Procorph: um Sistema de Apoio à Criação de Dicionários Históricos

Arnaldo[1] Candido Junior
NILC – ICMC – USP
Av. Trabalhador São-carlense, 400
CEP: 13560-970 - São Carlos - SP
55 16 3373-9628
arnaldocan at ig dot com dot br

Sandra Maria Aluísio
NILC – ICMC – USP
Av. Trabalhador São-carlense, 400
CEP: 13560-970 - São Carlos - SP
55 16 3373-9628
sandra at icmc dot usp dot br

ABSTRACT
Dictionaries are important sources of study and knowledge about languages. Historical Dictionaries allow to study language evolution, and also to support historical documents understanding. This study presents a dictionary writing system called Procorph, developed to support the creation of Historical Portuguese dictionaries entries. This tool also can be adapted to work with contemporaries dictionaries. The system stores several information about the entries, such as definitions, text samples, spelling variations, sub-entries, and others.

RESUMO
Dicionários Históricos são importantes fontes de estudo e compreensão das línguas, pois possibilitam estudar a sua evolução, além de servirem de apoio para a leitura de documentos históricos. Este trabalho apresenta o sistema Procorph, desenvolvido para auxiliar a redação de verbetes de dicionários de Português Histórico e que pode ser adaptado para trabalhar com dicionários contemporâneos. O sistema armazena diversas informações sobre os verbetes como acepções, abonações, variantes de grafia, sub-entradas (sub-verbetes), entre outros.

Categories and Subject Descriptors
I.7 [**Document and Text Processing**]: Document and Text Editing; Document management; Document Preparation; Electronic Publishing.

J.5 [**Arts and Humanities**]: Linguistics.

General Terms
Design, Standardization, Human Factors.

Keywords
Historical dictionaries, dictionary writing systems, support systems for lexicographers.

1. INTRODUÇÃO
Dicionários construídos através do trabalho lexicográfico são uma importante fonte de informação para o estudo e a compreensão das línguas. Entretanto, o custo para construir um dicionário é alto devido ao tempo e aos recursos necessários. A redação de verbetes pode ser considerada a etapa principal do trabalho e, conseqüentemente, a de maior duração. Nesse contexto, ferramentas computacionais são úteis para apoiar a tarefa, automatizando atividades simples e repetitivas como a formatação de entradas do dicionário ou a geração de diferentes versões dos verbetes. Além disso, com o uso de ferramentas computacionais adequadas é possível organizar os verbetes em bases de dados, o que permite realizar buscas eficientes além de simplificar o compartilhamento de informações entre os redatores, uma vez que a base de dados pode ser compartilhada via *Web*. Apesar do potencial das ferramentas computacionais para apoio à tarefa lexicográfica, o número de ferramentas disponíveis é pequeno, além de não suprirem adequadamente necessidades de projetos voltados para Língua Portuguesa. No caso de dicionários históricos, o problema é agravado, pois estes possuem necessidades não compartilhadas por dicionários contemporâneos, como o tratamento de abreviaturas, de variantes de grafia e de símbolos que caíram em desuso [4].

Este trabalho apresenta o sistema Procorph (PROcessador de CÓrpus de Português Histórico), desenvolvido para apoiar a criação de dicionários de Português Histórico. O nome do sistema está relacionado ao processamento de córpus, pois esta é uma das tarefas para as quais foi concebido (além da confecção do dicionário). Este trabalho é o único na área voltado para o apoio à construção de dicionários históricos em Português, segundo os conhecimentos dos autores.

O sistema foi desenvolvido no decorrer do projeto DHPB (Dicionário Histórico do Português do Brasil), que visa à construção de um dicionário de Português Histórico a partir de textos escritos no Brasil entre 1500 e 1808 (Seção 2). A Seção 3 apresenta sistemas desenvolvidos para apoiar a criação de dicionários. A Seção 4 apresenta as características e os recursos do sistema Procorph e a Seção 5 traz as conclusões do trabalho.

2. O PROJETO DHPB
O projeto 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 com base em documentos históricos escritos no período pré-imprensa do Brasil, ou seja, entre os séculos XVI a XIX (até 1808). O projeto tem duração de

1 Bolsista CNPq

três anos, com início dos trabalhos em 2006 e término em 2008, e conta com 41 pesquisadores pertencentes a 11 universidades. O objetivo é a criação de um dicionário histórico com 10 mil entradas, sendo este 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, talvez devido a diferenças culturais entre os dois países e a existência de línguas indígenas no Brasil (com seus termos próprios para descrever a fauna e a flora brasileira).

2.1. Projeto e Compilação do Córpus do DHPB

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. As decisões de projeto do córpus estão diretamente relacionadas com os objetivos de pesquisa que o córpus deve atender. É possível definir três subetapas para a **etapa de compilação** do córpus: obtenção permissão de uso de textos protegidos por direitos autorais, coleta dos textos e limpeza. A obtenção de permissão de uso trata-se de uma etapa não técnica e geralmente trabalhosa, dado que um córpus pode ser constituído por textos de diversos autores. A limpeza envolve o tratamento de dados pessoais (se existirem), de metadados e de formatação. Durante a **etapa de anotação**, etiquetas são inseridas nos textos para preservar diferentes metadados. Para que o córpus seja reusado em diversas de pesquisas e processado por diferentes ferramentas computacionais é desejável que este 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. Diversos padrões de anotação internacionais foram propostos para a anotação de córpus, como o TEI[2] e o XCES[3]. Durante a **etapa de uso**, o córpus pode atender a profissionais com diferentes perfis, entre eles: especialistas em linguagem (por exemplo, lexicógrafos e terminólogos), especialistas em conteúdo (por exemplo, historiadores e críticos literários) e especialistas em mídia (por exemplo, profissionais que atuam no processamento automático de textos). Além disso, a pesquisa envolvendo córpus pode ser aplicada a diversas áreas.

Os textos utilizados para compilar o córpus histórico no qual o dicionário é baseado são de autores brasileiros ou portugueses que viveram um grande período de tempo no Brasil. Exemplos de textos incluem cartas dos missionários jesuítas, documentos dos bandeirantes, relatos dos sertanistas, documentos da inquisição católica, inventários e testamentos, entre outros.

Para a coleta de textos foram utilizados documentos impressos, manuscritos e arquivos em formato PDF (*Portable Document Format*) contendo imagens. Os PDFs foram digitalizados a partir de documentos inacessíveis, provenientes de acervos únicos. As imagens contidas nos PDFs foram extraídas e os textos impressos foram digitalizados. As imagens resultantes foram convertidas em texto com formatação através do processo de Reconhecimento Óptico de Caracteres (*Optical Character Recognition*). Manuscritos foram analisados e transcritos para o formato eletrônico através de digitação. A seguir, os textos passaram por um pré-processamento semi-automático para que pudessem ser trabalhados com ferramentas de processamento de córpus. As sub-etapas envolvidas no pré-processamento incluem (a) remoção de formatação, (b) limpeza de metadados, (c) conversão para formato de texto puro, e (d) inserção de anotação em formato TEI-Lite [11]. A seguir, foram geradas versões do córpus para uso com as ferramentas *Unitex* [8] e *Philologic* [15], pois cada uma delas apresentou algumas vantagens exclusivas. As ferramentas, por sua vez, geram as concordâncias utilizadas como base para a redação de verbetes. 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 córpus foi totalmente compilado, contando com 2.458 textos e 7.5 milhões de formas simples[4]. O córpus do projeto DHPB não será publicamente disponibilizado inicialmente, pois é necessária a obtenção de autorização de distribuição dos textos das editoras que os publicaram (embora haja textos que já caíram em domínio público, a maioria deles vêm de edições atuais, com direitos autorais ainda protegidos). Mais informação sobre o córpus do projeto pode ser encontrada em [12].

Em [9] foram levantados alguns problemas comuns em textos históricos, também verificados no projeto DHPB, entre eles: ausência de hifenização, junções de palavras (por exemplo, "éamor"), símbolos tipográficos incomuns (por exemplo, "ſ" em "deſcobrio") para a grafia de palavras e a alta freqüência de abreviaturas. Estes problemas devem ser adequadamente tratados durante a construção de dicionários históricos. O tratamento de símbolos tipográficos que caíram em desuso pode ser feito com o uso de etiquetas para denotá-los ou através da escolha de uma codificação de caracteres apropriada. A codificação *Unicode* é recomendada devido ao grande número de símbolos que permite representar. O uso pervasivo de abreviaturas é comum em manuscritos e também nos primeiros materiais impressos e sua presença pode dificultar o entendimento dos textos. Além disso, também é possível encontrar nos textos variações na grafia de uma dada palavra, que ocorrem pois os textos foram escritos em uma época em que não havia um sistema ortográfico unificado para o Português do Brasil. Variações de grafia chegam a ocorrer dentro de um único texto, tratando-se de um fenômeno particularmente freqüente no século XVI, dificultando a criação de concordâncias no córpus para a redação de verbetes. Algumas práticas comuns em textos em Português anteriores ao século XVIII, levantadas por [7], são: consoantes dobradas, inconsistência no uso de acentuação e troca entre vogais.

3. SISTEMAS PARA CRIAÇÃO DE DICIONÁRIOS

Em [13] são descritos 31 sistemas para criação de dicionários e apoio a tarefas lexicográficas e terminológicas. A maioria dos sistemas são voltados para a área terminológica, mas parte deles permite a criação de dicionários de língua geral e dicionários multilíngües. Os sistemas são, em geral, escritos para o idioma inglês, de forma que podem ter limitações se usados para a criação de dicionários em Português. Em [6] foi constatado 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

2 http://www.tei-c.org/index.xml

3 http://www.xces.org/

4 Formas simples são compostas de letras pertencentes ao alfabeto do Português Histórico, criado no escopo do projeto DHPB.

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.

É desejável que as ferramentas sejam capazes de gerenciar bases de dados, pois a velocidade de acesso a essas bases 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* [1]. O sistema é dividido em módulos e conta com o módulo *Browser/Refiner*, responsável pelo gerenciamento de bases de dados terminológicas. A ferramenta Corplex [10] é focada no gerenciamento de verbetes e possui recursos para apoiar a lexicográfica corporativa (desenvolvida em empresas e organizações). Entre os recursos oferecidos pelo sistema, destacam-se as suas buscas. Para o Português, existe o Corpógrafo que permite a criação e o processamento de córpus, a extração de terminologia e o gerenciamento de base de dados terminológicas com relações semânticas e ontológicas e, atualmente, está sendo desenvolvido o ambiente e-termos [2], uma plataforma *web* colaborativa para criação de produtos terminológicos.

4. O SISTEMA PROCORPH

O sistema Procorph foi desenvolvido para tratar necessidades do projeto DHPB, como forma de melhorar o processo de redação de verbetes e centralizar a criação do dicionário já que, atualmente, os verbetes do projeto são redigidos no editor de textos *Microsoft Word*. O sistema ainda não foi implantado no projeto DHPB, embora tenha sido apresentado em uma reunião do projeto DHPB para os redatores de verbetes, e posteriormente testado por 4 dos 21 redatores. Segundo a avaliação dos usuários que testaram a ferramenta, o sistema apresentou um ótimo desempenho. Mas considera-se que um teste incluindo todos os lexicógrafos tem um potencial maior para indicar limitações e apontar melhorias possíveis de serem implementadas.

A construção da ferramenta foi motivada por algumas dificuldades encontradas por lexicógrafos durante a redação de verbetes, entre elas, problemas de formatação, ausência de um sistema para simplificar referências sobre textos abonados e ausência de um sistema para centralizar os verbetes redigidos simultaneamente por diferentes lexicógrafos. No caso de dicionários históricos, dificuldades extras podem ser encontradas como por exemplo a busca por variantes de grafia das entradas e o gerenciamento das datações das abonações. Além de simplificar tarefas realizadas pelos lexicógrafos, o sistema pode também ser utilizado por consulentes em geral. Em [3] é destacado o consenso na área de terminologia e lexicografia computacional sobre o fato de dicionários utilizáveis por máquina serem muito mais eficientes do que dicionários impressos.

Um dos objetivos durante o desenvolvimento da ferramenta foi torná-la capaz de tratar bases de dados históricas em geral, de forma que possa vir a ser utilizada em outros projetos para construção de dicionários históricos, com pequenas adaptações. Além disso, também é possível modificar a ferramenta para a criação de sistemas voltados para dicionários contemporâneos, já que o Procorph é um software livre, disponibilizado sobre a licença GPL[5] (*General Public License*). O programa e seu código fonte estão publicamente disponíveis[6], sem custos adicionais e modificações são permitidas livremente. O sistema possui uma interface *Web*, foi desenvolvido em linguagem PHP (*Hipertext PreProcessor*), usando o banco de dados *Mysql*. O uso de *Javascript* na interface permitiu tornar a edição de verbetes mais dinâmica e simples de usar. A vantagem da criação de um sistema *Web* é a centralização dos dados e a opção de compartilhamento de verbetes entre os redatores.

As duas principais telas do sistema permitem a consulta e a edição de verbetes, respectivamente. As informações armazenadas na base de dados incluem a classe gramatical do verbete, informações de gênero e flexão, as diferentes acepções (ou definições) da entrada, verbetes relacionados, observações e sub-entradas. Cada acepção é acompanhada por uma abonação (excerto de um texto no córpus no qual a entrada é um exemplo da acepção em questão), além de uma referência ao texto do córpus do qual a acepção foi retirada. A referência compreende a página em que o excerto ocorre e o código de texto. Através do código é possível obter o título, o ano de publicação e o autor do texto, gerando referências em um formato semelhante ao formato ABNT (Associação Brasileira de Normas Técnicas). Outras telas do sistema incluem a tela de listagem de textos usados durante a abonação, a tela de buscas por variantes de grafia e a tela de controle de usuários (apenas para usuários com poderes administrativos).

Além das informações comuns em dicionários contemporâneos, o sistema também permite a inserção de variantes de grafia. Abreviaturas podem ser utilizadas em conjunto com as variantes de grafia se o redator de verbete desejar. Outro recurso específico para dicionários históricos é o controle da primeira datação dos verbetes. A primeira datação da entrada no córpus fornece uma informação útil para estimar período aproximado do início de uso da palavra na Língua Portuguesa. Cada entrada pode ser acompanhada de sub-entradas, que são verbetes completos (contendo os mesmos atributos das entradas) associados a um verbete principal e geralmente consistem de lexias complexas. Por exemplo, para o verbete "mulher", as sub-entradas incluem "mulher do reino", "mulher ama", "mulher moça" e "mulher da terra". A Figura 1 mostra a tela de edição de verbetes para o verbete "prezuízo" (por questões de espaço, acepções e variantes da entrada foram removidas e o tamanho da figura foi reduzido).

Os verbetes são armazenados em codificação *Unicode*, pois seu uso permite representar todos os símbolos encontrados nos textos históricos coletados. Entretanto, não é possível digitar parte desses símbolos em teclados brasileiros. Uma possível solução é o uso de programas como o Mapa de Caracteres, disponível no *Microsoft Windows*. Entretanto, essa solução é pouco prática devido à dificuldade para localizar os caracteres desejados. A solução utilizada no sistema Procorph envolve o uso de conjuntos de caracteres para denotar símbolos *Unicode* de difícil digitação. A vantagem reside na facilidade em converter os conjuntos de caracteres em seus respectivos símbolos automaticamente no sistema. A Tabela 1 mostra cadeias de caracteres e seus respectivos símbolos denotados.

5 http://www.gnu.org/licenses/gpl.txt

6 http://nilc.icmc.usp.br/nilc/projects/procorph/

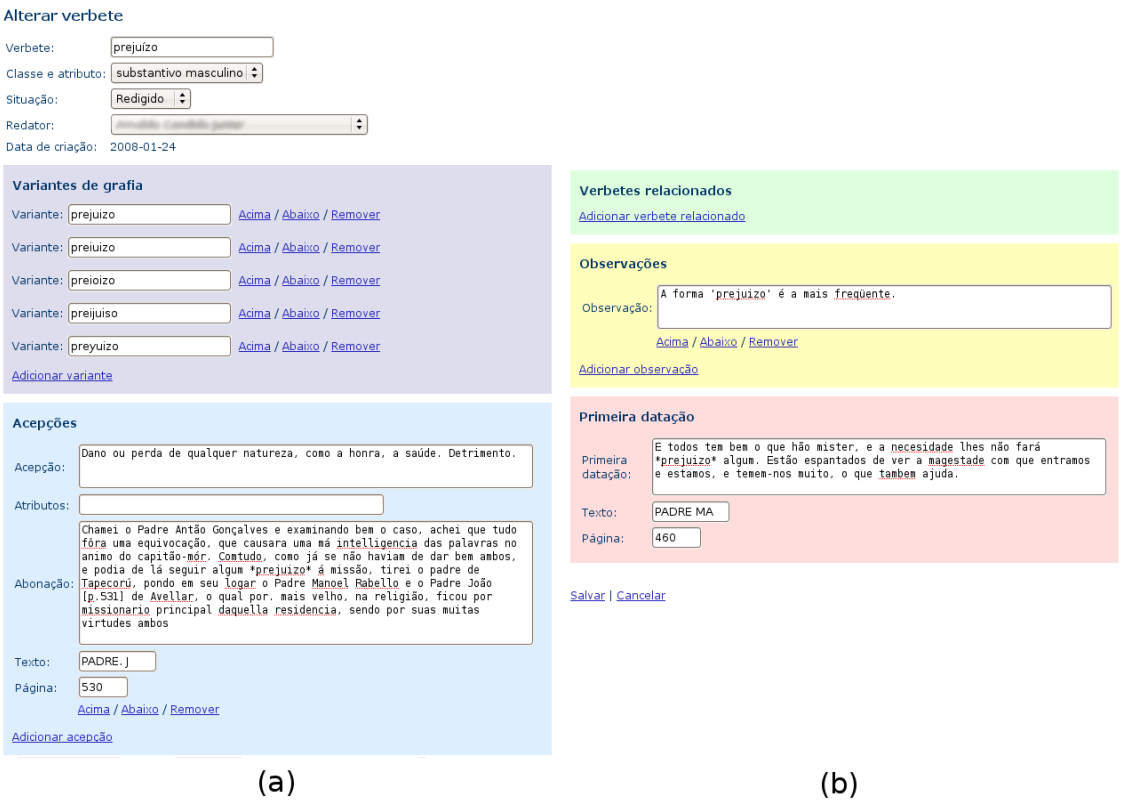


Figura 1: Tela de edição de verbetes

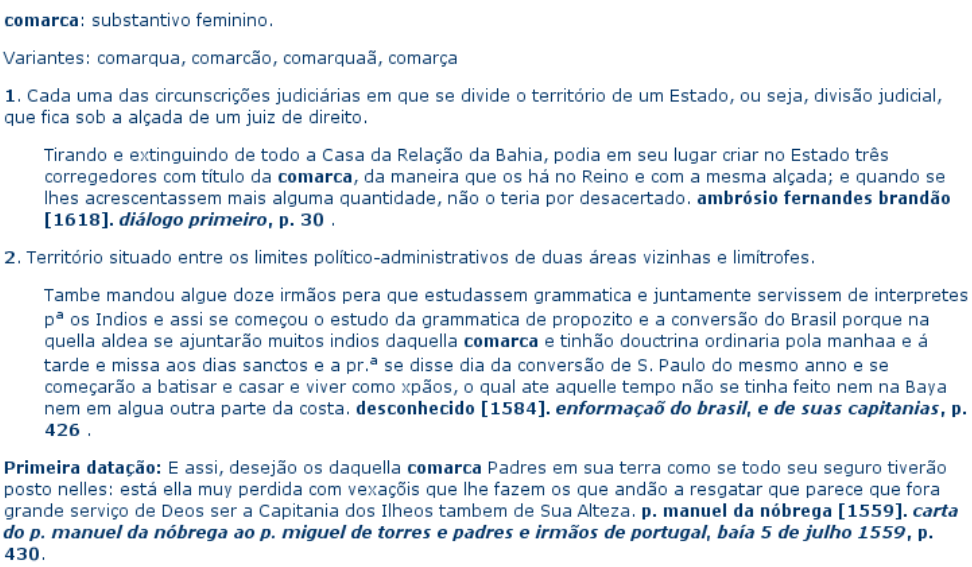
Tabela 1: Conversão de cadeias para *Unicode*

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

Durante a redação de verbetes são levantadas as diferentes variações de grafia da entrada em questão para permitir que as abonações mais relevantes para o dicionário sejam selecionadas e para informar o consulente do dicionário sobre as diferentes grafias que este pode encontrar quando consultar textos históricos. O número de variações pode ser grande (principalmente no século XVI), como, por exemplo para a entrada "prejuízo", que possuí 10 variantes conhecidas (prejuizo, preiuizo, preioizo, preijuiso, preyuizo, preyoizo, prejoizo, prejuiso, perjuizo, prejuiſo). Entretanto, é difícil levantar manualmente as variações de grafia no córpus. Para mitigar esse problema, o sistema Procorph disponibiliza um glossário de variantes de grafia detectadas automaticamente. Como a criação do glossário foi feita automaticamente e pode conter erros, as variantes não são inseridas automaticamente durante a redação de verbetes, pois precisam passar pela análise dos redatores. Além disso, o processo de levantamento automático de variantes não é capaz de detectar todas as possíveis variantes para uma determinada entrada.

A construção do glossário foi feita a partir do uso de Regras de Transformação para agrupamento de palavras com o uso com o uso do sistema Siaconf (Sistema de Apoio à Contagem de Freqüência em Córpus), também disponibilizado publicamente [5]. O glossário construído contém 18.082 agrupamentos de palavras, em um total de 41.170 variantes de grafia. Exemplos de agrupamentos contidos no glossário incluem as palavras "vila" com 5 variantes (villa, vyla, vjlla, vylla e vjla) e "não" com 3 variantes (naõ, nam e nao). O glossário é distribuído juntamente com o sistema. No caso do projeto DHPB, os redatores também usaram o sistema *Philologic* para recuperar variações de grafia automaticamente. O *Philologic* utiliza algoritmos de distância de edição fornecidos pelo utilitário de sistema *Agrep*.

Durante o decorrer do projeto DHPB a formatação de verbetes foi realizada pelos redatores. Entretanto, a tarefa é demorada e pode causar um impacto negativo na produtividade do dicionário. Esse problema é solucionado com o uso do Procorph, uma vez que o sistema formata as entradas automaticamente. Outra vantagem é o baixo custo para aplicar mudanças de formatação a todas as entradas simultaneamente, bem como a possibilidade de gerar



comarca: substantivo feminino.

Variantes: comarqua, comarcão, comarquaã, comarça

1. Cada uma das circunscrições judiciárias em que se divide o território de um Estado, ou seja, divisão judicial, que fica sob a alçada de um juiz de direito.

Tirando e extinguindo de todo a Casa da Relação da Bahia, podia em seu lugar criar no Estado três corregedores com título da **comarca**, da maneira que os há no Reino e com a mesma alçada; e quando se lhes acrescentassem mais alguma quantidade, não o teria por desacertado. **ambrósio fernandes brandão [1618].** ***diálogo primeiro*****, p. 30** .

2. Território situado entre os limites político-administrativos de duas áreas vizinhas e limítrofes.

Tambe mandou algue doze irmãos pera que estudassem grammatica e juntamente servissem de interpretes pª os Indios e assi se começou o estudo da grammatica de propozito e a conversão do Brasil porque na quella aldea se ajuntarão muitos indios daquella **comarca** e tinhão douctrina ordinaria pola manhaa e á tarde e missa aos dias sanctos e a pr.ª se disse dia da conversão de S. Paulo do mesmo anno e se começarão a batisar e casar e viver como xpãos, o qual ate aquelle tempo não se tinha feito nem na Baya nem em algua outra parte da costa. **desconhecido [1584].** ***informaçaõ do brasil, e de suas capitanias*****, p. 426** .

Primeira datação: E assi, desejão os daquella **comarca** Padres em sua terra como se todo seu seguro tiverão posto nelles: está ella muy perdida com vexaçõis que lhe fazem os que andão a resgatar que parece que fora grande serviço de Deos ser a Capitania dos Ilheos tambem de Sua Alteza. **p. manuel da nóbrega [1559].** ***carta do p. manuel da nóbrega ao p. miguel de torres e padres e irmãos de portugal, baía 5 de julho 1559*****, p. 430**.

Figura 2: verbete criado a partir do sistema Procorph

diferentes versões do dicionário se os coordenadores do projeto assim desejarem. É possível, por exemplo, modificar o sistema para gerar uma versão integral e uma versão resumida, na qual as abonações são removidas por questões de espaço. Adicionalmente, a ferramenta oferece uma opção para converter o verbete automaticamente para o formato do *Microsoft Word*, já que este tem sido usado ostensivamente no projeto DHPB. A geração do documento consiste em uma técnica simples na qual um arquivo HTML criado seguindo determinados padrões é gerado com a extensão DOC, e posteriormente aberto no *Word*. A Figura 2 contém o exemplo de um verbete após a formatação pelo sistema (as referências bibliográficas são inseridas automaticamente).

Como os verbetes são disponibilizados através da *Web*, faz-se necessário um controle de usuários com acesso à base e com privilégios para redigir verbetes. No sistema Procorph, quatro níveis de acesso ao sistema são permitidos: consulente, redator, revisor e administrador. Consulentes podem apenas navegar pela base e consultar verbetes, textos e variantes. Redatores possuem permissão para criar verbetes e modificar seus próprios verbetes. Revisores têm acesso completo à base de verbetes, e podem alterar verbetes redigidos por qualquer usuário. Administradores têm os poderes de revisores e também controlam os usuários cadastrados na base.

5. CONCLUSÕES

Neste trabalho foi apresentado o Procorph, um sistema *Web* criado no contexto do projeto DHPB para auxiliar a redação dos verbetes do dicionário de Português Histórico. O sistema é distribuído livremente e pode ser adaptado a outros projetos relacionados à criação de dicionários históricos. Com algumas modificações, pode vir a atender a projetos de dicionários contemporâneos. Entre os recursos oferecidos encontram-se um editor de verbetes, capaz de tratar variantes de grafia, abreviaturas, gerenciar símbolos *Unicode* e referenciar documentos históricos, preservando informações sobre datação e períodos de uso da entrada. Um glossário de variantes de grafia gerado automaticamente é incluído para apoiar a redação dos verbetes. Os verbetes são formatados automaticamente e podem, opcionalmente, serem convertidos para uso com o editor *Microsoft Word*.

Trabalhos futuros incluem o uso da ISO 9126 [14] para avaliação intrínseca dos recursos oferecidos pelo sistema. A ISO 9126 foi criada para avaliar a qualidade de software sob diferentes critérios a partir de 6 métricas (e sub-métricas) definidas para a tarefa. Além disso, são deixados como trabalhos futuros novos recursos que estendam a funcionalidade do sistema, permitindo o processamento de córpus históricos. Para tal, é proposto um módulo de geração de concordâncias e um módulo para contagem de freqüências. O concordanceador utilizado pode vir a ser o do *Philologic*, já que este também possui interface *Web*. Com os novos módulos, o sistema passa a exercer funcionalidades do *Unitex* e *Philologic*, tornando-se mais abrangente e centralizando o acesso ao córpus e ao dicionário histórico.

6. REFERÊNCIAS

[1] 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.

[2] 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.

[3] CORREIA, M. Terminologia e lexicografia computacional. Disponível em: <http://www.realiter.net/spip.php?article787#nb1>. Acesso em 08 set. 2008.

[4] CANDIDO JR, A.; ALUÍSIO, S.M. Um Ambiente Computacional para o Processamento de Córpus de Português Histórico. To be published in the Proceedings of the *IV Workshop on MSc Dissertation and PhD Thesis in Artificial Intelligence (WTDIA 2008) & VI Best MSc Dissertation/PhD Thesis Contest (CTDIA 2008)*, p. 1-10.

[5] GIUSTI, R.; CANDIDO JR, A.; MUNIZ, M.; CUCATTO, L.; A. ALUÍSIO, S. 2007. “Automatic detection of spelling variation in historical corpus: An application to build a

Brazilian Portuguese spelling variants dictionary". In: *Proceedings of the Corpus Linguistics 2007 Conference*, Matthew Davies, Paul Rayson, Susan Hunston, Pernilla Danielsson (eds.).

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

[7] 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.

[8] PAUMIER, S. *Unitex 1.2*: User Manual. June 2006. Disponível em: <http://www-igm.univ-mlv.fr/~unitex/UnitexManual.pdf>. Acesso em: 25 jul. 2008.

[9] 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.

[10] Simonsen, Henrik K. CorpLex: Blueprints of a Corporate Dictionary and Editing System. In: *Studies in Contrastive Linguistics*, Santiago de Compostela, September 2005, Universidade de Santiago de Compostela, pp. 453-460.

[11] TEI CONSORTIUM. *The TEI Guidelines*. Text Encoding Initiative Consortium, 2008. Disponível em: <http://www.tei-c.org/Guidelines2/>. Acesso em: 25 jul. 2008.

[12] VALE, O.; CANDIDO JR. A.; MUNIZ, M. ;BENGTSON, C.; CUCATTO, L.; ALMEIDA, G.; BATISTA, A.; PARREIRA, M.C.; BIDERMAN, M.T. ALUÍSIO, S. Building a large dictionary of abbreviations for named entity recognition in Portuguese historical corpora. In the *proceedings of the LREC 2008 Workshop on Language Technology for Cultural Heritage Data* (LaTeCH 2008), p. 1-10.

[13] UNIVERSITÄT LEIPZIG. *Terminology Management*. 2008. Disponível em: <http://www.uni-leipzig.de/~xlatio/software/soft-termiman.htm>. Acesso em: 25 jul. 2008.

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

[15] UNIVERSITY OF CHICAGO. *PhiloLogic User Manual*. 2008. Disponível em: <http://philologic.uchicago.edu/manual>. Acesso em: 25 jul. 2008.