiquetas TEI para denotar sobrescrito e números de página, atualmente denotados pelo circunflexo e por chaves (para o trabalho com o *Unitex*), respectivamente. O processo de extração da ficha catalográfica precisa de mudanças para abranger um número maior de metadados. Novas funcionalidades poderão ser acrescentadas às ferramentas Protew e Protej para aplicar as mudanças no córpus. Essas duas ferramentas poderão ainda ser avaliadas de acordo com as métricas definidas na ISO 9126.

Em relação aos glossários, foi observado que mais heurísticas poderão aumentar o conjunto de abreviaturas detectadas automaticamente e que novas regras de transformação deverão aumentar a abrangência do glossário de variantes de grafia. No glossário de variações de grafia, foi verificado que muitas variantes de grafia são oriundas do século XVI, o que indica a necessidade de um estudo em separado para o uso da grafia nesse século. O glossário de junções, útil para aumentar o desempenho das buscas e a contagem de freqüências, ainda precisa ser aplicado ao córpus, gerando uma versão do córpus com junções separadas. Para o glossário de abreviaturas, é necessária a inserção de informações morfossintáticas para as letras entre D e Z. Atualmente o processo está sendo feito manualmente por um bolsista do

projeto.

O módulo de extração automática de metadados poderá ser feito com base na proposta de Aires (2005). O software Weka pode ser utilizado para a tarefa de aprendizado supervisionado para extração dos metadados gênero e domínio. O módulo de extração automática de metadados deverá extrair os traços lingüísticos levantados por Jacqueline Souza automaticamente e utilizá-los como base para a classificação de domínios e gêneros. A classificação poderá ser feita manualmente para um subconjunto do córpus, utilizado para treinamento do classificador. O classificador criado será capaz de classificar textos restantes, se alcançar uma boa precisão para essa tarefa.

Em relação à redação de verbetes, a inclusão de sub-entradas deverá tornar a ferramenta Procorph mais completa. Além disso, é necessário permitir a inserção de atributos dos verbos (como a transitividade). Essas duas tarefas serão desenvolvidas durante os meses de Março e Abril de 2008.

Também seria interessante criar um módulo para acesso a glossários na ferramenta para agilizar a redação de verbetes. Por fim, um módulo concordanceador e um módulo contator de freqüências poderiam ser desenvolvidos para integrar as funcionalidades do *Philologic* e do *Unitex* e centralizar o acesso ao córpus e a redação de verbetes em um único sistema. O módulo concordanceador teria a vantagem do tratamento de notas do rodapé, já que as notas são removidas do *Unitex* e do *Philologic* para evitar problemas em buscas por concordâncias.

Referências

AHMAD, K. Language engineering and the processing of specialist terminology. In: *The Language Engineering Convention/Journees du Genie Linguistique*. Paris, France: European Network in Language and Speech (ELSNET), 1994.

AIRES, R. V. X. *Uso de marcadores estilísticos para a busca na Web em português*. 183 p. Tese (Doutorado) – ICMC-USP, São Carlos, 2005.

ALMEIDA, G. M. B.; OLIVEIRA, L. H. M.; ALUÍSIO, S. M. A terminologia na era da informática. *Ciência e Cultura (SBPC)*, v. 58, p. p.42–45, 2006.

ALUÍSIO, S. M.; ALMEIDA, G. M. de B. *O que é e como se constrói um córpus?* Lições aprendidas na compilação de vários corpora para pesquisa lingüística. Calidoscópio (UNISINOS). Vol. 4, n. 3 , p. 155-177, set/dez 2006. Disponível em: <http://www.unisinos.br/publicacoes_cientificas/images/stories/pdfs_calidoscopio/vol4n3/art04_aluisio.pdf >. Acesso em: 25 Fev. 2008.

ALUÍSIO, S. M. et al. An account of the challenge of tagging a reference corpus of Brazilian Portuguese. In: *PROPOR 2003*, Lecture Notes on Artificial Intelligence. Faro, Portugal: Springer Verlag, 2003. v. 1.

ALUÍSIO, S. M. et al. The lacio-web project: overview and issues in Brazilian Portuguese corpora creation. In: *Corpus Linguistics 2003* (also as UCREL Technical Report, Vol 16 Part). Lancaster, UK: Lancaster University, 2003. v. 16, p. 14–21.

ALUÍSIO, S. M. et al. The lácio-web: Corpora and tools to advance brazilian portuguese language investigations and computational linguistic tools. In: *LREC 2004*. Lisboa, Portugal: Elra, 2004. p. 1779–1782.

ALVES, C. D. C.; FINGER, M. Etiquetagem do português clássico baseada em corpus. In: *IV Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada (PROPOR 99)*. Évora, Portugal: Universidade de Évora, 1999. p. 17–31.

ARCHER, D., ERNST-GERLACH A., KEMPKEN S., PILZ T., RAYSON P. The identification of spelling variants in English and German historical texts: manual or automatic. In: D*igital Humanities*, 2006, Paris: Sorbonne, 2006. p. 3-5.

ASTON, G.; BURNARD, L. *Introduction to SARA98*. 2001. Disponível em: <http://www.oucs.ox.ac.uk/rts/xaira/Doc/saratut.html>. Acesso em: 22 set. 2006.

ATKINS, S. et al. From resources to applications. designing the multilingual ISLE lexical entry. In: *LREC 2002 - Third Inter national Conference*. Las Palmas, Ilhas Canárias: Elra, 2002. p. 687–693.

ATKINS, S.; CLEAR, J.; OSTLER, N. Corpus design criteria. *Journal of Literary and Linguistic Computing*, v. 7, n. 1, 1992.

BIBER, D. *Variation across speech and writing*. New York: Cambridge University Press, 1988.

BIBER, D. Representativeness in corpus design. *Literary and Linguistic Computing*, v.

8, n. 4, p. 1–15, 1993.

BIBER, D. Using register-diversified corpora for general language studies. *Computational Linguistics*, v. 19, n. 2, p. 219–241, 1993.

BIBER, D. *Sociolinguistic Perspective on Register*. New York: Oxford University Press, 1994.

BIBER, D. *Dimensions of Register Variation*: A cross-linguistic comparison. 1. ed. Cambridge: Cambridge University Press, 1995.

BICK, E. *The parsing system "Palavras"*: Automatic grammatical analysis of portuguese in a constraint grammar framework. Tese (Doutorado) — University of Århus, Aarhus, Denmark, 2000.

BRITTO, H.; FINGER, M. Constructing a parsed corpus of historical portuguese. In: *International Humanities Computing Conference ACH-ALLC'99*. Charlottesville, Virginia: University of Virginia, 1999. p. 234–235.

BURNAGE, G.; DUNLOP, D. Encoding the British national corpus. In: *13th International Conference on English Language research on computerised corpora*. Amsterdam: Rodopi, 1992. Disponível em: <http://www.natcorp.ox.ac.uk/archive/papers/Burnage93a.htm>. Acesso em: 16 set. 2006.

CHRIST, O. A modular and flexible architecture for an integrated corpus query system. In: *COMPLEX*. Amsterdam: [s.n.], 1994. Disponível em: <http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/Papers/christ:complex94.ps.gz>. Acesso em: 16 set. 2006.

CUNNINGHAM, H., D. MAYNARD, K. BONTCHEVA, V. TABLAN, C. URSU, M. DIMITROV, M. DOWMAN, N. ASWANI, I. ROBERTS, Y. LI AND A. SHAFIRIN. *Developing Language Processing Components with GATE Version 4* (a User Guide). Disponível em: <http://gate.ac.uk/sale/tao/index.html>. Acesso em: 29 out. 2007.

FAIRON, C.. 1999. Parsing a Web site as a corpus, in C. Fairon (ed.). Analyse lexicale et syntaxique: Le système INTEX, *Lingvisticae Investigationes Tome XXII*. John Benjamins Publishing, Amsterdam/Philadelphia, p. 327-340.

FIRTH, J. R. The technique of semantics. In: *Papers in Linguistics 1934-1951*. London: Oxford University Press, 1935.

FLEXOR, M. H. O. *Abreviaturas*: Manuscritos dos séculos xvi ao xix. 2. ed. [S.l.]: UNESP, 1991. 468 p.

GALVES, C.; BRITTO, H. A construção do corpus anotado do português histórico tycho brahe. In: *IV Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada (PROPOR 99)*. Évora, Portugal: Universidade de Évora, 1999. p. 81–92.

GARNER, S. WEKA: The waikato environment for knowledge analysis. In: *New Zealand Computer Science Research Students Conference*. Hamilton, New Zealand: University of Waikato, 1995. p. 57–64.

GIOULI, V.; PIPERIDIS, S.. *Corpora and HLT: Current trends in corpus processing and annotation*. Disponível em: <http://www.larflast.bas.bg/balric/eng_files/corpus_deliverable_final.htm>. Acesso em: 25 fev. 2008.

GIUSTI, R.; CANDIDO JR, A.; MUNIZ, M. C. M.; CUCATTO, L. A.; ALUÍSIO, S. M. Automatic detection of spelling variation in historical corpus: An application to build a Brazilian Portuguese spelling variants dictionary. In: *Corpus Linguistics, 2007*, Londres. Corpus Linguistics, 2007.

GRÖNQVIST, L. *TEI or XCES*? Porting the Göteborg Spoken Language Corpus to XML. 2003. Disponível em: <http://www.gslt.hum.gu.se/˜ leifg/gslt/doc/lingres.ps>. Acesso em: 16 set. 2006.

HADDAD, R. *Survey of the Canadian Translation Industry*. Moncton, Canada: Canadian Translation Industry Sectoral Committee, 1999. Technical report.

HAREM. HAREM: Avaliação de Reconhecimento de Entidades Mencionadas. Disponível em: <http://poloxldb.linguateca.pt/harem.php>. Acesso em: 25 fev. 2008.

HIROHASHI, A. S. *Aprendizado de regras de substituição para normatização de textos históricos*. Dissertação (Mestrado) — Instituto de Matemática e Estatística, USP, São Paulo, 2004.

IDE, N. Encoding linguistic corpora. In: *Sixth Workshop on Very Large Corpora*. Montreal: University of Montreal, 1998. p. 9–17.

IDE, N. *Linguistic Annotation Format*. 2006. Disponível em: <http://www.tc37sc4.org/new_doc/ISO_TC_37_SC4_N311_Linguistic_Annotation_Framework.pdf>. Acesso em: 16 set. 2006.

IDE, N.; BONHOMME, P.; ROMARY, L. XCES: An xml-based encoding standard for linguistic corpora. In: *Second International Language Resources and Evaluation Conference*. Paris: European Language Resources Association, 2000. Disponível em: <http://www.cs.vassar.edu/˜ide/papers/xces-lrec00.pdf>. Acesso em: 16 set. 2006.

IDE, N.; BREW, C. Requirements, tools, and architectures for annotated corpora. In: *Data Architectures and Software Support for Large Corpora*. Paris: European Language Resources Association, 2000. p. 1–5.

IDE, N.; ROMARY, L. International standard for a linguistic annotation framework. *Journal of Natural Language Engineering*, v. 10, n. 3, p. 211–225, 2004.

IDE, N.; SUDERMAN, K. The American National Corpus first release. In: *Fourth Language Resources and Evaluation Conference (LREC)*. Lisbon: Elra, 2004. p. 1681–1684.

JESUS, M. C.; NUNES, M. das G. V. Autômatos finitos e representação de grandes léxicos. In: *V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR 2000)*. Atibaia, SP: ICMC/USP, 2000. p. 29–41.

KAUFFMANN, C. H. *O Corpus do Jornal*: Variação lingüística, gêneros e dimensões da imprensa diária escrita. Dissertação (Mestrado) — LAEL - PUC, São Paulo, 2005.

KENNEDY, G. *An Introduction to Corpus Linguistics*. New York: Longman, 1998.

LEWICKI, P.; HILL, *T.. Statistics Methods and Applications*. Disponível em: <http://www.statsoft.com/textbook/stathome.html?stdiscan.html&1>. Acesso em: 25 fev. 2008.

MACLEOD, C.; IDE, N.; GRISHMAN, R. The american national corpus: A standardized resource for american english. In: *Second International Conference on Language Resources and Evaluation (LREC 2000)*. Athens: Elra, 2000.

MARCUSCHI, L. A. Gêneros textuais: definição de funcionalidade. In: DIONÍSIO, A. P.; MACHADO, A. R.; BEZERRA, M. *Gêneros Textuais & Ensino*. Rio de Janeiro: Lucerna, 2002.

MARCUSCHI, L. A.; XAVIER, A. C. *Hipertexto e gêneros digitais*. Rio de Janeiro: Lucerna, 2005.

MCENERY, T.; WILSON, A. *Corpus Linguistics*. Edinburgh: Edinburgh University Press, 1996.

MENEGATTI, T. A. *Regras Lingüísticas para Tratamento Computacional da Variação de Grafia e Abreviaturas do Corpus Tycho Brahe*. Campinas: Universidade de Campinas, 2002. Relatório técnico.

MONARD, M. C. et al. *Uma introdução ao aprendizado simbólico de máquina por exemplos. São Carlos: ICMC, outubro 1997*. Notas didáticas do ICMC. Disponível em: <http://www.icmc.usp.br/ biblio/download/not_did/ND_29.pdf>. Acesso em: 11 fev 2007.

MUNIZ, M. C. M. *A construção de recursos lingüístico-computacionais para o português do Brasil: o projeto Unitex-PB*. Dissertação (Mestrado) – Instituto de Ciências Matemáticas de São Carlos, USP, fev. 2004.

MUNIZ, M.; PAULOVICH, F. V.; MINGHIM, R.; INFANTE, K.; MUNIZ, F.; VIEIRA, R.; ALUÍSIO, S. Taming the tiger topic: an XCES compliant corpus Portal to generate subcorpus based on automatic text topic identification. In: *Corpus Linguistics 2007 conference*, 27-30 July 2007, University of Birmingham. Proceedings of the Corpus Linguistics 2007 conference, 2007.

MUNIZ, M. C. M.; NUNES, M. G. V.; LAPORTE, E. Unitex-pb, a set of flexible language resources for Brazilian Portuguese. In: *Workshop on Technology of Information and Human Language (TIL)*. UNISINOS, São Leopoldo, Brazil, 2005. p. 2059–2068.

NUNES, M.; OLIVEIRA JR., O. N. O processo de desenvolvimento do revisor gramatical ReGra. In: *XXVII SEMISH (XX Congresso Nacional da Sociedade Brasileira de Computação)*. Curitiba: PUC-PR, 2000. v. 1, p. 6. Disponível em: <http://www.nilc.icmc.usp.br/nilc/download/Nunesfinal.zip>. Acesso em: 23 fev. 2007.

PAKHOMOV, S. Semi-supervised maximum entropy-based approach to acronym and abbreviation normalization in medical texts. In: *Medical Texts Proceedings of ACL 2002*. Philadelphia: University of Pennsylvania, 2002. Disponível em: <http://citeseer.ist.psu.edu/542155.html>. Acesso em: 23 fev. 2007.

PAUMIER, S. *Unitex 1.2*: User Manual. June 2006. Disponível em: <http://www-igm.univ-mlv.fr/~unitex/UnitexManual.pdf>. Acesso em: 16 set. 2006.

PINHEIRO, G. M.; ALUÍSIO, S. M. Corpus NILC: *Descrição e análise crítica com vistas ao projeto Lacio-Web*. 2003. Relatório Técnico NILC-TR-03-03. Disponível em: <http://www.nilc.icmc.usp.br/lacioweb/downloads/NILC-TR-03-03.zip>. Acesso em: 23 set. 2006.

RAYSON, P. E. *Matrix*: A statistical method and software tool for linguistic analysis through corpus comparison. Tese (Doutorado) – Lancaster University, september 2002.

RAYSON, P., D. ARCHER AND N. SMITH.VARD versus Word: A comparison of the UCREL variant detector and modern spell checkers on English historic corpora, In *Proceedings of Corpus Linguistics 2005*, vol. 1, no. 1. Birmingham: Birmingham University.

ROCHE, E. Text disambiguation by finite state automata, an algorithm and experiments on corpora. In: *14th conference on Computational linguistics*. Morristown, NJ, USA: Association for Computational Linguistics, 1992. p. 993–997.

RYDBERG-COX, J. A. Automatic disambiguation of Latin abbreviations in early modern texts for humanities digital libraries. In: *Joint Conference on Digital Libraries*. Houston, USA: IEEE Press, 2003. v. 3, p. 372–373.

SANTOS, D.; BICK, E. Providing internet access to portuguese corpora: the AC/DC project. In: *Second International Conference on Language Resources and Evaluation (LREC 2000)*. Athens: Elra, 2000. p. 205–210.

SANTOS, D.; RANCHHOD, E. Ambientes de processamento de corpora em português: comparação entre dois sistemas. In: *PROPOR'99*. [S.l.]: Evora, 2002.

SARDINHA, T. B. *Lingüística de Corpus*. Barueri, SP: Manole, 2004.

SCHULZE, B. M. et al. *Comparative State-of-the-Art Survey and Assessment Study of General Interest Corpus-oriented Tools*. Institut fur maschinelle Sprachverarbeitung, Universitat Stuttgart. 1994. Disponível em: <http://citeseer.ist.psu.edu/rd/0%2C458857%2C1%2C0.25%2CDownload/http://citeseer.ist.psu.edu/cache/papers/cs/22993/http:zSzzSzwww.rxrc.xerox.comzSzpubliszSzmlttzSzmltt-014.pdf/schulze94comparative.pdf>. Acesso em: 25 Fev. 2008.

SCHWARTZ, A. M.; HEARST, M. A simple algorithm for identifying abbreviation definitions in biomedical texts. In: *Pacific Symposium on Biocomputing (PSB)*. Hawaii: Universität Trier, 2003. Disponível em: <http://citeseer.ist.psu.edu/schwartz03simple.html>. Acesso em: 22 fev. 2007.

SEBASTIANI, F.. Machine learning in automated text categorization. *ACC Computing Surveys*, v. 1, n. 1, p. 1-47. 2002.

SENO, E.; RINO, L. Co-referential chaining for coherent summaries through retorical and linguistic modeling. In: *Workshop on Crossing Barriers in Text Summarization Research/RANLP*. Borovets, Bulgaria: [s.n.], 2005. Disponível em: <http://www.nilc.icmc.usp.br/nilc/download/SenoRino-RANLP05.pdf>. Acesso em: 22 fev. 2007.

SILBERZTEIN, M. D. Intex: a corpus processing system. In: *COLING 94 Proceedings*. Kyoto, Japan: [s.n.], 1994.

SILVA, J. F. F. da. *Extração de Unidades Textuais, Agrupamento, Caracterização e Classificação de Documentos*. Tese (Doutorado) – Universidade Nova de Lisboa, 2003. Disponível em: <http://terra.di.fct.unl.pt/~jfs/publicacoes/tese_final.ps>. Acesso em: 09 fev. 2007.

SILVA, M. C. da. A Noção de Gênero em Swales: Revisitando Conceitos. Recorte - *Revista de Linguagem, Cultura e Discurso* (Ano 2, n. 3), 2005. Disponível em: <http://www.unincor.br/recorte/artigos/edicao3/3artigo_marta.htm>. Acesso em: 14 nov. 2006.

SINCLAIR, J. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press, 1991.

SINCLAIR, J. *Preliminary recommendations on Corpus Typology*. EAGLES, 1996. Disponível em: <http://www.ilc.cnr.it/EAGLES96/pub/eagles/corpora/corpustyp.ps.gz>. Acesso em: 16 fev. 2007.

SWALES, J. M. *Genre analysis*: English in academic and research settings. Cambridge: Cambridge University Press, 1990.

TEI CONSORTIUM. *The TEI Guidelines*. Text Encoding Initiative Consortium, 2006. Disponível em: <http://www.tei-c.org/Guidelines2/>. Acesso em: 16 set. 2006.

TERADA, A.; TOKUNAGA, T.; TANAKA, H. Automatic expansion of abbreviations by using context and character information. *Inf. Process. Management*, Pergamon Press, Inc., Tarrytown, NY, USA, v. 40, n. 1, p. 31–45, 2004. ISSN 0306-4573.

UNICODE CONSORTIUM. *The Unicode Standard version 4.0.1*. 2006. Disponível em: <http://www.tei-c.org/Guidelines2/>. Acesso em: 24 abr. 2006.

UNIVERSITÉ DE GENÈVE. *The ISO 9126 Standard*. 2006. Disponível <http://www.issco.unige.ch/ewg95/node1.html>. Acesso em: 14 nov. 2006.

UNIVERSITÄT LEIPZIG. *Terminology Management*. 2007. Disponível em: <http://www.uni-leipzig.de/~xlatio/software/soft-termiman.htm>. Acesso em: 12 fev. 2007.

UNIVERSITY OF CHICAGO. *PhiloLogic User Manual*. 2006. Disponível em: <http://philologic.uchicago.edu/manual>. Acesso em: 22 set. 2006.

VALE, O. A. ; CANDIDO JR, A. ; MUNIZ, M. C. M.; BENGTSON, C. G. ; CUCATTO, L. A.; ALMEIDA, G. M. B.; BIDERMAN, M. T. ; ALUÍSIO, S. M. Building a large dictionary of abbreviations for named entity recognition in Portuguese historical corpora. In: *American Association for Corpus Linguistics (AACL) 2008*. Proceedings of American Association for Corpus Linguistics, 2008.

VASSAR COLLEGE. *XCES Home*. 2006. Disponível em: <http://www.cs.vassar.edu/XCES/>. Acesso em: 17 set. 2006.

W3C CONSORTIUM. *Extensible Markup Language (XML) 1.0*. 2006. Disponível em: <http://www.w3.org/TR/2006/REC-xml-20060816/>. Acesso em: 14 nov. 2006.

WILKS, Y. et al. Machine tractable dictionaries as tools and resources for natural language processing. In: *Colling'88*. Budapest: John von Neumann Society for Computing Sciences, 1988. p. 750–755.

WOLFF, M.; ANDREEV, L.; OLSEN, M. Ate: Artfl text encoding. In: *ACH-ALLC'99 International Humanities Computing Conference*. Charlottesville, Virginia: IATH, 1999.

WYNNE, M. (Ed.). *Developing Linguistic Corpora*: a guide to good practice. Oxford: Oxbow Books, 2005. Disponível em: <http://ahds.ac.uk/linguistic-corpora/>. Acesso em: 23 fev. 2007.

XIAO, R. *Xaira*: an XML aware indexing and retrieval architecture. 2005. Disponível em: <www.lancs.ac.uk/postgrad/xiaoz/papers/xaira_review.pdf>. Acesso em: 17 set. 2006.

YU, H.; HRIPCSAK, G.; FRIEDMAN, C. Mapping abbreviations to full forms in biomedical articles. *J Am Med Inform Assoc*., v. 9, n. 3, p. 262–272, 2002.

Apêndice A – Páginas de projetos e organizações

A Tabela A.1 contém o endereço de organizações e projetos para a construção de recursos e softwares citados neste trabalho.

Tabela A.1: Páginas de organizações e projetos

Projeto	**Acesso em**	**Página**
Abbyy	16 Fev. 2008	http://www.abbyy.com/
AC/DC	16 Fev. 2008	http://www.linguateca.pt/ACDC/
Agrep	23 Fev. 2008	http://www.tgries.de/agrep/
ANC	16 Fev. 2008	http://www.americannationalcorpus.org/
BNC	16 Fev. 2008	http://www.natcorp.ox.ac.uk/
BNC (CDIF)	16 Fev. 2008	http://xml.coverpages.org/bnc-encoding2.html
CNC	16 Fev. 2008	http://ucnk.ff.cuni.cz/english/index.html
Corpógrafo	24 Fev. 2008	www.linguateca.pt/Corpografo/
Córpus do Português	16 Fev. 2008	http://www.corpusdoportugues.org/
DICIWEB	16 Fev. 2008	http://clp.dlc.ua.pt/DICIweb/
EAGLES	16 Fev. 2008	http://www.ilc.cnr.it/EAGLES/home.html
Emacs	16 Fev. 2008	http://www.gnu.org/software/emacs/
Emdross	16 Fev. 2008	http://emdros.org/
EXPLORA	16 Fev. 2008	http://www.dc.ufscar.br/~lucia/PROJECTS/EXPLOSA.htm
FRANTEXT	16 Fev. 2008	http://www.lib.uchicago.edu/efts/ARTFL/databases/TLF/
Fundação Andrew W Mellon	16 Fev. 2008	http://www.mellon.org/
GATE	16 Fev. 2008	http://www.ontotext.com/gate/index.html
GistSumm	16 Fev. 2008	http://www.icmc.usp.br/~taspardo/GistSumm.htm
HTDig	16 Fev. 2008	http://www.htdig.org/
HTTrack	16 Fev. 2008	http://www.httrack.com/
ICE	16 Fev. 2008	http://www.ucl.ac.uk/english-usage/ice/
IMS	16 Fev. 2008	http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/
Intex	16 Fev. 2008	http://intex.univ-fcomte.fr/
ISLE	16 Fev. 2008	http://www.ilc.cnr.it/EAGLES96/isle/ISLE_Home_Page.htm
KwiCFinder	16 Fev. 2008	http://www.kwicfinder.com/
Lácio-Web	16 Fev. 2008	http://www.nilc.icmc.usp.br/lacioweb/
LADL	16 Fev. 2008	http://infolingu.univ-mlv.fr/DonneesLinguistiques/Dictionnaires/delac.html
Linguateca	16 Fev. 2008	http://www.linguateca.pt
LX Lemmatizer	16 Fev. 2008	http://lxlemmatizer.di.fc.ul.pt/
Much.more	16 Fev. 2008	http://muchmore.dfki.de/

Projeto	Acesso em	Página
MXPOST	16 Fev. 2008	http://www.inf.ed.ac.uk/resources/nlp/local_doc/MXPOST.html
NILC	16 Fev. 2008	http://acdc.linguateca.pt/acesso/historiaCorpusNILC.html
NILC (córpus)	16 Fev. 2008	http://acdc.linguateca.pt/acesso/historiaCorpusNILC.html
NILC (Tradutor UNL)	16 Fev. 2008	http://www.nilc.icmc.usp.br/nilc/projects/unl.htm
NSP	16 Fev. 2008	http://www.d.umn.edu/~tpederse/nsp.html
OLIF	16 Fev. 2008	http://www.olif.net/
Open XML Editor	16 Fev. 2008	http://www.philo.de/xmledit/
Philologic	16 Fev. 2008	http://philologic.uchicago.edu/
PLN-BR	16 Fev. 2008	http://www.nilc.icmc.usp.br/plnbr/
Projeto ARTFL	16 Fev. 2008	http://humanities.uchicago.edu./orgs/ARTFL/
Programa Dicionário	21 Fev. 2008	http://www.nilc.icmc.usp.br/nilc/projects/unitex-pb/web/prog_dicionario.html
ReGra	16 Fev. 2008	http://www.nilc.icmc.usp.br/nilc/projects/regra.htm
RheSumaRST	16 Fev. 2008	http://www.nilc.icmc.usp.br/nilc/projects/SummRST.htm
rsnsr	24 Fev. 2008	http://www.is.informatik.uni-duisburg.de/projects/rsnsr/index.html.en
RSTTool	16 Fev. 2008	http://www.wagsoft.com/RSTTool/
SALT	16 Fev. 2008	http://www.loria.fr/projets/SALT/
SciPo	16 Fev. 2008	http://www.nilc.icmc.usp.br/nilc/projects/scipo.htm
Senter	16 Fev. 2008	http://www.icmc.usp.br/~taspardo/Senter.htm
System Quirk	16 Fev. 2008	http://www.computing.surrey.ac.uk/SystemQ/
TBL Tagger	16 Fev. 2008	http://www.cs.jhu.edu/~brill/code.html
Tenka Text	21 Fev. 2009	http://sourceforge.net/projects/corsis/
TEI Guidelines	16 Fev. 2008	http://www.tei-c.org/
The Bank of English	16 Fev. 2008	http://www.lingsoft.fi/doc/engcg/Bank-of-English.html
TIGER-XML	16 Fev. 2008	http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/doc/html/TigerXML.html
Tree Tagger	16 Fev. 2008	http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
Tycho-Brahe	16 Fev. 2008	http://www.ime.usp.br/~tycho/corpus/
Unitex	16 Fev. 2008	http://www-igm.univ-mlv.fr/~unitex/
Unitex-PB	16 Fev. 2008	http://www.nilc.icmc.usp.br/nilc/projects/unitex-pb/web/index.html
VIEW	16 Fev. 2008	http://view.byu.edu/
Webcorp	16 Fev. 2008	http://www.webcorp.org.uk/
Wordsmith Tools	16 Fev. 2008	http://www.lexically.net/wordsmith/
Xaira	16 Fev. 2008	http://www.oucs.ox.ac.uk/rts/xaira/
XCES	16 Fev. 2008	http://www.xml-ces.org/
XPDF	16 Fev. 2008	http://www.foolabs.com/xpdf/

Anexo A – Exemplo de anotação XCES

A seguir são mostrados os arquivos para um texto do córpus PLN-BR GOLD com anotação *stand-off*. Os arquivos compreendem o texto sem anotação (Figura A.1), o cabeçalho (Figura A.2), a anotação lógica (capítulos e parágrafos) (Figura A.3) e anotação de sentenças (Figura A.4). Adicionalmente, é ilustrado a versão mesclada do texto com as anotações (Figura A.5). No cabeçalho observa-se os códigos de categorias as quais o texto pertence pelo uso da etiqueta "<catRef>". As palavras chaves que caracterizam o assunto do texto se encontram na estrutura "<keywords>".

Membros de torcidas uniformizadas do Corinthians emboscaram na madrugada de ontem o ônibus em que a delegação do clube viajava para São Paulo, após a derrota por 1 a 0 para o Santos, na Vila Belmiro, pelo Brasileiro.
No km 45, após o trecho de serra da rodovia dos Imigrantes (sentido São Paulo), torcedores com camisa da Gaviões atravessaram um ônibus em que viajavam na pista, transformando-o numa barricada.
Quando o ônibus dos jogadores chegou, a torcida investiu contra ele, armada com pedras, paus e galhos arrancados de árvores.
O ônibus estava sem proteção da Polícia Rodoviária, porque a diretoria não fez o pedido.
No ataque, os agressores xingavam os jogadores de mercenários e visavam especialmente Souza, Mirandinha e Donizete _embora os dois últimos nem estivessem ali.
Orientando pelos seguranças do clube, os jogadores fecharam as cortinas e deitaram no corredor.
O ataque durou cerca de dez minutos e deixou dois feridos: o meia Rincón, que recebeu estilhaços de vidro na perna, e o motorista do ônibus, com corte no supercílio.
O vice de Futebol, José Mansur Farhat, e o diretor Jorge Neme não estiveram no cerco. Temendo represálias, deixaram o estádio antes do fim do jogo.
A Gaviões negou ter sido autora do ataque, mas seus diretores se contradizeram sobre o ocorrido.
Um diretor corintiano disse que viu diretores da Gaviões e o próprio presidente, Douglas Deúngaro, no ataque. Mas pediu para não ser identificado: "Sei do que eles são capazes de fazer".
Deúngaro, por outro lado, disse que chegou ao local por acaso e tentou conter os torcedores, que, segundo ele, não eram da Gaviões.
Mas a tônica no clube ontem era tentar abafar o caso. Ainda não registrou queixa na polícia e é provável que nem o faça.
Os grupos de oposição política no clube criticaram o elo entre a atual administração e a Gaviões.
O ataque surge em hora crítica para o Corinthians e para a Gaviões. O time está em 20º lugar no Brasileiro e corre risco de rebaixamento. Já a Gaviões, proibida como todas as uniformizadas de frequentar estádios paulistas, negociava com a PM e o Ministério Público um modo de retornar.
LEIA mais sobre o ataque ao ônibus do Corinthians nas págs. 4-3 e 4-4

Figura A.1: Texto original sem anotação

```
<?xml version="1.0" encoding="UTF-8"?>
<cesHeader xmlns="http://www.xces.org/schema/2003" xmlns:xlink="http://www.w3.org/1999/xlink"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://www.xces.org/schema/2003"
version="1.0.4">
        <fileDesc>
        <titleStmt>
                <title>1997out_6114</title>
                <respStmt>
                        <respType>Criação do Header</respType>
                        <respName type="person">Kleber Infante</respName>
                </respStmt>
                <respStmt>
                        <respType>Criação do Header</respType>
                        <respName type="person">Marcelo Muniz</respName>
                </respStmt>
        </titleStmt>
        <extent>
                <wordCount>372</wordCount>
                <byteCount units="bytes">4376.0</byteCount>
                <extNote>2</extNote>
        </extent>
        <publicationStmt>
                <pubAddress>Av. Trabalhador São-carlense, 400 - Centro, Caixa Postal: 668 - CEP: 13560-970 - São
Carlos - SP</pubAddress>
                <telephone>+55 16 33739663</telephone>
                <eAddress type="www">http://www.nilc.icmc.usp.br</eAddress>
                <pubDate>2006</pubDate>
        </publicationStmt>
        <sourceDesc>
                <biblStruct>
                        <monogr>
                                <title>Corinthians sofre ataque de guerrilha</title>
                                <title>Ônibus do clube, parado em rodovia (....)</title>
                                <author>da Reportagem Local</author>
                                <respStmt>
                                        <respType>crédito</respType>
                                        <respName type="institution">DA REPORTAGEM LOCAL</respName>
                                </respStmt>
                                <imprint>
                                        <pubPlace>Folha de São Paulo</pubPlace>
                                        <publisher type="org">Empresa Folha da Manhã S.A.</publisher>
                                        <pubDate>16/10/97</pubDate>
                                        <pubAddress>São Paulo</pubAddress>
                                </imprint>
                                <biblNote>ESPORTE</biblNote>
                                <biblScope type="PP">4-1</biblScope>
                        </monogr>
                </biblStruct>
        </sourceDesc>
        </fileDesc>
        <encodingDesc>
        <projectDesc>O projeto Recursos e Ferramentas para a Recuperação de (...)</projectDesc>
        <samplingDecl>PLN-BR GOLD é o córpus gold standard do Projeto PLN-BR (...)</samplingDecl>
        </encodingDesc>
        <profileDesc>
        <textClass>
                <catRef target="genero.8 genero.8.18 genero.8.18.10  distribuicao.12 tipotextual.35 " />
                <keywords>
                        <keyTerm>VIOLÊNCIA</keyTerm>
                        <keyTerm>AGRESSÃO</keyTerm>
                        <keyTerm>ATAQUE</keyTerm>
                        <keyTerm>JOGADOR</keyTerm>
                        <keyTerm>FUTEBOL</keyTerm>
                        <keyTerm>CORINTHIANS</keyTerm>
                        <keyTerm>CLUBE</keyTerm>
                        <keyTerm>GAVIÕES DA FIEL</keyTerm>
                        <keyTerm>TORCIDA ORGANIZADA</keyTerm>
                </keywords>
        </textClass>
        <annotations>
                <annotation type="logical"  ann.loc="ESPORTE_1997_640-logical.xml"  >Logical
markup</annotation>
                <annotation type="s"  ann.loc="ESPORTE_1997_640-s.xml"  >Sentence boundaries</annotation>
                <annotation type="content"  ann.loc="ESPORTE_1997_640.txt"  >Document content</annotation>
        </annotations>
        </profileDesc>
</cesHeader>
```

Figura A.2: Cabeçalho

```
<?xml version="1.0" encoding="UTF-8" ?>
 <cesAna xmlns="http://www.xces.org/schema/2003" version="1.0.4">
 <struct type="cesDoc" from="0" to="2193">
 <feat name="version" value="1.0.4" />
 <feat name="id" value="ESPORTE_1997_640" />
 <feat name="xmlns:xsi" value="http://www.w3.org/2001/XMLSchema-instance" />
 <feat name="xmlns:xlink" value="http://www.w3.org/1999/xlink" />
 <feat name="xmlns" value="http://www.xces.org/schema/2003" />
 </struct>
 <struct type="text" from="0" to="2192" />
 <struct type="body" from="1" to="2191" />
 <struct type="div" from="2" to="2190">
 <feat name="type" value="materia" />
 </struct>
 <struct type="p" from="3" to="219">
 <feat name="id" value="p1" />
 </struct>
 <struct type="p" from="220" to="413">
 <feat name="id" value="p2" />
 </struct>
 <struct type="p" from="414" to="538">
 <feat name="id" value="p3" />
 </struct>
 <struct type="p" from="539" to="627">
 <feat name="id" value="p4" />
 </struct>
 <struct type="p" from="628" to="786">
 <feat name="id" value="p5" />
 </struct>
 <struct type="p" from="787" to="882">
 <feat name="id" value="p6" />
 </struct>
 <struct type="p" from="883" to="1048">
 <feat name="id" value="p7" />
 </struct>
 <struct type="p" from="1049" to="1196">
 <feat name="id" value="p8" />
 </struct>
 <struct type="p" from="1197" to="1293">
 <feat name="id" value="p9" />
 </struct>
 <struct type="p" from="1294" to="1482">
 <feat name="id" value="p10" />
 </struct>
 <struct type="p" from="1483" to="1614">
 <feat name="id" value="p11" />
 </struct>
 <struct type="p" from="1615" to="1735">
 <feat name="id" value="p12" />
 </struct>
 <struct type="p" from="1736" to="1833">
 <feat name="id" value="p13" />
 </struct>
 <struct type="p" from="1834" to="2119">
 <feat name="id" value="p14" />
 </struct>
 <struct type="p" from="2120" to="2189">
 <feat name="id" value="p15" />
 </struct>
 </cesAna>
```

Figura A.3: Anotação lógica

```
<?xml version="1.0" encoding="UTF-8" ?>
  <cesAna xmlns="http://www.xces.org/schema/2003" version="1.0.4">
  <struct type="s" from="3" to="219">
  <feat name="id" value="p1s1" />
  </struct>
  <struct type="s" from="220" to="413">
  <feat name="id" value="p2s1" />
  </struct>
  <struct type="s" from="414" to="538">
  <feat name="id" value="p3s1" />
  </struct>
  <struct type="s" from="539" to="627">
  <feat name="id" value="p4s1" />
  </struct>
  <struct type="s" from="628" to="786">
  <feat name="id" value="p5s1" />
  </struct>
  <struct type="s" from="787" to="882">
  <feat name="id" value="p6s1" />
  </struct>
  <struct type="s" from="883" to="1048">
  <feat name="id" value="p7s1" />
  </struct>
  <struct type="s" from="1049" to="1134">
  <feat name="id" value="p8s1" />
  </struct>
  <struct type="s" from="1135" to="1196">
  <feat name="id" value="p8s2" />
  </struct>
  <struct type="s" from="1197" to="1293">
  <feat name="id" value="p9s1" />
  </struct>
  <struct type="s" from="1294" to="1403">
  <feat name="id" value="p10s1" />
  </struct>
  <struct type="s" from="1404" to="1482">
  <feat name="id" value="p10s2" />
  </struct>
  <struct type="s" from="1483" to="1614">
  <feat name="id" value="p11s1" />
  </struct>
  <struct type="s" from="1615" to="1668">
  <feat name="id" value="p12s1" />
  </struct>
  <struct type="s" from="1669" to="1735">
  <feat name="id" value="p12s2" />
  </struct>
  <struct type="s" from="1736" to="1833">
  <feat name="id" value="p13s1" />
  </struct>
  <struct type="s" from="1834" to="1901">
  <feat name="id" value="p14s1" />
  </struct>
  <struct type="s" from="1902" to="1971">
  <feat name="id" value="p14s2" />
  </struct>
  <struct type="s" from="1972" to="2119">
  <feat name="id" value="p14s3" />
  </struct>
  <struct type="s" from="2120" to="2179">
  <feat name="id" value="p15s1" />
  </struct>
  <struct type="s" from="2180" to="2189">
  <feat name="id" value="p15s2" />
  </struct>
  </cesAna>
```

Figura A.4: Anotação de sentenças

```
<?xml version="1.0" encoding="UTF-8" ?>
 <cesDoc version="1.0.4" id="ESPORTE_1997_640" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns:xlink="http://www.w3.org/1999/xlink" xmlns="http://www.xces.org/schema/2003">
 <text>
 <body>
 <div type="materia">
 <p id="p1">
 <s id="p1s1">Membros de torcidas uniformizadas do Corinthians emboscaram na madrugada de ontem o ônibus em que a
delegação do clube viajava para São Paulo, após a derrota por 1 a 0 para o Santos, na Vila Belmiro, pelo Brasileiro.</s>
 </p>
 <p id="p2">
 <s id="p2s1">No km 45, após o trecho de serra da rodovia dos Imigrantes (sentido São Paulo), torcedores com camisa da
Gaviões atravessaram um ônibus em que viajavam na pista, transformando-o numa barricada.</s>
 </p>
 <p id="p3">
 <s id="p3s1">Quando o ônibus dos jogadores chegou, a torcida investiu contra ele, armada com pedras, paus e galhos
arrancados de árvores.</s>
 </p>
 <p id="p4">
 <s id="p4s1">O ônibus estava sem proteção da Polícia Rodoviária, porque a diretoria não fez o pedido.</s>
 </p>
 <p id="p5">
 <s id="p5s1">No ataque, os agressores xingavam os jogadores de mercenários e visavam especialmente Souza, Mirandinha e
Donizete _embora os dois últimos nem estivessem ali.</s>
 </p>
 <p id="p6">
 <s id="p6s1">Orientando pelos seguranças do clube, os jogadores fecharam as cortinas e deitaram no corredor.</s>
 </p>
 <p id="p7">
 <s id="p7s1">O ataque durou cerca de dez minutos e deixou dois feridos: o meia Rincón, que recebeu estilhaços de vidro na
perna, e o motorista do ônibus, com corte no supercílio.</s>
 </p>
 <p id="p8">
 <s id="p8s1">O vice de Futebol, José Mansur Farhat, e o diretor Jorge Neme não estiveram no cerco.</s>
 <s id="p8s2">Temendo represálias, deixaram o estádio antes do fim do jogo.</s>
 </p>
 <p id="p9">
 <s id="p9s1">A Gaviões negou ter sido autora do ataque, mas seus diretores se contradizeram sobre o ocorrido.</s>
 </p>
 <p id="p10">
 <s id="p10s1">Um diretor corintiano disse que viu diretores da Gaviões e o próprio presidente, Douglas Deúngaro, no ataque.</
s>
 <s id="p10s2">Mas pediu para não ser identificado: "Sei do que eles são capazes de fazer".</s>
 </p>
 <p id="p11">
 <s id="p11s1">Deúngaro, por outro lado, disse que chegou ao local por acaso e tentou conter os torcedores, que, segundo ele,
não eram da Gaviões.</s>
 </p>
 <p id="p12">
 <s id="p12s1">Mas a tônica no clube ontem era tentar abafar o caso.</s>
 <s id="p12s2">Ainda não registrou queixa na polícia e é provável que nem o faça.</s>
 </p>
 <p id="p13">
 <s id="p13s1">Os grupos de oposição política no clube criticaram o elo entre a atual administração e a Gaviões.</s>
 </p>
 <p id="p14">
 <s id="p14s1">O ataque surge em hora crítica para o Corinthians e para a Gaviões.</s>
 <s id="p14s2">O time está em 20º lugar no Brasileiro e corre risco de rebaixamento.</s>
 <s id="p14s3">Já a Gaviões, proibida como todas as uniformizadas de frequentar estádios paulistas, negociava com a PM e o
Ministério Público um modo de retornar.</s>
 </p>
```

Figura A.5: Texto mesclado com anotações

Anexo B – Domínios, subdomínios e gêneros e subgêneros utilizados no Projeto DHPB

A seguir são mostrados os domínios, subdomínios, gêneros e subgêneros definidos para o Projeto DHPB e levantados por Jacqueline Souza.

Domínio Discursivo 1 – Religioso

1.1.Auto de confissão
1.2.Breve (Rescrito papalino que contém uma decisão de caráter particular)
1.3.Carta pastoral
1.4.Epístola
1.5.Memento (Cada uma das duas preces do cânon da missa)
1.6.Oração
1.7.Sermão
1.8.Voto

Domínio Discursivo 2 – Jurídico

2.1. Legislativo
2.1.1.Decreto
2.1.2.Lei
2.1.3.Medida provisória
2.1.4.Portaria
2.1.5.Resolução
2.2. Jurisprudência
2.2.1.aresto
2.2.2.minuta
2.2.3.petição
2.2.4.sentença
2.2.5.súmula
2.3. Jurídico-Administrativo
2.3.1.Abaixo-assinado
2.3.2.alvará
2.3.3.assento (7.Anotação, registro, apontamento. 8.Termo de qualquer ato oficial. In Aurélio Eletrônico, 2004)
2.3.4.ata
2.3.5.Atestação (Declaração escrita e assinada sobre a verdade de um fato, para servir a outrem de documento; atestado, testemunho In Aurélio Eletrônico, 2004)
2.3.6.atestado
2.3.7.auto (Registro escrito e autenticado de qualquer ato In Aurélio Eletrônico, 2004)
2.3.7.1. auto de abertura
2.3.7.2. auto de anulação
2.3.7.3. auto de assento
2.3.7.4. auto de averiguação
2.3.7.5. auto de demarcação
2.3.7.6. auto de diligência
2.3.7.7. auto de inquirição
2.3.7.8. auto de justificação
2.3.7.9. auto de posse
2.3.7.10. auto de vereação
2.3.8. Despacho

2.3.9. foral (1.Carta de lei que regulava a administração duma localidade ou concedia privilégio a indivíduos ou corporações. 2.Carta de aforamento de terras; foro. In Aurélio Eletrônico, 2004)
2.3.10. Inventário
2.3.11. Lançamento (Jur. Ato pelo qual, em certos casos, o juiz afasta da ação penal pública o acusador privado (querelante), por não haver ele apresentado o libelo no devido prazo, declarando-a perempta ou devolvendo-a ao Ministério Público. In Aurélio Eletrônico, 2004)
2.3.12. Notificação
2.3.13. Ofício
2.3.14. Procuração
2.3.15. Provimento (5.Jur. Manifestação dos tribunais superiores ao receberem e julgarem favoravelmente o recurso interposto contra decisões dos juízes inferiores. 6.Jur. Instruções ou determinações administrativas baixadas pelo corregedor ao realizar as correições. In Aurélio Eletrônico, 2004)
2.3.16. Termo (13.Jur. Peça em que se formaliza determinado ato processual. In Aurélio Eletrônico, 2004)
2.3.16.1. Termo de declaração
2.3.16.2. Termo de vereação
2.3.16.3. Termo em junta
2.4. Jurídico Comercial
2.4.1. Contrato
2.4.2. Escritura (ou compromisso)
2.4.3. Registro
2.4.4. Representação (13.Jur. Contrato remunerado, firmado entre dois comerciantes ou empresas comerciais, para que uma parte promova a venda de produtos da outra, efetuando negócios em nome dela, ou realize aproximação de fregueses, etc., mediante condições variáveis em cada caso. Bras. Jur. Pedido que a vítima de certos delitos — ou seus representantes legais — formula à autoridade policial ou judiciária, e bem assim ao órgão do Ministério Público, para que se proceda contra o delinqüente, sem o que será nula a ação penal que se intentar na espécie. In Aurélio Eletrônico, 2004)
2.5. Jurídico Civil
2.5.1 Certidão
2.5.1.1. certidão de justificação
2.5.2. Certificado
2.5.3. Testamento
2.6 Jurídico comunicacional
2.6.1. Edital
2.6.2. Édito (1.Ordem judicial publicada por anúncios ou editais. In Aurélio Eletrônico, 2004)
2.6.3. Exposição de motivos (1.Na linguagem burocrática, ofício dirigido por Ministro de Estado ao Presidente da República. In Aurélio Eletrônico, 2004)

Domínio Discursivo 3 – Científico

3.1. Divulgação
3.1.1. artigo
3.1.2. ensaio
3.1.3. resenha
3.1.4. resumo
3.2. Pesquisa
3.2.1. dissertação
3.2.2. monografia
3.2.3. projeto
3.2.4. tese

Domínio Discursivo 4 – Informativo

4.1. Jornalístico
4.1.1. editorial
4.1.2. entrevista
4.1.3. reportagem
4.2. Informe

4.2.1. aviso
4.2.2. boletim
4.2.3. comunicado

Domínio Discursivo 5 – Referencial

5.1. catálogo
5.2. índex
5.5. verbete

Domínio Discursivo 6 – Instrucional

6.1. Didático
6.1.1. apostila
6.1.2. livro-texto
6.2. Procedimental
6.2.1. bula
6.2.2. manual
6.2.3. receita

Domínio Discursivo 7 – Técnico administrativo e/ou oficial b.7

7.1. Comunicacional – entre 2 ou mais pessoas, informando, solicitando, exigindo....
7.1.1. ato
7.1.1.1. ato de nomeação
7.1.1.2. ato de sujeição e obediência e vassalagem
7.1.1.3. aviso
7.1.1.3.1. aviso público
7.1.2. carta
7.1.2.1.carta de apresentação
7.1.2.2.carta régia
7.1.2.3.carta de abrasão de armas de nobreza e fidalguia
7.1.2.4.carta de confirmação
7.1.2.5.carta de conta
7.1.2.6.carta de diligência
7.1.2.7.carta de doação
7.1.2.8.carta de examinação
7.1.2.9.carta de mercê
7.1.2.10.carta de nomeação
7.1.2.11.carta de ofício
7.1.2.12.carta de ordenança
7.1.2.13.carta de prego (Carta fechada, na qual se determina o que o comandante de um navio deve fazer, e que ele só deve abrir quando fora da barra.In Aurélio Eletrônico, 2004)
7.1.2.14.carta de privilégio
7.1.2.15.carta de propriedade
7.1.2.16.carta de sentença
7.1.2.17.carta oficial
7.1.2.18.carta patente
7.1.3. circular
7.1.4. declaração
7.1.5. despacho
7.1.6. informação de serviço
7.1.7. memorando
7.1.8. ofício
7.1.9. provisão (5.Documento oficial em que o governo confere cargo, mercê, dignidade, ofício, etc., autoriza o exercício de uma profissão ou expede instruções. In Aurélio Eletrônico, 2004)
7.1.10. requerimento
7.1.11. solicitação

7.2. Descritivo
7.2.1.ata
7.2.2.auto de exame médico
7.2.3.balanço (financeiro)
7.2.4.diário (Livro onde se registram, em ordem cronológica, todas as operações contabilizáveis de uma empresa. In Aurélio Eletrônico, 2004)
7.2.4.1. diário (de viagem)
7.2.4.2. diário do governo
7.2.5.informe (Bras. Mil. Qualquer documento, fotografia, mapa, relatório ou observação, relativos ao inimigo ou a uma conjuntura complexa, e que pode contribuir para esclarecer a situação dele ou dela. In Aurélio Eletrônico, 2004)
7.2.6.levantamento
7.2.7.ordem do dia (2.Mil. Conjunto de determinações e instruções divulgadas diariamente por comandante militar. In Aurélio Eletrônico, 2004)
7.2.8.panejamento
7.2.9.regimento
7.2.9.1. regimento das fronteiras
7.2.10.relatório (relação, lista, quadros demonstrativos...)
7.3. Comercial
7.3.1. conhecimento (Econ. Documento representativo de mercadoria depositada ou entregue para transporte, e que, se endossado, pode ser negociado como título de crédito. In Aurélio Eletrônico, 2004)
7.3.2. contrato
7.3.3. nota
7.3.4. recibo

Domínio Discursivo 8 – Literário b.8

8.1. Prosa
8.1.1. biografia
8.1.2. conto
8.1.3. crônica
8.1.4. ensaio
8.1.5. libelo (artigo ou escrito de caráter satírico ou difamatório; panfleto In Aurélio Eletrônico, 2004)
8.1.6. novela
8.1.7. resenha
8.1.8. romance
8.2. Poesia/poema
8.2.1. elegia
8.2.2. ode
8.2.3. soneto
8.3. Teatro
8.3.1. ato (Cada uma das maiores partes em que se divide a peça, e cujo número pode variar, ger., de um a cinco. In Aurélio Eletrônico, 2004)
8.3.2. peça (Texto e/ou representação teatral. In Aurélio, 2004)

Domínio Discursivo 9 – Pessoal

9.1. Correspondência
9.1.1. anotação
9.1.2. bilhete
9.1.3. carta (ou missiva)
9.1.4. dedicatória
9.1.5. memento (2.Marca que serve para lembrar qualquer coisa. 3.Papel ou caderneta onde se anotam coisas que devem ser lembradas; memorial, memorando, memória. 4.Essa anotação; apontamento, memória. 5.Livrinho onde se acham resumidas as partes essenciais de uma questão. In Aurélio Eletrônico, 2004)
9.1.6. elogio

9.2. Documento Pessoal
9.2.1. certidão
9.2.2. certificado
9.2.3. diploma