# Um Ambiente Computacional para o Processamento de Córpus de Português Histórico

Arnaldo Candido Junior (ICMC -USP)
Sandra Maria Aluísio (ICMC -USP)
{arnaldoc, sandra} at icmc.usp.br

**Resumo**: A utilização de córpus tem crescido progressivamente em áreas como a Lingüística e o Processamento de Línguas Naturais. Entretanto, o processamento computacional de córpus históricos apresenta vários problemas não encontrados em córpus contemporâneos. Neste trabalho, são apresentados os desafios encontrados para o processamento de córpus de Português Histórico, no escopo do projeto Dicionário Histórico do Português do Brasil. Um ambiente computacional para processamento de córpus, criação de glossários, e redação de verbetes foi desenvolvido, podendo também ser utilizado em outros projetos de criação de dicionários históricos ou de descrição lingüística.

**Palavras chave**: Processamento de Línguas Naturais, Processamento de Córpus, Córpus Históricos

Dissertação defendida em 02/04/2008
Disponível em: <http://www.nilc.icmc.usp.br/nilc/publications.htm#Theses>

## 1 Introdução[1]

Córpus 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 desta língua ou uma variedade lingüística [19]. A construção e uso de córpus eletrônicos ainda estão em sua infância, embora um grande progresso em relação ao projeto de córpus tenha ocorrido. Existem diversos projetos de córpus, citados por diferentes autores [1, 4, 10, 19], com objetivos e finalidades distintas, alguns exemplos de projetos são: (a) *Brown Corpus of Standard American English* (para o Inglês); (b) *The Bank of English* (para o Inglês) (c) FRANTEXT (para o Francês); ((d) Córpus do NILC [14] (para o Português); (e) Córpus do projeto Lácio-Web [2] (para o Português) e (e) Tycho-Brahe [7] (para o Português).

O projeto Dicionário Histórico do Português do Brasil (DHPB) 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 histórico. O córpus já está totalmente compilado, contando com 2.458 textos e 7.5 milhões de formas simples, com anotação em TEI

1 O presente trabalho foi realizado com apoio do CNPq, Conselho Nacional de Desenvolvimento Científico e Tecnológico - Brasil

[21]. 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 produzirá o primeiro dicionário histórico voltado para a variante brasileira do Português, que começou a diferir de sua variante européia já nos primeiros séculos da nossa história. No Brasil, iniciativas para a construção de córpus históricos são particularmente importantes, pois possibilitam a preservação da história do país e do seu registro lingüístico, além de possibilitar o estudo da evolução da língua no período estudado. Adicionalmente, projetos com este perfil são raros.

As ferramentas computacionais são um importante auxílio para a construção e o estudo de um córpus. Um comparativo entre várias ferramentas disponíveis livremente é apresentado na Seção 2. Durante o desenvolvimento do projeto DHPB foi constatado que, apesar da ampla gama de ferramentas computacionais disponíveis para processamento de córpus, poucas delas cobriam de maneira desejável os requisitos para a construção de córpus históricos do Português. 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 (ou léxicos computacionais) para apoiar a tarefa lexicográfica. As Seções 3, 4 e 5 cobrem as necessidades de processamento de córpus históricos. O objetivo deste trabalho de mestrado foi levantar os problemas típicos de córpus históricos para informar a construção de recursos, (por exemplo, o próprio córpus em formato computacional), metodologias e ferramentas para o projeto DHPB, com a expectativa de aumentar a produtividade dos pesquisadores envolvidos. Para isso, um ambiente computacional foi proposto (veja Seção 6), sendo útil não apenas ao projeto DHPB, mas também a outros projetos baseados em córpus históricos em Português. O ambiente computacional foi concebido com o intuito de agilizar o trabalho com córpus e dicionários históricos, reduzindo o tempo para o desenvolvimento dessas pequisas e aumentando a sua qualidade. Além disso, recursos prontos são fornecidos na forma de ferramentas e glossários. As conclusões do trabalho são apresentadas na Seção 7.

## 2 Sistemas de Processamento de Córpus

Existem diferentes comparativos entre ferramentas para processamento de córpus [15, 18, 20, 23], mas, em geral, não são focados em ferramentas livres. Neste trabalho, é feito um comparativo com foco em ferramentas livres, abrangendo cinco ferramentas: GATE, *Philologic*, *Tenka Text*, *Unitex* e *Xaira*. **GATE** é um sistema para engenharia da linguagem desenvolvido e mantido pela Universidade de *Sheffield*, capaz de fornecer uma infra-estrutura robusta para o desenvolvimento e a distribuição de softwares para PLN (Processamento de Língua Natural). O sistema, desenvolvido em *Java*, teve sua primeira versão lançada em 1995 e tem sido usado em grandes organizações para pesquisas variadas, como criação de ontologias e anotação semântica. ***Philologic*** é uma ferramenta *Web* para análise de córpus desenvolvida por pesquisadores da universidade de Chicago como uma das metas do projeto ARTFL. A ferramenta foi originalmente desenvolvida para gerenciar textos em francês, mas devido à codificação *Unicode*, diversos idiomas são permitidos. ***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*, disponibilizado para 5 sistemas operacionais. A ferramenta oferece uma biblioteca e uma interface com duas ferramentas: lista de palavras e um concordanceador. ***Unitex*** [13] 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 na Universidade de *Marne-la-Vallée* na França como uma implementação livre do software *Intex*. Em [12], foram construídas regras de resolução de ambigüidades, bibliotecas para acesso a glossários compactados e ferramentas para a Língua Portuguesa. ***Xaira*** 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.

Os cinco processadores de córpus foram avaliados quanto à qualidade de software pelas seis métricas definidas na ISO 9126 [23]: funcionalidade, confiabilidade, usabilidade, eficiência, manutenibilidade e portabilidade. A Tabela 1 mostra parte dos critérios utilizados na análise dos processadores de córpus. Mais detalhes sobre as métricas e os critérios da Tabela 1 podem ser encontrados em [5].

Tabela 1: Comparativo entre ferramentas para processamento de córpus

| **Critério** | **GATE (build 2752)** | **Philologic 3.1** | **Unitex 2.0 beta** | **Tenka 0.1.3.2** | **Xaira 1.23** |
|---|---|---|---|---|---|
| concordanceador | sim | sim | sim | sim | sim |
| contador de freqüência | não | sim | sim | sim | sim |
| busca orientada a glossário | sim | sim | não | sim | sim |
| processamento de anotação | sim (XCES) | sim (TEI-Lite) | Parcial (gramatical) | Parcial (gramatical) | sim (TEI ou similar) |
| colocações ou *n*-gramas | sim | sim | não | não | sim |
| codificação de caracteres | UTF-8 | UTF-8 | UTF-16 | UTF, ISO, etc | UTF-8/16 |
| tempo de pré-processamento (segs.) | 663 | 61,5 | 19,5 | 0 | 36,9 |
| tempo do concordanceador | 212 | 1,5 | 8 | 13,5 | 0,7 |

No caso de uso de córpus para tarefas lexicográficas, a geração de concordâncias, é um importante recurso. Neste requisito as ferramentas *Philologic*, *Unitex* e *Xaira* apresentaram-se como boas opções. O GATE não possui um concordanceador por padrão, mas um pode ser criado a partir dos recursos oferecidos pela ferramenta. Entretanto, o procedimento é complexo e o tempo de resposta do concordanceador é alto. Por outro lado, o GATE possui bons recursos, podendo ser utilizado para outras pesquisas, por exemplo, análise sintática e etiquetagem de córpus. O *Tenka*, por sua vez, possui um concordanceador de fácil uso, mas apresentou problemas de desempenho, uma vez que não indexa os textos. Outro problema do *Tenka* é ainda estar em fase inicial de desenvolvimento, possuindo poucos recursos disponíveis. Seu uso é mais indicado para usuários que procuram uma alternativa ao *WordSmith Tools*.

Para o projeto DHPB, as ferramentas mais indicadas foram o *Philologic*, o *Unitex* e o *Xaira*. O *Xaira*, apesar de possuir um grande número de recursos para as buscas, não foi utilizado uma vez que sua interface é de difícil uso para usuários iniciantes. Foram escolhidos então o *Philologic* e o *Unitex*. A escolha do *Philologic* se deu por sua interface de fácil utilização e pela centralização dos dados oferecida por ferramentas *Web*, além de permitir o levantamento de variantes de grafia através de distância de edição (com auxílio do utilitário *Agrep*). O *Unitex* foi escolhido devido ao processamento de glossários, simplificando buscas por abreviaturas e variantes de grafia. As ferramentas, em geral, possuem um grande número de recursos, entretanto, poucos são voltados para a construção de córpus históricos. A Seção 3 descreve parte dos problemas encontrados em córpus históricos, que podem ser tratados com o uso de ferramentas adequadas.

## 3 Processamento de Córpus Históricos para Tarefas Lexicográficas

O trabalho com córpus históricos é mais difícil do que com córpus contemporâneos, uma vez que os primeiros possuem características não encontrados nos segundos. Algumas características comuns em textos históricos foram levantadas em [17], entre elas: ausência de hifenização, gerando não-palavras, variações de grafia, junções de palavras (exemplo: "éamor"), uso símbolos tipográficos incomuns e a alta freqüência de abreviaturas. No córpus do projeto DHPB ocorrem problemas semelhantes, discutidos a seguir.

**Símbolos tipográficos incomuns** podem ser tratados através da escolha de uma boa codificação de caracteres e o uso de etiquetas adequadas para denotá-los, como a etiqueta "<symbol>" do padrão TEI. O uso de *Unicode* é particularmente importante em córpus históricos, pois é comum encontrar caracteres não permitidos pelos padrões de codificação usuais, por exemplo, o símbolo "æ" (união de "a" e "e") e o símbolo "m̃" ("com̃ercio"). No projeto DHPB, optou-se pelo uso da codificação *Unicode* para tratar esses símbolos, já que este é capaz de representar todos os símbolos encontrados nos textos históricos. O uso de **abreviaturas** é comum em manuscritos e também nos primeiros materiais impressos. O processamento de abreviaturas é particularmente difícil, pois estas são ambíguas e podem ter um grande número de expansões. Por exemplo, entre as expansões da abreviatura "A" estão "alteza", "alvará", "Amaro", "Ana", "anima", entre outras. Além disso, uma única palavra pode possuir um grande número de abreviaturas diferentes. Por exemplo, "Janeiro" pode ser abreviada como "Jan.ro", "Janr.o", "Jnro", entre outras. Embora existam técnicas para expansão automática de abreviaturas para as línguas contemporâneas [22], há pouca pesquisa sobre o assunto para tratamento de abreviaturas em textos históricos. 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 abordagem alternativa é o uso de glossários de abreviaturas para auxiliar a expansão de abreviatura durante as pesquisas no córpus. Está última abordagem foi a escolhida no projeto DHPB, por ser mais rápida para ser implementada e pouco propensa a erros.

As **variações de grafia** de uma mesma palavra ocorrem nos textos do córpus

DHPB, pois esses textos foram escritos em uma época em que não havia um sistema ortográfico unificado para o idioma Português do Brasil. Algumas práticas comuns em textos em Português anteriores ao século XVIII, levantadas por [11], são: consoantes dobradas, inconsistência no uso de acentuação e troca entre vogais. 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 ocorrências de uma lexia. Existem diversos trabalhos para detecção automática de variantes de grafia [3, 9, 16]. Em [8], propomos uma abordagem de uso de regras de transformação criadas manualmente, usada por um sistema para detecção automática de variação de grafias. A proposta foi baseada nos trabalhos de [9] e [11]. Um exemplo de regra de transformação é a tripla (ea, e, ei), aplicada a grafias que possuam as letras "ea". Se aplicada à grafia "aldea" a regra substituirá "e" por "ei", formando a nova grafia "aldeia", o que permite agrupar "aldea" junto com a grafia "aldeia". As **junções** também dificultam as buscas no córpus. A solução mais adequada neste caso é a separação das junções. O padrão TEI permite a anotação de junções através da etiqueta "<choice>". 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"). A Seção 4 apresenta a metodologia utilizada para lidar com os problemas encontrados no córpus do projeto DHPB.

## 4 Metodologia Utilizada

A metodologia apresentada a seguir foi baseada nas decisões do grupo de pesquisadores em reuniões do projeto DHPB. As reuniões foram organizadas com periodicidade média de seis meses. Inicialmente, o córpus foi pré-processado. A seguir, glossários foram criados para tratar os fenômenos de abreviaturas, junções de palavras e variantes de grafia. Durante o processo, ferramentas foram desenvolvidas ou adaptadas para processar o córpus, acessar os textos e permitir a redação de verbetes no dicionário histórico. As ferramentas e os glossários apresentados nesta seção estão disponíveis publicamente (http://www.nilc.icmc.usp.br/nilc/projects/procorph).

As tarefas de **pré-processamento do córpus** DHPB são a limpeza e a anotação dos textos digitalizados. Os textos foram digitalizados em formato DOC e convertidos para um formato TXT (para isso foi desenvolvida a ferramenta Protew [5]). O formato TXT, por sua vez, permite a geração de córpus em XML simplificado usado para a criação de um córpus em formato TEI ou em formato de texto puro com informações catalográficas. As versões do córpus são usadas nos processadores de córpus *Philologic* e *Unitex*. Para a geração dos diferentes formatos do córpus, foi desenvolvida a ferramenta Protej [5]. Exemplos de tarefas realizadas por essas duas ferramentas são: conversão da ficha catalográfica dos textos para XML, remoção automática de hifenização quando possível e tratamento de numeração de linhas e de parágrafos. O processamento pode ser automático ou semi-automático, conforme a tarefa a ser realizada.

Os **glossários** desenvolvidos foram três: (a) um glossário de abreviaturas e suas

expansões (chamado de glossário F), (b) um glossário de junções de palavras (criado manualmente) e (c) um glossário de variantes de grafia para auxiliar a busca por concordâncias e a contagem de freqüências. Os glossários de abreviaturas e de variantes seguem o formato DELA [13], utilizado pelo *Unitex*. As palavras iniciadas pelas letras A, B e C do glossário F receberam também informações morfossintáticas e semânticas [24], para criação de buscas mais elaboradas na ferramenta *Unitex* (por exemplo, buscas por abreviaturas de entidades nomeadas). Esse glossário foi extraído de [6] e de listas de abreviaturas anexas a textos do córpus DHPB. Um quarto glossário (chamado de glossário C) contendo abreviaturas extraídas automaticamente do córpus (sem as expansões) foi criado para avaliar a abrangência do glossário F em relação ao córpus. Foram utilizadas 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 (sucedido por até quatro caracteres), por exemplo: "jan.ro" (janeiro) e (c) palavras terminadas por algumas consoantes e sucedidas por ponto final, por exemplo "av." (avenida). O glossário de variantes de grafia foi criado com a ajuda da Ferramenta Siaconf, desenvolvida por um participante do projeto DHPB. A metodologia para detecção de variações de grafia se baseia na aplicação de regras de transformação ao córpus. Ao todo, foram criadas 43 regras de transformação. Além do glossário de grafias, as variantes também podem ser procuradas no córpus através do *Philologic*, que utiliza algoritmos de distância de edição para agrupar grafias semelhantes. Um exemplo de algoritmo de distância de edição é o utilizado pelo revisor ortográfico do *MS Word*.

As **ferramentas utilizadas** podem ser divididas em três grupos: pré-processamento do córpus (Protew e Protej), acesso ao córpus (*Philologic* e *Unitex*) e redação de verbetes. Para a criação do dicionário histórico, foi desenvolvida a ferramenta Procorph [5], pois os verbetes vem sendo redigidos com o auxílio do *MS Word*. Entretanto, o uso do *MS Word* não é adequado para a tarefa, pois um redator não tem acesso ao verbete dos demais (os verbetes não são distribuídos para evitar problemas de sincronização) e a formatação do verbete é não é automática. Outro problema se relaciona a variações na forma e no conteúdo dos verbetes, dificultando sua padronização. O *Procorph* é disponibilizado via *Web*, simplificando o acesso dos participantes do projeto, além de permitir a padronização dos verbetes. Quatro níveis de acesso são fornecidos aos usuários: consulente (acessa e consulta a base), redator (inclui verbetes na base e altera os próprios verbetes), revisor (pode alterar verbetes durante a tarefa de revisão de verbetes) e administrador (pode cadastrar usuários). A ferramenta foi testada por 4 dos 21 redatores do projeto DHPB, que concluíram que o sistema apresentou um bom desempenho para a função para a qual foi desenvolvido. Seção 5 apresenta a avaliação do córpus, das ferramentas desenvolvidas, e dos glossários construídos.

## 5 Avaliação

O tamanho do córpus é um indicativo para avaliar a robustez das ferramentas usadas em seu processamento. A Tabela 2 contém informações sobre o córpus. Os dados da Tabela 2 foram estimados pelo *Unitex*. *Tokens* incluem palavras, números, sinais de pontuação, espaço, entre outros. *Types* correspondem a *tokens* sem as repetições.

Formas simples são as palavras do texto, cuja formação é definida em um arquivo de configuração do *Unitex* "Alphabet.txt". Formas simples únicas representam as formas simples sem as repetições. O fato de poucas mudanças terem sido necessárias nas ferramentas durante todo o processo de compilação do córpus sugere que estas são robustas e capazes de pré-processar diversos textos históricos com poucas adaptações.

Tabela 2: Dados 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 |

A Figura 1 mostra o gráfico percentual da distribuição do córpus por século. Há poucos textos do século XVI, pois neste período existiam poucos brasileiros alfabetizados e parte dos documentos se perderam devido à ação do tempo. O problema é menor no século XVII. O século XVIII é o que possui mais textos. Poucos textos do século XIX foram incluídos, pois o córpus contém documentos apenas até 1808.

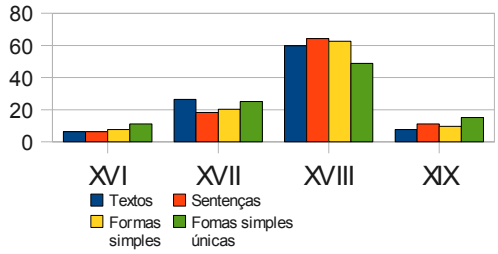


Figura 1: Distribuição do córpus por séculos

Em relação aos glossários, o glossário de junções totalizou 10.369 junções; o glossário F totalizou 16.067 abreviaturas (18% delas ocorreram no córpus); e o glossário C contém 7.045 abreviaturas. Observa-se no glossário F um predomínio maior de abreviaturas com sobrescrito. Devido ao levantamento automático de abreviaturas no glossário F, existem erros presentes (por exemplo, alguns números em romanos foram considerados como abreviaturas). Os glossários C e F apresentaram 2.473 abreviaturas em comum. Como F contém muitas abreviaturas que não estão C, é possível concluir que o glossário de [6] é bem abrangente. Entretanto, como C

também contém muitas abreviaturas que não estão em F, outra conclusão é que o glossário de [6] poderia ser melhorado com o uso de heurísticas como as apresentadas aqui.

O glossário de variantes foi construído com 18.082 palavras com variações de grafias ou agrupamentos, num total de 41.170 variações através de regras de transformação. As técnicas são complementares, possibilitando um aumento no número de abreviaturas levantadas automaticamente. Um comparativo entre as duas técnicas foi feito através das medidas precisão e cobertura comparativa. A cobertura comparativa é uma alternativa quando a cobertura não é conhecida. Para o cálculo dessa medida, as abreviaturas levantadas pelas duas técnicas são unidas em um único conjunto (as repetições são eliminadas), e então verifica-se a proporção de abreviaturas levantadas por cada técnica.

O experimento realizado no córpus consistiu em analisar variantes de grafia geradas pelas duas técnicas. Para tal, foram escolhidas 23 palavras no glossário de variantes. Como a análise do experimento pode ser longa, optou-se por um número pequeno de palavras na amostra utilizada. A Tabela 3 mostra as médias das precisões e das coberturas comparativas para as 23 palavras. Verdadeiros positivos representam variantes autênticas e falsos positivos são erros devido processo automático de levantamento de variantes. É possível observar uma alta precisão da técnica de regras de transformação e uma alta cobertura comparativa da técnica de distância de edição.

Tabela 3: Precisão e cobertura comparativa

| **Técnica** | **Verdadeiros positivos** | **Falsos positivos** | **Precisão** | **Cobertura comparativa** |
|---|---|---|---|---|
| Regras de transformação | 36 | 0 | 100% | 72% |
| Distância de edição | 41 | 196 | 21% | 84% |

O glossário de variantes de grafia também passou por uma avaliação com os usuários, na qual constatou-se que ainda existem uma parcela considerável de variantes de grafia não detectadas pelo método de regras de transformação. Novas regras serão criadas para aumentar a cobertura comparativa desse método. A Seção 6 traz a proposta de um ambiente para processamento de córpus históricos a partir das técnicas propostas neste trabalho.

# 6 O Ambiente Proposto

Um ambiente para processamento de córpus históricos foi concebido a partir das experiências obtidas no projeto DHPB. Um protótipo do ambiente utiliza as ferramentas descritas nas Seções 4 e 5. É dado um enfoque a atividades lexicográficas, mas o ambiente pode ser adaptado para atender às necessidades com objetivos variados. O ambiente é constituído por módulos, agrupados em duas arquiteturas: arquitetura para processamento de córpus e criação de glossários e arquitetura para acesso a córpus, glossários e redação de verbetes. A arquitetura para

compilação do córpus e criação de glossários é composta por seis módulos: (a) limpeza e anotação; (b) detecção de erros; (c) extração de abreviaturas; (d) extração de metadados, (e) geração de versões e (f) extração de variações de grafia. A arquitetura para acesso a córpus é baseada em ambiente *Web* e tem como vantagem a centralização de dados, característica de sistemas do tipo cliente-servidor. Essa arquitetura é dividida em 3 módulos: (a) acesso ao córpus; (b) acessos a glossários e (c) redação de verbetes. A vantagem da arquitetura modular é a possibilidade de substituir ou modificar módulos de acordo com as necessidades de diferentes projetos. Cada ferramenta pode ser responsável por um módulo. Mais detalhes do ambiente podem ser encontrados em [5].

## 7 Conclusões

Este trabalho 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. As contribuições deste trabalho englobam a metodologia, os recursos (os glossários e o córpus como um todo), as ferramentas desenvolvidas para processá-los, a ferramenta desenvolvida para redação de verbetes do dicionário, objetivo principal do projeto DHPB. No melhor do nosso conhecimento, este projeto de mestrado é inovador, sendo o único a disponibilizar um ambiente computacional com todas as funcionalidades aqui apresentadas para o tratamento de córpus históricos. 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. As contribuições foram disponibilizadas publicamente para uso em outros projetos (com exceção de material protegido por direitos autorais). 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. As contribuições apresentadas neste trabalho só foram possíveis graças à ajuda de inúmeros participantes do projeto DHPB. Trabalhos futuros incluem: melhorias no ambiente proposto para identificação de mais variantes de grafia; levantamento de mais abreviaturas; testes mais profundos com os usuários utilizando técnicas de IHC (Interação Humano Computador), melhorias para a conversão da ficha catalográfica para o formato TEI; e extração automática de metadados do córpus através de técnicas de aprendizado de máquina.

## Referências


1 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.

2 ALUISIO, S., PINHEIRO, G.M., MANFRIM, A.M.P, OLIVEIRA, L. H. M. de, L. C. GENOVES Jr., TAGNIN, S. E. O. The Lácio-Web: Corpora and Tools to advance Brazilian Portuguese Language Investigations and Computational Linguistic Tools. In LREC 2004. *Proceedings of LREC, 2004*, Lisboa, Portugal, p. 1779-1782.

3 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: *Digital Humanities*, 2006, Paris: Sorbonne, 2006. p. 3-5.
4 ATKINS, S.; CLEAR, J.; OSTLER, N. Corpus design criteria. *Journal of Literary and Linguistic Computing*, v. 7, n. 1, 1992.
5 CANDIDO, A. C. Criação de um ambiente para o processamento de córpus de Português Histórico. Dissertação (Mestrado) - Instituto de Ciências Matemáticas e de Computação,, USP, São Carlos, 2008.
6 FLEXOR, M. H. O. *Abreviaturas: Manuscritos dos séculos XVI ao XIX*. 2. ed. [S.l.]: UNESP, 1991. 468 p.
7 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.
8 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.
9 HIROHASHI, A. S. *Aprendizado de regras de substituição para normatização de textos históricos. Dissertação* (Mestrado) - Instituto de Matemática e Estatística, USP, São Paulo, 2004.
10 MCENERY, T.; WILSON, *A. Corpus Linguistics*. Edinburgh: Edinburgh University Press, 1996.
11 MENEGATTI, T. A. *Regras Lingüísticas para Tratamento Computacional da Variação de Grafia e Abreviaturas do Corpus Tycho Brahe*.Campinas: UNICAMP, 2002.Relatório Técnico.
12 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 e de Computação,, USP, São Carlos, 2004.
13 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.
14 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.
15 RAYSON, P. E. *Matrix*: A statistical method and software tool for linguistic analysis through corpus comparison. Tese (Doutorado) – Lancaster University, September 2002.
16 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.
17 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.
18 SANTOS, D.; RANCHHOD, E. Ambientes de processamento de corpora em português: comparação entre dois sistemas. In: *PROPOR'99*. [S.l.]: Evora, 2002.
19 SARDINHA, T. B. *Lingüística de Corpus*. Barueri, SP: Manole, 2004.
20 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.
21 TEI CONSORTIUM. The TEI Guidelines. Text Encoding Initiative Consortium, 2006. Disponível em: <http://www.tei-c.org/Guidelines2/>. Acesso em: 16 set. 2006.
22 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.
23 UNIVERSITÉ DE GENÈVE. *The ISO 9126 Standard*. 2006. Disponível <http://www.issco.unige.ch/ewg95/node1.html>. Acesso em: 14 nov. 2006.
24 VALE, O. A. ; CANDIDO JR, A. ; MUNIZ, M. C. M.; BENGTSON, C. G. ; CUCATTO, L. A.; ALMEIDA, G. M. B.; BATISTA, A. PARREIRA, M. C.; BIDERMAN, M. T. ; ALUÍSIO, S. M. Building a large dictionary of abbreviations for named entity recognition in Portuguese historical corpora. In: *LATECH* 2008. Paris: ELRA, 2008. v. 1. p. 1-10.