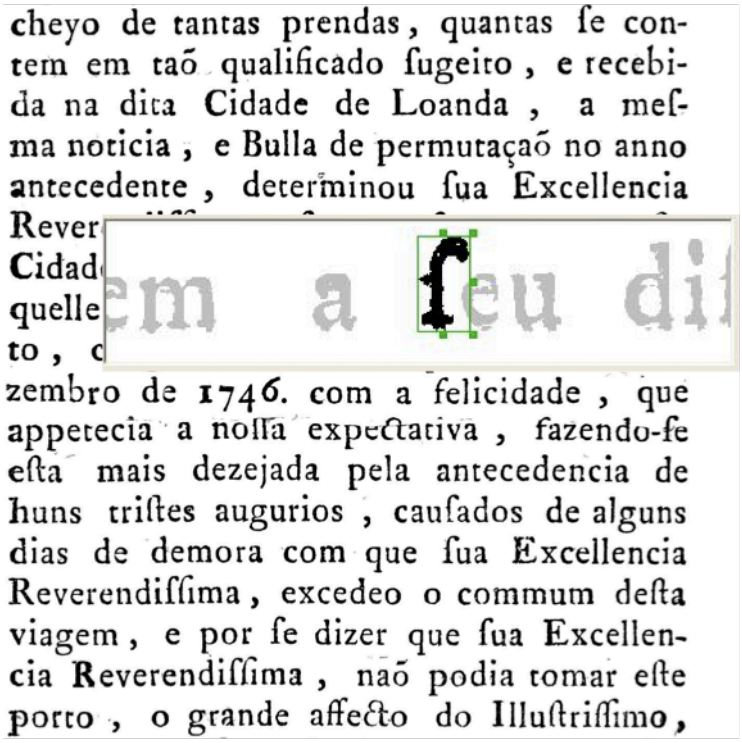
Universidade de São Paulo
Pró-Reitoria de Graduação
Programa Ensinar com Pesquisa

Edições Filológicas na Brasiliana Digital


cheyo de tantas prendas, quantas ſe con-
tem em taõ qualificado ſugeito, e recebi-
da na dita Cidade de Loanda, a meſ-
ma noticia, e Bulla de permutaçaõ no anno
antecedente, determinou ſua Excellencia
Rever
Cidad
quelle
to, c
zembro de 1746. com a felicidade, que
appetecia a noſſa expectativa, fazendo-ſe
eſta mais dezejada pela antecedencia de
huns triſtes augurios, cauſados de alguns
dias de demora com que ſua Excellencia
Reverendiſſima, excedeo o commum deſta
viagem, e por ſe dizer que ſua Excellen-
cia Reverendiſſima, naõ podia tomar eſte
porto, o grande affecto do Illuſtriſſimo,

Relatório de Pesquisa
2010

Universidade de São Paulo
Pró-Reitoria de Graduação
Programa Ensinar com Pesquisa

Projeto 1754

Edições Filológicas na Brasiliana Digital

Relatório de Pesquisa
2010

Coordenador
Maria Clara Paixão de Sousa
Departamento de Letras Clássicas e Vernáculas
Faculdade de Filosofia, Letras e Ciências Humanas
Universidade de São Paulo

Bolsistas
Fabiana Marcondes Ferraz
Jáderson Johnattan Porto
Leila Rosa de Oliveira
Márcia Aparecida Santos Mendes
Thelma Tavares Dias

Apresentação

Este documento é parte do Relatório Final do projeto Edições Filológicas na Brasiliana Digital, registrado sob o número 1754 no programa Ensinar com Pesquisa, da Pró-reitoria de Graduação da Universidade de São Paulo.

O Relatório completo é composto de duas partes:

(I) Relatório Científico

No presente documento, apresentam-se as considerações gerais sobre o andamento do projeto, elaboradas por sua coordenadora.

(II) Anexos: Documentação

Outros documentos relevantes para a avaliação do projeto estão apresentados sob forma de anexos eletrônicos - particularmente, os relatórios individuais de cada bolsista, os resultados do trabalho de pesquisa (i.e., os textos editados), e a documentação referente a reuniões e eventos científicos ligados ao projeto.

Os anexos eletrônicos encontram-se armazenados no seguinte endereço-web:

http://lampiao.brasiliana.usp.br/lingua/EnsinarComPesquisa2010Relatorio

I. Relatório Científico

1. Introdução

Esta proposta de pesquisa insere-se no contexto maior dos projetos *Brasiliana USP* e *Brasiliana Digital*. O *Projeto Brasiliana* é uma iniciativa da Reitoria da Universidade de São Paulo, com a missão de custodiar e desenvolver a *Biblioteca Brasiliana*, reunindo cerca de 500.000 volumes de inestimável valor histórico, fruto da união entre os acervos do *Instituto de Estudos Brasileiros*, órgão com a tradição de mais de 40 anos dedicados à curadoria de material histórico na USP, e da *Biblioteca Brasiliana Guita e José Mindlin,* fundada em janeiro de 2005[1] como abrigo da coleção reunida pelo bibliófilo José Mindlin e generosamente doado à USP em maio de 2006. Os acervos serão transferidos em conjunto para o edifício especialmente construído para este fim no coração da Universidade, previsto para ser terminado em 2012.


Maquete do Edifício Brasiliana USP 1

A esta nova condição física, favorecedora da consulta pública, soma-se a construção da *Biblioteca Brasiliana Digital*, em http://www.brasiliana.usp.br, que tem por objetivo ampliar o alcance do acervo para fins de pesquisa geral e acadêmica, sob os moldes propostos no projeto "*Por uma Biblioteca Brasiliana Digital*" (Puntoni, 2007).

Desde 2008, o Grupo de Pesquisas *Língua Brasiliana* veio somar-se às iniciativas da *Brasiliana* USP, investigando caminhos para revelar e preparar o potencial dos textos do Acervo como fonte da língua e sobre a língua no Brasil – uma área para a qual este material naturalmente se vocaciona. Temos colaborado com o grupo maior da Brasiliana USP em pesquisas experimentais iniciais junto ao *Laboratório Brasiliana*, inaugurado no início de 2009 junto ao canteiro de obras do edifício em construção. Nosso objetivo de longo prazo é desenvolver instrumentos para pesquisas lingüísticas com base no acervo, por meio da prospecção de materiais de interesse e do desenvolvimento e aplicação de métodos de preparação editorial e de instrumentação computacional para extração automática de informação dos textos mais antigos. Esperamos com isso fundar os alicerces para a exploração do acervo *Brasiliana USP* por estudiosos da história da língua, bem como formar as bases humanas, tecnológicas e materiais para o aproveitamento futuro do acervo pela comunidade de pesquisa. A documentação das atividades do grupo está disponível em http://lampiao.brasiliana.usp.br/lingua/.

O projeto Edições Filológicas na Brasiliana Digital, realizado sob os auspícios da Pró-reitoria de Graduação da Universidade, no âmbito do programa Ensinar com Pesquisa, representou a concretização inicial dos objetivos deste grupo, ao reunir alunos da graduação em letras em torno do desafio inicial de preparar edições filológicas em meio digital que permitam o tratamento computacional dos textos mais antigos do acervo para essas futuras pesquisas. Neste relatório, apresentam-se os principais resultados das pesquisas realizadas no projeto ao longo de 2010.

[1] cf. Resolução da Reitoria da Universidade de São Paulo Nº 5172, 23.12.2004. D.O.E, 24.12.2004; e <http://leginf.uspnet.usp.br/resol/r5172m.htm>

2. Metas e Resultados

Os objetivos deste projeto fundaram-se na meta de longo prazo do grupo *Língua Brasiliana:* revelar o potencial da Biblioteca como fonte de estudos lingüísticos, graças à criação de instrumentos apropriados de pesquisa, contribuindo assim para o conhecimento sobre a língua portuguesa e sobre a formação lingüística do Brasil; e dar início a um centro de formação de recursos acadêmicos (humanos e tecnológicos) para a exploração do acervo em três áreas de pesquisa: Filologia, Linguística Histórica e Linguística Computacional. Para o ano de 2010, apresentamos duas metas pontuais (cf. Projeto Original, http://lampiao.brasiliana.usp.br/lingua/sites/default/files/projeto_ensinar_com_pesquisa_mcpsousa_0.pdf)

Metas para o ano de 2010:

- Treinamento de sofwtare de reconhecimento de caracteres para tratamento de textos em português impressos nos séculos XVI e XVII;
- Produção de um glossário de variações ortográficas com base em textos em português impressos nos séculos XVI e XVII, para uso em softwares de reconhecimento de caracteres e em programações de buscas.

No projeto original apresentado ao programa Ensinar com Pesquisa, essas metas foram justificadas tendo em vista dois fatores principais: de um lado as características da construção de um acervo digitalizado de textos antigos como a Brasiliana Digital; de outro lado, o valor e o interesse desse tipo de acervo para o trabalho em linguística histórica. Lembramos, na ocasião, que a intensificação da circulação de textos antigos no meio digital representa um campo imensamente promissor para as pesquisas atuais sobre a história da língua, em particular para aquelas que vêm procurando aliar os saberes tradicionais da filologia e da crítica textual com os avanços recentes da linguística computacional e de corpus (como é o caso da linha que seguimos no grupo Língua Brasiliana, a partir de Paixão de Sousa 2005, 2006, 2007). Salientamos entretanto, na mesma ocasião, que a profusão de textos antigos em circulação no meio eletrônico traz também, para os estudos sobre a língua, alguns dos desafios mais importantes da filologia e da crítica textual em todos os tempos: a preocupação com a reprodução fidedigna dos originais. Esses desafios, no contexto digital, precisam ser enfrentados pela combinação entre o trabalho filológico e o conhecimento do funcionamento das tecnologias computacionais de processamento. Por conta disso, neste projeto tomamos o problema do reconhecimento e processamento de caracteres como questão central a ser trabalhada. A metodologia concebida para este enfrentamento, conforme exposta no projeto original, previa uma combinação entre o trabalho de desenvolvimento de um programa de reconhecimento automático e o trabalho de edição filológica dos textos - unidos, fundamentalmente, pelo objetivo de construir o produto principal do projeto: um glossário de variações ortográficas.

As metas e a metodologia assim estabelecidas e justificadas no projeto inicial foram atingidas neste primeiro ano de pesquisas, conforme se relata com maior detalhamento na sessão seguinte. Neste ponto do relatório, resumem-se de forma pontual os principais resultados obtidos. Observe-se que ao glossário de variações ortográficas inicialmente prometido veio somar-se um segundo glossário, composto pelos erros do programa de reconhecimento de caracteres, que resultou do trabalho de edição dos textos e que,

a nosso ver, pode também ser útil para futuras pesquisas.

Tabela 1 - Resumo dos resultados obtidos

Texto trabalhado	Número de palavras total	Itens com variação ortográfica catalogados	Erros de reconhecimento catalogados
(1) Cartas do Marquês..., 1642	1,213	606	305
(2) Tratado de tréguas..., 1642	3,454	1,902	1,133
(3) Sucesso de la guerra..., 1646	4,285	2,125	1,371
(4) Sucesso da armada...			
(5) Nova Lusitânia, 1675	21, 725	4,775	1,714
(6) Relação da entrada..., 1747	3,783	1,339	542
(7) Queda que as mulheres..., 1861	4,554	804	237
(Totais)	**39,014**	**11,551**	**5,302**

Os números totais dos nossos resultados, expostos na tabela acima, somam portanto um glossário de 11.551 itens no glossário de variações ortográficas e 5.302 itens no glossário de erros de reconhecimento, num universo total de 39.014 palavras catalogadas. Os procedimentos técnicos utilizados para a obtenção desses resultados, bem como a metodologia que fundou o trabalho de edicão dos textos, são detalhados na seção seguinte. Os textos integralmente editados e os glossários de variação ortográfica e de erros de reconhecimento obtidos em cada um deles estão disponíveis sob forma de anexos, no endereço http://lampiao.brasiliana.usp.br/lingua/EnsinarComPesquisa2010Relatorio. A opção pela apresentação desses resultados na forma de armazenamento em endereço web justifica-se, primordialmente, por conta do tamanho dos arquivos gerados, que sobrecarregariam o envio do documento de relatório no sistema.

3. Detalhamento

Nesta seção, descrevem-se as etapas da pesquisa que levaram aos resultados resumidos acima, buscando-se indicar, onde relevante, os principais desafios técnicos e metodológicos enfrentados neste processo. A apresentação se inicia com uma lista dos textos utilizados como corpus da pesquisa. Depois disso, segue-se a ordem cronológica das etapas conduzidas: primeiro, apresentamos o trabalho com o software de reconhecimento de caracteres, e em seguida, o trabalho de edição filológica conduzido com a ferramenta eletrônica E-Dictor. Na seção 4 mais adiante, procuramos apresentar um balanço dos resultados obtidos, e apontar as perspectivas que se colocam para o prosseguimento da pesquisa - essencialmente, no contexto da aprovação de um novo projeto no programa Ensinar com pesquisa, já em andamento desde fevereiro de 2011.

3.1 Corpus da pesquisa

Esta pesquisa foi realizada com base em um corpus de sete obras portuguesas impressas entre os séculos XVII e XIX, digitalizadas na Brasiliana Digital. A escolha das obras atendeu a critérios fundamentalmente técnicos: selecionamos textos impressos, representativos de sua época, sob o ponto de vista linguístico e material. A lista completa dessas obras está apresentada abaixo. Para cada texto, são listados os endereços eletrônicos dos resultados das edições. Nas seções que seguem, cada obra será referida mediante seu título abreviado.

1) Cartas do Marquês de Montalvão (1642)

Texto trabalhado por Márcia Aparecida Santos Mendes


CARTAS
QVE ESCREVEO
O MARQVEZ DE MONTALVAM SEN-
do Viſo Rey do Eſtado do Braſil, ao Conde de Naſ-
ſau, que gouernaua as armas em Pernambuco
dandolhe auiſo da felice acclamação de
ſua Mageſtade o Senhor Rey Dõ
Ioão o IV. neſtes ſeus Reynos
de Portugal, è reposta do
Conde de Naſſau.
COM OVTRA CARTA QVE O MARICHAL
ſeu filho trouxe para ſe apreſentar cõ ella a ſua Mageſtade.
EM LISBOA.
Com todas as licenças neceſſarias.
Na Officina de Domingos Lopez Roſa. Anno de 164?.

Ficha Catalográfica:
Autor: Montalvão, Jorge Mascarenhas, Marquês de, d. 1652
Título: Cartas que escreveo o marquez de Montalvam sendo Viso Rey do Estado do Brasil, ao Conde de Nassau, que governava as armas em Pernambuco dandolhe aviso de felice acclamação de sua Magestade [...]
Título: [...] o Senhor Rey Dõ João o IV nestes seus Reynos de Portugal, é reposta do Conde de Nassau. Com outra carta que o Marichal seu filho trouxe para apresentar cõ ella a sua Magestade
Local de Publicação: Lisboa : Officina de Domingos Lopez Rosa
Ano de Publicação: 1642
Descrição Física: 7 p. (numeradas A2, A3)
Idioma: Português
Resumo: Carta em que o Marquês de Montalvão escreve ao Conde Maurício de Nassau relatando a subida ao trono de D. João IV. Em seguida carta com a resposta do Conde Maurício de Nassau parabenizando a aclamação de Sua Majestade. Por último, cópia da carta, para que o Marechal, seu filho, se apresente à Sua Majestade.
URI: http://www.brasiliana.usp.br/bbd/handle/1918/01202700
Tipo: Carta

Arquivos Relacionados a este texto no Anexo Eletrônico:

- Texto resultante do reconhecimento automático (HTML)
- Texto com reconhecimento automático corrigido (HTML)
- Texto modernizado (HTML)
- Lista das correções de reconhecimento automático (HTML)
- Lista completa das edições realizadas (HTML)
- Antroponímia do texto (HTML)
- Toponímia do Texto (HTML)
- Texto de Base Anotado (TXT)

Tratado das tréguas ... (1642)

Texto trabalhado por Jáderson J. Porto


Treſlado do Latin na lin-
goa Portugeza.

Trattado das Tregoas eſuspenſaó de todo o acto de hoſtilidade ebem aſſi de navegaçaó, Comercio ejuntamente Socorro, feito, começado eaccabado em Haya de Hollande a xij. de Iunho 1641. por tempo de des annos entre o Senhor Triſtaó de Mendoça Furtado, *do Conſelho e Embaixador do Sereniſſimo epoderoſiſſimo* Dom Ioaõ IV *deſte nome Rey de Portugal e dos Algarves, Eos Senhores Deputados dos Muito poderoſos Senhores Eſtados Geraés das Provincias Vnidas dos Paizes Baixos.*

Em a HAYA,
Em caza da Viuva e Erdeiros de Ilebrandt Iacobſon van Wouw, Imprimidor Ordinario dos Muy altos e poderoſos Senores Eſtados Generais, Anno 1642. Cum Privilegio.

Ficha Catalográfica:

Título: Trattado das tregoas e suspensao do todo o acto de hostilidade e bem assi de navegaçao, commercio e juntamente socorro, feito começado e accabado em Haya de Hollande a Xij de iunha 1641 ...

Título: ... Por tempo de des annos entre o senhor Tristao de Mendoça Furtado, do conselho e exbaixador do Serenissimo e Poderosissimo dom Ioao IV. deste nome Rey de Portugal e dos Algarves, e os senhores deputados dos muito poderosos senhores estados geraes das privincias unidas dos paizes baixos. Treslado do Latin na lingoa Portugueza. Em a Haya. Em caza da Viuva e Erdeiros de Ilebrandt Iacobson van Wouw, Imprimidor Ordinario dos Muy altos e poderosos Senores Estados Generais

Local de Publicação: Haia : Em caza da Viuva e Erdeiros de Ilebrandt Iacobson van Wouw, Imprimidor Ordinario dos muy altos e poderosos Senores Estados Generais

Ano de Publicação: 1642

Descrição Física: 16 p.

Idioma: Português

Resumo: Tratado de trégua e suspensão de todo ato de hostilidade, assim como de navegação, comércio e socorro, assinado em Haia em 12 de junho 1641, com duração de dez anos, entre o embaixador português, Tristão de Mendoça Furtado, e os Estados Gerais das Províncias Unidas dos Países Baixos. Traduzido do original em Latim. Existe tradução em holandês, publicada em Haia em 1642.

Direitos: Domínio público

URI: http://www.brasiliana.usp.br/bbd/handle/1918/01936100

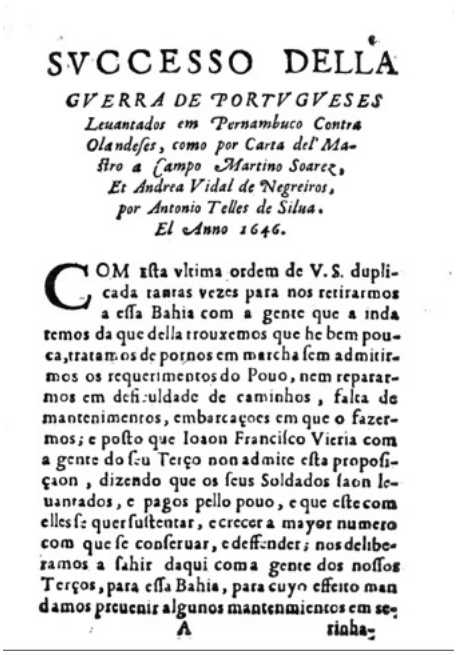
Tipo: Folheto

Arquivos Relacionados a este texto no Anexo Eletrônico:

- Texto resultante de reconhecimento automático (HTML)
- Texto com reconhecimento automático corrigido (HTML)
- Texto modernizado (HTML)
- Lista das correções de reconhecimento automático (HTML)
- Lista completa das edições realizadas (HTML)
- Arquivo de Base Anotado (XML)
- Arquivo de Base Anotado (TXT)

Sucesso de la guerra ... (1646)

Texto trabalhado por Fabiana M. Ferraz e Jaderson J. Porto



SVCCESSO DELLA
GVERRA DE PORTVGVESES
Leuantados em Pernambuco Contra
Olandeſes, como por Carta del' Ma-
ſtro a Campo Martino Soarez,
Et Andrea Vidal de Negreiros,
por Antonio Telles de Silua.
El Anno 1646.

COM eſta vltima ordem de V. S. dupli-
cada tantas vezes para nos retirarmos
a eſſa Bahia com a gente que a inda
temos da que della trouxemos que he bem pou-
ca, tratamos de pormos em marcha ſem admitir-
mos os requerimentos do Pouo, nem reparar-
mos em deficuldade de caminhos, falta de
mantenimentos, embarcaçoes em que o fazer-
mos; e poſto que Ioaon Francisco Vieria com
a gente do ſeu Terço non admite eſta propoſi-
çaon, dizendo que os ſeus Soldados ſaon le-
uantados, e pagos pello pouo, e que eſte com
elles ſe quer ſuſtentar, e crecer a mayor numero
com que ſe conſeruar, e deffender; nos delibe-
ramos a ſahir daqui coma gente dos noſſos
Terços, para eſſa Bahia, para cuyo effeito man
damos preuenir algunos mantenmientos em ſe-
A rinha-

Ficha Catalográfica:
Título: Sucesso della guerra de portugueses levantados em Pernambuco contra Olandeses, como por Carta del' Maestro a Campo Martino Soarez, et Andrea Vidal de Negreiros, por Antonio Telles da Silva. El Anno 1646
Local de Publicação: [Bahia] : [s.n.]
Ano de Publicação: 1646
Descrição Física: 20 p.
Idioma: Português
Conteúdo: Contém : Carta de Martim Soares Moreno e André Vidal de Negreiros para Antonio Telles da Silva, Governador Geral da Bahia, datada de 3 de setembro de 1646; Carta de João Fernandes Vieira a Antonio Telles da Silva, datada de 2 de dezembro de 1646; "Copia da Carta que os Ministros da Campanhia Governadores no Recife de Pernambuco escriveraon a os Mestres de Campo, Governadores de quela Capitania depois de ser chegado o Sigismondo."; "Resposta que os Mestres de Campo Governadores em Pernambuco deraon a sobre dita Carta dos Ministros da Companhia", datada de 11 setembro de 1646
URI: http://www.brasiliana.usp.br/bbd/handle/1918/01752400
Tipo: Folheto

Arquivos Relacionados a este texto nos anexos eletrônicos:

Texto resultante do reconhecimento automático (HTML)

Texto com reconhecimento automático corrigido (HTML)

Texto modernizado (HTML)

Lista das correções de reconhecimento automático (HTML)

Lista completa das edições realizadas (HTML)

Antroponímia do texto (HTML)

Toponímia do Texto (HTML)

Texto de Base Anotado (TXT)

Texto de Base Anotado (XML)

Nova Lusitânia ... (1675)

Texto trabalhado por Márcia Aparecida dos Santos Mendes e Leila Rosa de Oliveira

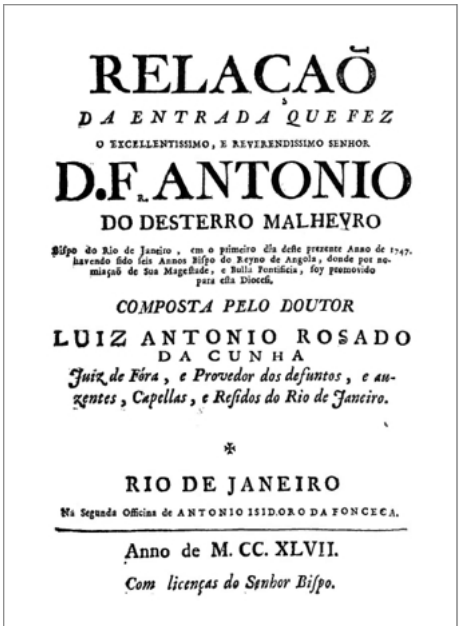
Ficha Catalográfica:
Autor: Freire, Francisco de Brito, ca. 162?-1692
Colaborador: Bérain, Jean, (il.)
Título: Nova Lusitania, historia da guerra Brasilica [...]
Título: [...] a purissima alma e saudosa memoria do serenissimo principe Dom Theodosio pincipe de Portugal, e principe do Brasil, por Francisco de Brito Freyre. Decada primeira
Local de Publicação: Lisboa: Officina de Joam Galram
Ano de Publicação: 1675
Descrição Física: 460 p.; front.; possui índice
Idioma: Português
Descrição: Frontispício foi gravado por Bérain. O livro possui dedicatória do autor.
Resumo: A História da Guerra Brasílica é uma das melhores fontes portuguesas para os acontecimentos ocorridos durante o período holandês no Brasil do século XVII. Foi escrito durante os 6 anos em que o autor esteve no cativeiro devido a desavenças políticas. O texto é construído com base na cronologia dos eventos, nas fontes documentais e nos relatos dos acontecimentos. Ao mesmo tempo, tem características da literatura ficcional quando o autor interpreta e dá voz a seus personagens históricos. O primeiro capítulo trata das conquistas ultramarinas e das glórias de Portugal, mas o Brasil – ou Nova Lusitânia – é o tema central do livro. Nos capítulos subsequentes trata da guerra holandesa na Bahia, das lutas dos holandeses e portugueses, da instalação dos holandeses em Pernambuco e da retirada dos moradores que se recusavam a permanecer sob seu domínio.
URI: http://www.brasiliana.usp.br/bbd/handle/1918/00727000
Tipo: Livro

Arquivos Relacionados:

- Texto resultante do reconhecimento automático (HTML)
- Texto com reconhecimento automático corrigido (HTML)
- Texto modernizado (HTML)
- Lista das correções de reconhecimento automático (HTML)
- Lista completa das edições realizadas (HTML)

Relação da entrada ... (1747)

Texto trabalhado por Fabiana Marcondes Ferraz



RELAÇAÕ

DA ENTRADA QUE FEZ

O EXCELLENTISSIMO, E REVERENDISSIMO SENHOR

D.F. ANTONIO

DO DESTERRO MALHEYRO

Bispo do Rio de Janeiro, em o primeiro dia deste prezente Anno de 1747. havendo sido seis Annos Bispo do Reyno de Angola, donde por nomiaçaõ de Sua Magestade, e Bulla Pontificia, foy promovido para esta Diocesi.

COMPOSTA PELO DOUTOR

LUIZ ANTONIO ROSADO DA CUNHA

Juiz de Fóra, e Provedor dos defuntos, e auzentes, Capellas, e Resíduos do Rio de Janeiro.

RIO DE JANEIRO

Na Segunda Officina de ANTONIO ISIDORO DA FONSECA.

Anno de M. CC. XLVII.

Com licenças do Senhor Bispo.

Ficha Catalográfica:
Autor: Cunha, Luís Antonio Rosado da
Título: Relação da entrada que fez o excellentissimo, e reverendissimo senhor D. Fr. Antonio do Desterro Malheyro [...]
Título: [...] Bispo do Rio de Janeiro, em o primeiro dia deste prezente Anno de 1747 havendo sido seis Annos Bispo do Reyno de Angola, donde por nomiaçaoo de Sua Magestade, e Bulla Pontificia, foy promovido para esta Diocesi. Composta pelo doutor Luiz Antonio Rosado da Cunha Juiz de Fóra, e Provedor dos defuntos, e auzentes, Capellas, e Residuos do Rio de Janeiro
Local de Publicação: Rio de Janeiro : Na Segunda Officina de Antonio Isidoro da Fonseca
Ano de Publicação: 1747
Descrição Física: 20 p.
Idioma: Português
Resumo: Trata-se de um folheto de 24 páginas escrito pelo juiz-de-fora Antonio Rosado da Cunha, o qual narra as celebrações decorrentes da vinda de D. Antonio do Desterro Malheyro, novo diocesano da cidade do Rio de Janeiro, em 1747. Nele, são citados o desembarque, a hospedagem, a visita à ópera, as passagens processionais e as homenagens feitas ao bispo. O tema do documento é, portanto, passageiro, mas seu valor histórico advém, principalmente, das circunstâncias de sua impressão, se dando esta num momento que antecede em mais de meio século a vinda definitiva da imprensa para o Brasil, quando a prática ainda era interdita na colônia. Antonio Isidoro da Fonseca foi impressor em Lisboa e fundou, no Rio de Janeiro, uma oficina tipográfica de onde saíram, comprovadamente, três obras, dentre elas, a Relação da entrada [...], considerado este o primeiro livro impresso no Brasil. O exemplar da Relação da entrada [...] disponibilizado pela Brasiliana pertenceu a Rubens Borba de Moraes. A julgar pelo exemplar, este foi encadernado e possivelmente adquirido em Paris.
URI: http://www.brasiliana.usp.br/bbd/handle/1918/03908100
Tipo: Folheto

Arquivos Relacionados:

Texto resultante do reconhecimento automático (HTML)

Texto com reconhecimento automático corrigido (HTML)

Texto modernizado (HTML)

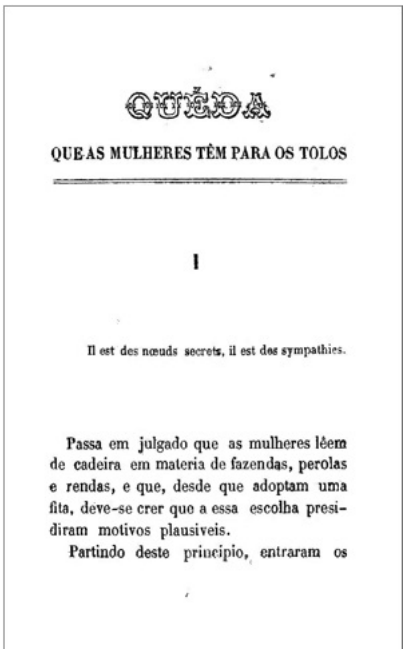
Lista das correções de reconhecimento automático (HTML)

Lista completa das edições realizadas (HTML)

Queda que as mulheres têm ... (1861)

Texto trabalhado por Thelma Tavares Dias

Ficha Catalográfica:



QUEDA

QUE AS MULHERES TÊM PARA OS TOLOS

I

Il est des nœuds secrets, il est des sympathies.

Passa em julgado que as mulheres lêem de cadeira em materia de fazendas, perolas e rendas, e que, desde que adoptam uma fita, deve-se crer que a essa escolha presidiram motivos plausiveis.

Partindo deste principio, entraram os

Colaborador: Assis, Machado de, 1839-1908 (trad.)
Título: Queda que as mulheres têm para os tolos
Local de Publicação: Rio de Janeiro : Typographia de F. de Paula Brito
Ano de Publicação: 1861
Descrição Física: 43 p.
Idioma: Português
Resumo: Pequeno texto de crítica de costumes, publicado primeiramente no periódico A Marmota (Rio de Janeiro, 19, 23, 26 e 30/04 e 03/05/1861) e depois neste pequeno volume de 43 páginas, pela mesma Typographia de Francisco de Paula Brito – onde Machado trabalhou como revisor e colaborador. O texto, durante muito tempo foi creditado ao próprio Machado. Recentemente, foi estabelecido que se trata de uma versão do panfleto anônimo (atribuído a Victor Henaux), De l'amour des femmes pour les sots (Liège, F. Renard, 1859).
Assunto:
Literatura brasileira - Séc. XIX
Teatro (Literatura) - Séc. XIX
Assunto:
Brazilian literature - 19th century
Drama - 19th century
URI: http://www.brasiliana.usp.br/bbd/handle/1918/00211100
Tipo: Livro

Arquivos Relacionados

Texto resultante do reconhecimento automático (HTML)

Texto com reconhecimento automático corrigido (HTML)

Texto modernizado (HTML)

Lista das correções de reconhecimento automático (HTML)

Lista completa das edições realizadas (HTML)

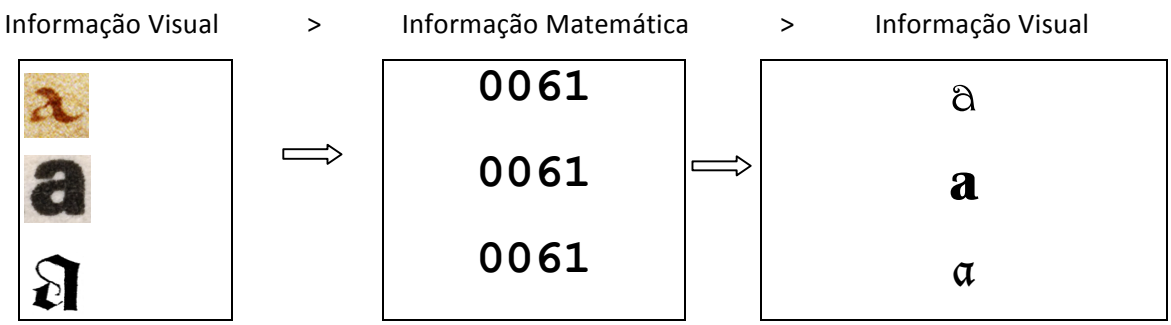
3.2 Trabalho de aprimoramento do reconhecimento automático de caracteres

A primeira etapa do trabalho no projeto teve por objetivo apurar o padrão de acerto de um programa de reconhecimento automático de caracteres aplicados aos textos listados acima, dentro das possibilidades do próprio software, que oferece um modulo de "treinamento" customizável. Para uma compreensão integral deste trabalho, valerá lembrar alguns aspectos técnicos relativos ao reconhecimento automático de caracteres, em especial no que remete ao processamento de textos mais antigos.

3.2.1 O reconhecimento automático de caracteres em textos antigos

O meio digital, por suas características técnicas inerentes, configura uma cadeia de difusão textual complexa - fundamentalmente, por se dar em forma de sobreposição de cópias automáticas (cf. Paixão de Sousa 2007[a]), trazendo o problema da confiabilidade na manutenção das formas e conteúdos. No caso específico da difusão de textos reproduzidos a partir de outros suportes (ou seja, da transposição de textos do meio impresso ou manuscrito em papel para o meio digital), opera-se uma tradução entre tecnologias fundamentalmente diversas, que encerra o problema essencial da fidelidade ao original. Na produção de edições digitais de textos transportados de outros suportes, ao potencial de processamento se choca um fator antagônico: a fidelidade ao original. A reprodução mais fiel aos originais são as imagens ou fac-símiles digitais, justamente os documentos com menor potencial de processamento, uma vez que não permitem buscas por conteúdo: as digitalizações sob forma de imagens são "textos" apenas do ponto de vista humano, não do ponto de vista computacional, o que impede seu processamento por ferramentas automáticas de buscas em seu conteúdo. As digitalizações produzidas como seqüências de caracteres codificados, em contrapartida, são factíveis de instrumentalização para fins de busca e pesquisa por conteúdos.

Os arquivos de textos digitais contemporâneos têm conseguido transpor esse antagonismo entre fidelidade e processamento pela aplicação de tecnologias de reconhecimento óptico de caracteres (OCR, *Optical Character Recognition*), combinada à tecnologia de reunião entre imagem e texto do formato PDF (Portable Document File). Os programas de OCR traduzem imagens de textos em seqüências de caracteres de texto legíveis por computadores, com base nas quais os sistemas de busca funcionam, realizando operações de equivalência entre a seqüência de caracteres da entrada fornecida e a seqüência de caracteres a ser buscada no texto. Trata-se, fundamentalmente, de um processo de tradução de informações gráfico-visuais em informações matemáticas, como ilustra o quadro a seguir:



Os processos de OCR vêm sendo desenvolvidos desde a década de 1950, tendo sido aplicadas diferentes metodologias de reconhecimento. De início, as metodologias eram essencialmente baseadas em reconhecimento de padrões gráficos por análise de estruturas , “”template matching” ou “feature matching” (Mori, 1992; Lui & Fijisawa, 2008). A partir dos anos 1990, foram desenvolvidas tecnologias inteligentes, que incluem algoritmos de reconhecimento por probabilidade, em especial com o recurso a sistemas de reconhecimento por aprendizado (como as redes neurais). Isso gerou uma aproximação entre as comunidades de pesquisa em reconhecimento de padrões e em aprendizado automático, formando o campo complexo hoje dedicado ao reconhecimento automático. Uma primeira característica importante dos modelos de reconhecimento atuais é a ampliação da janela de abordagem: ao contrário dos primeiros sistemas, que trabalhavam “caractere por caractere”, os sistemas atuais abordam unidades maiores, usando o entorno de cada caractere para aprimorar seu reconhecimento (chegando em muitos casos a apoiarem-se na própria organização lógica do documento). Um segundo fator importante no desenvolvimento tecnológico do reconhecimento de caracteres nas últimas décadas é sua relação com a Linguística Computacional. Os programas de reconhecimento atuais incluem dicionários que auxiliam imensamente o reconhecimento de padrões. De fato, podemos dizer que, em comparação com os sistemas antigos (de reconhecimento simples por padrões estruturais), os sistemas atuais por aprendizado são **linguisticamente contingenciados**. Existem programas comerciais altamente eficientes neste aspecto, que conseguem reconhecer documentos em diversas línguas (os mais abrangentes deles, da empresa Abbyy, incluem suporte para 179 idiomas). Notemos, entretanto, que este volume não representa nada mais que a inclusão de 179 dicionários na programação (isto é: o programa não é “multilíngue”, apenas inclui dicionários para cada língua). O recurso às tecnologias de aprendizado possibilitou enormes avanços ao campo do reconhecimento de textos, de modo que os programas de reconhecimento hoje disponíveis comercialmente apresentam taxas de acerto bastante elevadas – chegando a 99%.

Assim, nos acervos de textos contemporâneos, a introdução das modernas técnicas de reconhecimento de caracteres resultou na possibilidade de se oferecerem aos usuários documentos a um tempo fidedignos e processáveis por buscas. De fato, a partir dos arquivos processados por OCR, é possível formar acervos de textos tão fiéis como as imagens, mas que guardam camadas encaixadas, computacionalmente processáveis, com o recurso ao formato *Portable Document File*, PDF, da *Adobe*. Entretanto, no caso dos acervos de textos mais antigos, os dois lados desta equação – reconhecimento de caracteres e buscas baseadas em equivalências de grafias – tornam-se bastante desafiadores.

A taxa de acerto dos programas atuais de reconhecimento de caracteres depende de dois fatores fundamentais: a qualidade e clareza das imagens de base, e a intensidade do treinamento realizado. Isso significa, naturalmente, que os tipos de texto com maior potencial de bons processamentos são aqueles que se originam de informações visuais iniciais mais claras, e cujos padrões tenham sido mais intensivamente alimentados aos programas de treinamento (nisso se inclui: cuja língua seja mais facilmente reconhecida). Por outro lado, os textos com a menor probabilidade de serem bem processados pelos programas de reconhecimento disponíveis hoje são os textos com menor qualidade/clareza da imagem inicial, e/ou os textos escritos em línguas menos exploradas. Exemplo de textos que combinam os dois

fatores de dificuldade seriam os textos mais antigos escritos em Português.

Exemplos de ligaduras da escrita humanística [2]

Nesse tipo de texto, a aplicação da técnica de OCR é desafiante por conta já das particularidades materiais dos textos impressos mais antigos – em especial a formação grafemática e a ortografia distintas da atual. Os programas disponíveis hoje não reconhecem elementos tipográficos obsoletos. Na figura ao lado alguns desses elementos são ilustrados: desenhos de tipos em desuso, e caracteres em desuso (ligaduras, etc).

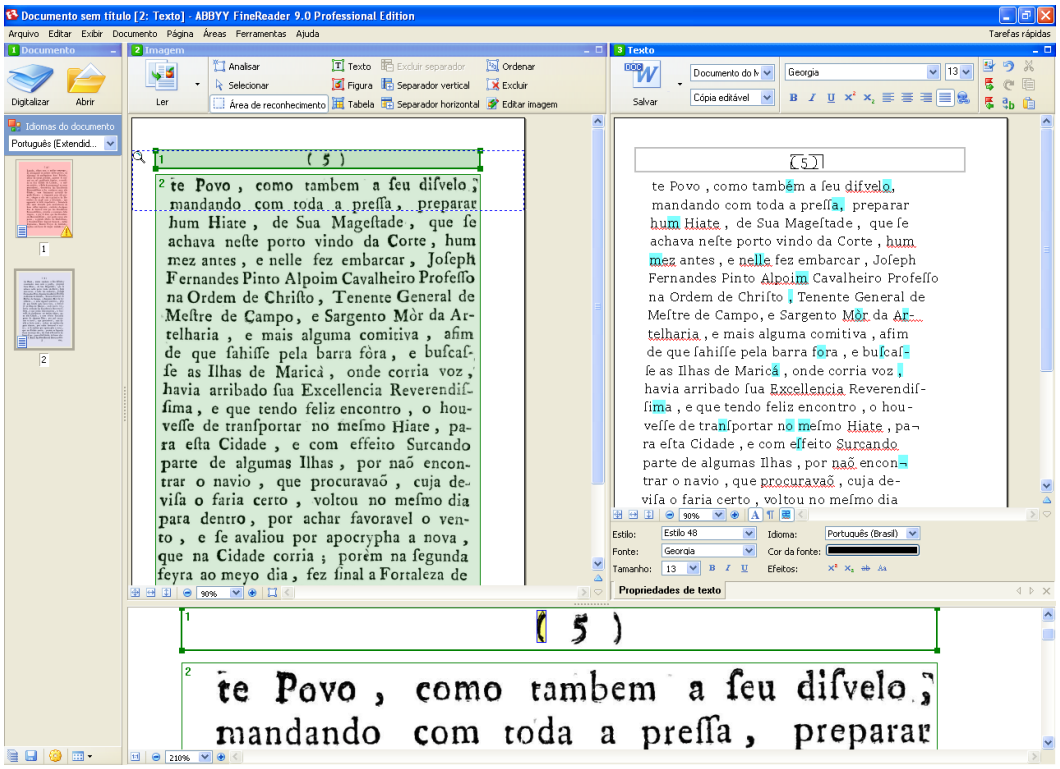
Um segundo aspecto com impacto direto sobre o processamento da informação dos textos antigos está para além do reconhecimento óptico: a diferença entre as grafias antigas e a atual. As grafias antigas trazem dois problemas principais para os acervos digitais. Em primeiro lugar, note-se que, ainda que o reconhecimento de caracteres em um texto antigo seja perfeito, este texto ainda apresentaria dificuldades de leitura consideráveis para o leitor moderno não especializado, que teria grandes dificuldades em aproveitar a possibilidade de buscas por palavras neste documento (a qual, lembremos, é a razão principal pela opção de apresentar os textos de uma biblioteca como textos processáveis, e não imagens). De fato, supõe-se que os consulentes de uma biblioteca digital fornecerão entradas de busca nas formas ortográficas atuais, e os mecanismos de busca simples não estão preparados para realizar a correspondência com as formas antigas. Interessa notar que ao usuário da Biblioteca é oferecido um arquivo PDF, com a transcrição em (b) acima oculta – assim, os usuários não vêem os erros, mas enfrentam obstáculos para realizar buscas profícuas; a impressão superficial é a de que "*as buscas não funcionam*" - quando, de fato, são os subsídios fornecidos ao programa de buscas que estão imperfeitos.Nota-se, portanto, que o problema das grafias tem uma implicação adicional ao problema do reconhecimento de caracteres, pois não remete a uma limitação técnica dos programas de processamento, mas sim às diferenças entre a língua atual e a língua antiga. Além desta questão da dificuldade da leitura humana, a diferença de grafias configura um novo aspecto de dificuldade para o próprio reconhecimento de caracteres – tendo em vista que, como dito, os programas modernos fazem uso de dicionários embutidos para auxiliar o processo de reconhecimento de padrões gráficos. Significa dizer que as grafias antigas são problemáticas para a leitura humana e para o processamento computacional dos textos, já que esse processamento funciona com base em regras ou algoritmos criados com base nas grafias modernas.

3.2.2 O Treinamento de um software de reconhecimento automático

Nossa primeira proposta para abordar essas duas dificuldades no processamento de textos antigos portugueses foi o trabalho de aperfeiçoamento de uma ferramenta automática de reconhecimento de caracteres . Reconhecemos que os caminhos para a solução do problema do reconhecimento da tipografia antiga pelas tecnologias de reconhecimento óptico de caracteres atuais passa, necessariamente, por uma conjunção entre três campos de pesquisa: a filologia, a lingüística computacional, e a engenharia de sinais (área em que se insere o desenvolvimento das tecnologias de reconhecimento de imagens). Nossas pesquisas preliminares mostram que esta conjunção é inexplorada no Brasil. No cenário internacional, algumas experiências notáveis neste sentido vêm sendo desenvolvidas nos últimos anos. Destacam-se, em primeiro lugar, as experiências que podem ser examinadas no periódico "Digital Medievalist"[2], e a experiência de parceria entre o projeto METAe[3] e a empresa Abbyy, resultando na construção do *Abbyy FineReader XIX*, um OCR capaz de reconhecer caracteres góticos do século XIX[4]. O software utilizado na Biblioteca é do mesmo fabricante, e nosso grupo vem procurando adaptá-lo para a leitura de textos em português dos séculos XVI a XVIII. Desenvolver tecnologias de reconhecimento capazes de lidar com a tipografia mais antiga seria um projeto coletivo de monta, no qual o apoio de estudiosos da filologia, especialistas em textos portugueses da imprensa mais antiga, será fundamental – residindo neste ponto a principal colaboração do presente projeto.

Neste projeto, procuramos lançar as bases iniciais para esta possibilidade ao treinar o software Abbyy Fine Reader 9.0 para a leitura de textos portugueses impressos nos séculos XVI a XVIII. Os testes sistemáticos realizados ao longo de 2010 utilizaram o módulo de "Treinamento de padrão" oferecido pelo programa.

2 Cf. <http://www.digitalmedievalist.org/>
3 Cf. <http://meta-e.aib.uni-linz.ac.at/>
4 Cf. <http://www.frakturschrift.com/>

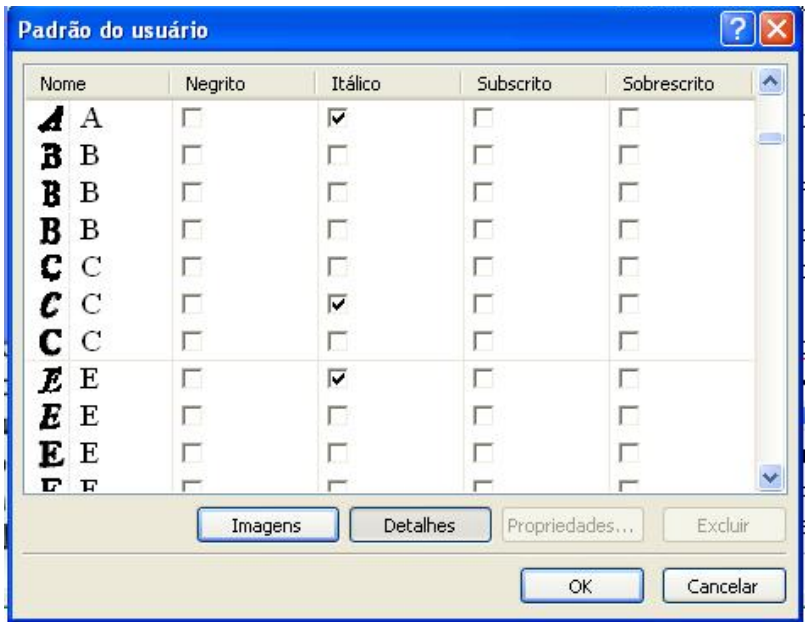


Janela de Leitura - Abbyy 9.0 1

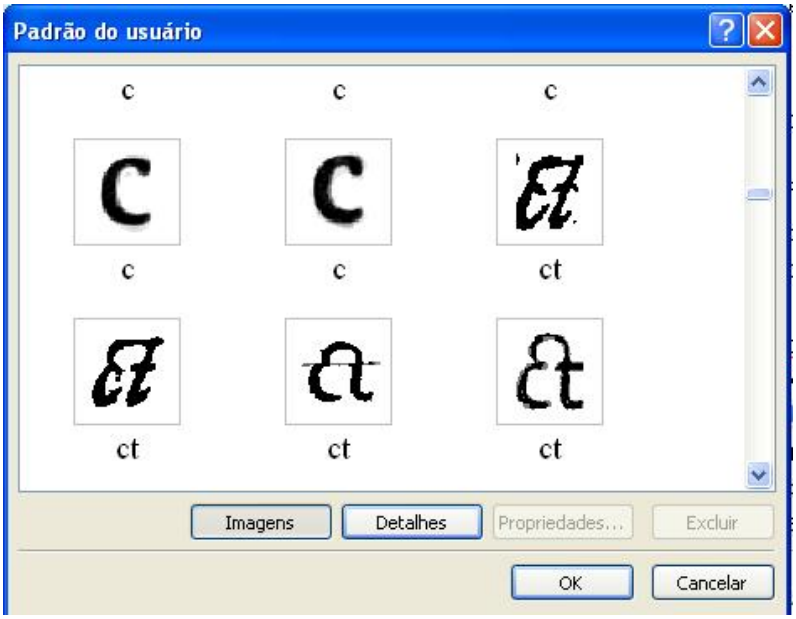
Neste módulo, é possível visualizar, caractere a caractere, os erros de leitura do programa, e corrigi-los um a um. O trabalho nesta etapa consiste, fundamentalmente, no exame detido de cada caractere reconhecido pelo programa (em uma janela de diálogo como a exemplificada abaixo). Caso um caractere não tenha sido reconhecido corretamente, o caractere correto é inserido pelo editor..

Janela de treinamento - Abbyy 9.0 1

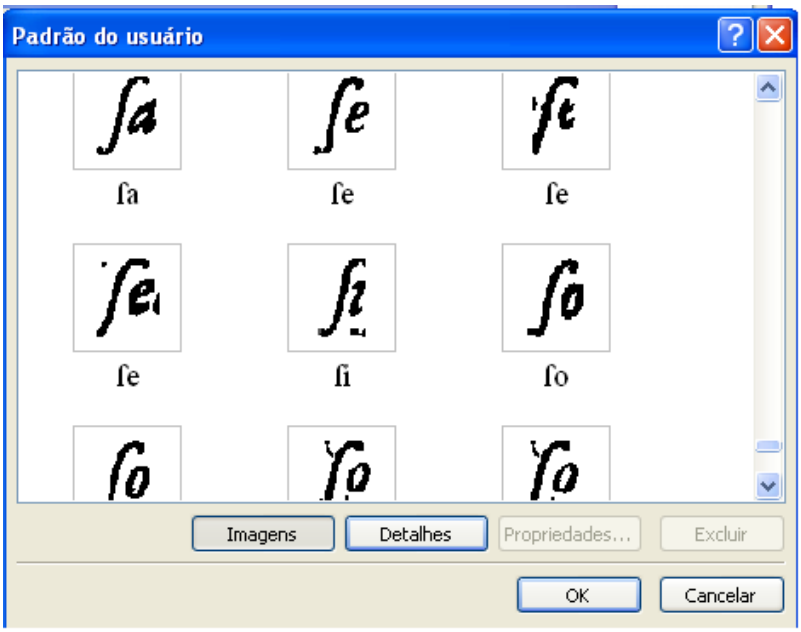
Neste processo, o programa armazena em sua memória a relação entre a imagem isolada no texto e o caractere (ou seja: o símbolo matemático) correto. Como resultado do treinamento, o programa passa a contar com um inventário expandido de imagens possíveis para cada símbolo matemático (correspondentes a a, b, c, etc), como ilustra (I.2). Observe-se que, neste ponto, é possível inserir, inclusive, uma relação de caracteres em desuso - como as ligaduras, o S longo, o til sobre o o, etc. (cf. I.3) Para isso, precisamos pesquisar nas tábuas de caracteres completas atualmente disponíveis (como o UTF, que utilizamos nas nossas edições - cf. http://unicode.org/charts/) os códigos correspondentes aos caracteres antigos, e armazená-los no banco de dados do programa.



Inventário do Padrão - Abbyy 9.0 1



Inventário do Padrão - Abbyy 9.0 2



Inventário do Padrão - Abbyy 9.0 3

Os inventários de caracteres "treinados" podem ser armazenados e utilizados para novas leituras do mesmo texto ou de textos semelhantes - resultando, em tese, em um melhor reconhecimento automático. Assim, a característica mais interessante deste módulo, segundo anuciada pelo próprio fabricante do programa, seria a de corresponder a um efetivo aprendizado por parte do programa – de modo que os resultados se tornam cumulativamente mais precisos, conforme o padrão treinado vai sendo ajustado às próximas leituras. Os resultados que obtivemos com esta técnica são descritos abaixo.

3.2.3 Resultados desta etapa

Neste ano, tratamos quatro obras com este método. Escolhemos inicialmente três das obras mais antigas do acervo, por conta da constatação de que nestas obras os resultados do reconhecimento automático inicial eram os piores possíveis. Paralelamente, trabalhamos também com uma obra mais moderna, impressa no século XIX, para verificarmos o desempenho do método com materiais menos desafiadores. As obras assim escolhidas foram:

(i) [Autor Desconhecido], 1642- Trattado das tregoas e suspensao do todo o acto de hostilidade ....

(ii) Britto Freire, Francisco de , 1675 - Nova Lusitania, historia da guerra Brasilica [...]

(iii) Cunha, Luís Antonio Rosado da, 1747 - Relação da entrada que fez o excellentissimo, e reverendissimo senhor D. Fr. Antonio do Desterro Malheyro ....

(iv) Assis, J.M Machado de (tradução), 1861 - A Queda que as mulheres têm para os tolos.

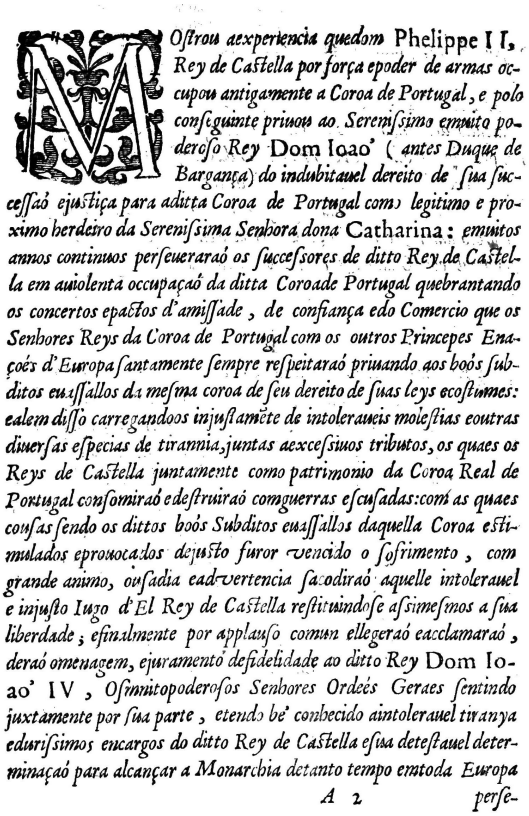
Chegamos a uma melhora sensível entre a primeira leitura automática (sem treinamento) e a leitura final (com treinamento). A taxa média de acertos nesses textos era **59%** no início do treinamento, e subiu para **86%** no final dos trabalhos. Esta taxa média, entretanto, não reflete os contrastes dos resultados a depender dos textos, o que se nota melhor na tabela abaixo:

Texto	Taxa de arcertos do OCR antes do treinamento	Taxa de acertos do OCR depois do treinamento
(2) Tratado de tréguas..., 1642	57%	67%
(5) Nova Lusitânia, 1675	81%	92%
(6) Relação da entrada..., 1747	77%	86%
(7) Queda que as mulheres..., 1861	87%	95%

Uma primeira constatação possibilitada pela comparação dos resultados é a de que de fato, conforme previámos, os resultados foram melhores no texto mais moderno (impresso em 1861). Este não é, entretanto, o único fator relevante - note-se o contraste entre o acerto final de 92% no texto de 1675 e o acerto final de 67% no texto relativamente próximo, de 1642, e o acerto final de 86% no texto mais moderno, de 1747. Podemos creditar a taxa relativamente elevada de acerto nesta obra de 1675 a suas características materiais: Nova Lusitânia é uma obra-prima da impressão seiscentista (cf. Antunes, 2009), e o exemplar digitalizado pela Brasiliana está em excelente estado de conservação. Comparem-se as imagens de páginas exemplares em cada caso nas figuras abaixo (uma melhor apreciação das diferenças pode ser vista no site do projeto, nos endereços respectivos dados ao lado das figuras).

Página de Nova Lusitânia, impresso em 1675

(Obra completa em http://www.brasiliana.usp.br/bbd/handle/1918/00727000)



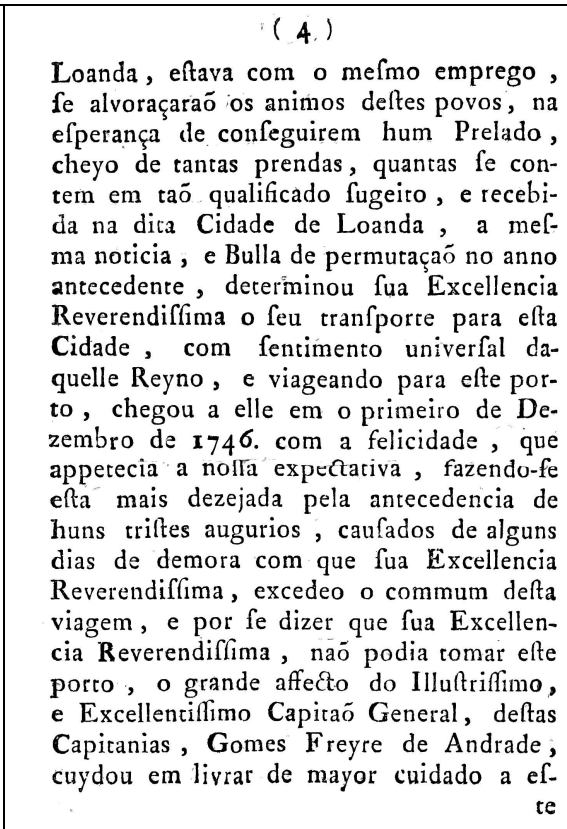
Mostrou aexperiencia quedom Phelippe II, Rey de Castella por força epoder de armas occupou antigamente a Coroa de Portugal, e polo conſiguinte priuou ao Sereniſsimo emuito poderoſo Rey Dom Ioaõ (antes Duque de Bargança) do indubitauel dereito de ſua ſucceſſaõ ejuſtiça para aditta Coroa de Portugal com legitimo e proximo herdeiro da Sereniſsima Senhora dona Catharina: emuitos annos continuos perſeueraraõ os ſucceſſores de ditto Rey de Caſtella em auiolenta occupaçaõ da ditta Coroade Portugal quebrantando os concertos epactos d'amiſſade, de confiança edo Comercio que os Senhores Reys da Coroa de Portugal com os outros Princepes Enaçoés d'Europa ſantamente ſempre reſpeitaraõ priuando aos boõs ſubditos euaſſallos da meſma coroa de ſeu dereito de ſuas leys ecoſtumes: ealem diſſo carregandoos injuſtamẽte de intoleraueis moleſtias eoutras diuerſas eſpecias de tirannia, juntas aexceſsiuos tributos, os quaes os Reys de Caſtella juntamente como patrimonio da Coroa Real de Portugal conſomiraõ edeſtruiraõ comguerras eſcuſadas: com as quaes couſas ſendo os dittos boõs Subditos euaſſallos daquella Coroa eſtimulados eprouocados dejuſto furor vencido o ſofrimento, com grande animo, ouſadia eadvertencia ſacodiraõ aquelle intoleraueł e injuſto Iugo d'El Rey de Caſtella reſtituindoſe aſsimeſmos a ſua liberdade; efinalmente por applauſo comun ellegeraõ eacclamaraõ, deraõ omenagem, ejuramento defidelidade ao ditto Rey Dom Ioaõ IV, Oſmuitopoderoſos Senhores Ordeés Geraes ſentindo juxtamente por ſua parte, etendo bẽ conhecido aintoleraueł tirannya eduriſsimos encargos do ditto Rey de Caſtella eſua deteſtauel determinaçaõ para alcançar a Monarchia detanto tempo emtoda Europa

A 2 perſe-

Página digitalizada de *Tratado...,* impresso em 1642

(Obra completa:

http://www.brasiliana.usp.br/bbd/handle/1918/01936100)



(4)

Loanda, eſtava com o meſmo emprego, ſe alvoraçaraõ os animos deſtes povos, na eſperança de conſeguirem hum Prelado, cheyo de tantas prendas, quantas ſe contem em taõ qualificado ſugeito, e recebida na dita Cidade de Loanda, a meſma noticia, e Bulla de permutaçaõ no anno antecedente, determinou ſua Excellencia Reverendiſſima o ſeu tranſporte para eſta Cidade, com ſentimento univerſal daquelle Reyno, e viageando para eſte porto, chegou a elle em o primeiro de Dezembro de 1746. com a felicidade, que appetecia a noſſa expectativa, fazendo-ſe eſta mais dezejada pela antecedencia de huns triſtes augurios, cauſados de alguns dias de demora com que ſua Excellencia Reverendiſſima, excedeo o commum deſta viagem, e por ſe dizer que ſua Excellencia Reverendiſſima, naõ podia tomar eſte porto, o grande affecto do Illuſtriſſimo, e Excellentiſſimo Capitaõ General, deſtas Capitanias, Gomes Freyre de Andrade, cuydou em livrar de mayor cuidado a eſte

Página digitalizada de *Relação...,* impresso em 1747.

(Obra completa:

http://www.brasiliana.usp.br/bbd/handle/1918/03908100)

O treinamento do programa de reconhecimento, em resumo, resultou em um aprimoramento relativo das taxas de acerto - cerca de 10% de elevação nas taxas - mas não conduziu a um patamar de leitura excelente. Uma taxa de acerto de 92% (nosso melhor resultado para um texto mais antigo) significa que, em 100 palavras buscadas no texto, apenas 92 serão encontradas; consideramos que, para um estudo linguístico ou filológico, esse resultado é ainda bastante sofrível. Levamos ainda em conta que este percentual de acerto máximo de 92%, obtido com o texto Nova Lusitânia, é excepcional (o que, como dissemos, remete às condições materiais ímpares desta obra). No cenário mais geral dos impressos seiscentistas e setecentistas portugueses, permaneceríamos na faixa de 67% a 86% de acertos - o que significaria que, no mínimo, 14 em cada 100 palavras permaneceriam invisíveis para buscas automáticas, perspectiva inaceitável.

Nossa conclusão, com este experimento, foi a de que mesmo os melhores resultados do treinamento ainda precisariam ser aprimorados. Consideramos ainda que o treinamento foi exaustivo, tendo tomado a metade do tempo dedicado à pesquisa neste período. Sem acesso ao funcionamento interno do programa (que é proprietário e fechado), não nos pareceu possível almejar resultados melhores - noutros termos, concluímos ter chegado ao limite de resultados do programa. Assim, em um balanço realizado ao final do primeiro semestre de pesquisas, decidimos avançar para uma nova etapa na busca de um melhor processamento automático dos textos: a preparação editorial, descrita na seção seguinte.

3.3 Trabalho de intervenção editorial nos textos resultantes do reconhecimento automático

De posse das melhores leituras produzidas pelo Abbyy 9.0, procedemos à sua correção via intervenção editorial, com auxílio da ferramenta E-Dictor, como descrevemos a seguir. Além dos textos já citados, incluímos nesta etapa o resultado da leitura automática das seguintes obras:

(i) Montalvão, Marquês de, 1642 - Cartas que o Marquez de Montalvam, sendo Viso Rey...

(ii) [Autor Desconhecido], 1646 - Successo de la guerra de portugueses Levantados

(iii) Melo, Francisco Manuel de, 1650 - Relaçam dos sucessos da armada, que a Companhia Geral

3.3.1 A opção pela intervenção editorial

O segundo desafio no processamento computacional de textos antigos apontado na seção anterior foi a questão da variação de grafias. Nossas experiências anteriores de pesquisa indicam duas abordagens possíveis para solucionar esse desafio: a intervenção editorial, e a aplicação de programas automáticos de reconhecimento de variação. No primeiro caso, os documentos são tratados por um pesquisador humano, que interpreta e moderniza o texto. No segundo caso, programaram-se sistemas de busca especialmente treinados para textos antigos, que realizam equivalências entre as formas antigas e modernas, resultando assim em buscas satisfatórias. No campo dos programas automáticos de busca, destacam-se, para os textos antigos em português, as propostas de Aluísio (2007), aplicadas segundo a técnica de Candido Jr (2008) ao Corpus fundamental do *Dicionário Histórico do Português do Brasil,* DHPB (Biderman, 2005). O desenvolvimento de programas automáticos de reconhecimento de grafias antigas apresenta diversas vantagens, sendo a mais evidente a potencial economia de tempo e recursos humanos. Entretanto, no atual estágio, esse sistema apresenta a desvantagem da baixa precisão em textos mais complexos: a ferramenta não suporta variações mais idiossincráticas como, por exemplo, as que caracterizam os textos manuscritos ou os textos impressos quinhentistas, seiscentistas e mesmo setecentistas.

Já a técnica da intervenção editorial consiste, fundamentalmente, na estratégia tradicional da edição filológica. O diferencial, neste caso, pode ser o uso de uma metodologia computacional no trabalho, nos moldes do projeto *Memórias do Texto*[5] (Paixão de Sousa 2004, 2005, 2006[b], 2006[c], 2007[b]; Trippel & Paixão de Sousa 2006). Nesse método, a partir do texto original, ou de uma digitalização, é elaborado um arquivo XML de base (b), onde se anotam as variações grafemáticas e de grafia. Esse arquivo-base gera versões automáticas, como edições diplomáticas ou modernizadas, acessíveis para a leitura de um público amplo e, ao mesmo tempo, para a leitura por máquinas[6]. O sistema é baseado em Linguagem de Marcação Extensível (XML) e suas tecnologias correlatas, Transformação em Folhas de Estilo Extensível (XSLT) e Busca Extensível (XQ) (cf. W3C 2008 [a],[b],[c]), de código aberto, independente de aplicativos comerciais e plataformas operacionais. Essa técnica já havia sido aplicada a uma coleção de textos portugueses dos

5 Projeto de pós-doutorado (Fapesp, 04/03462-4), <http://www.ime.usp.br/~tycho/participants/psousa/memorias>

6 Cf. Edição Edição Eletrônica integral: <http://www.tycho.iel.unicamp.br/~tycho/corpus/texts/xml/g_008.xml>

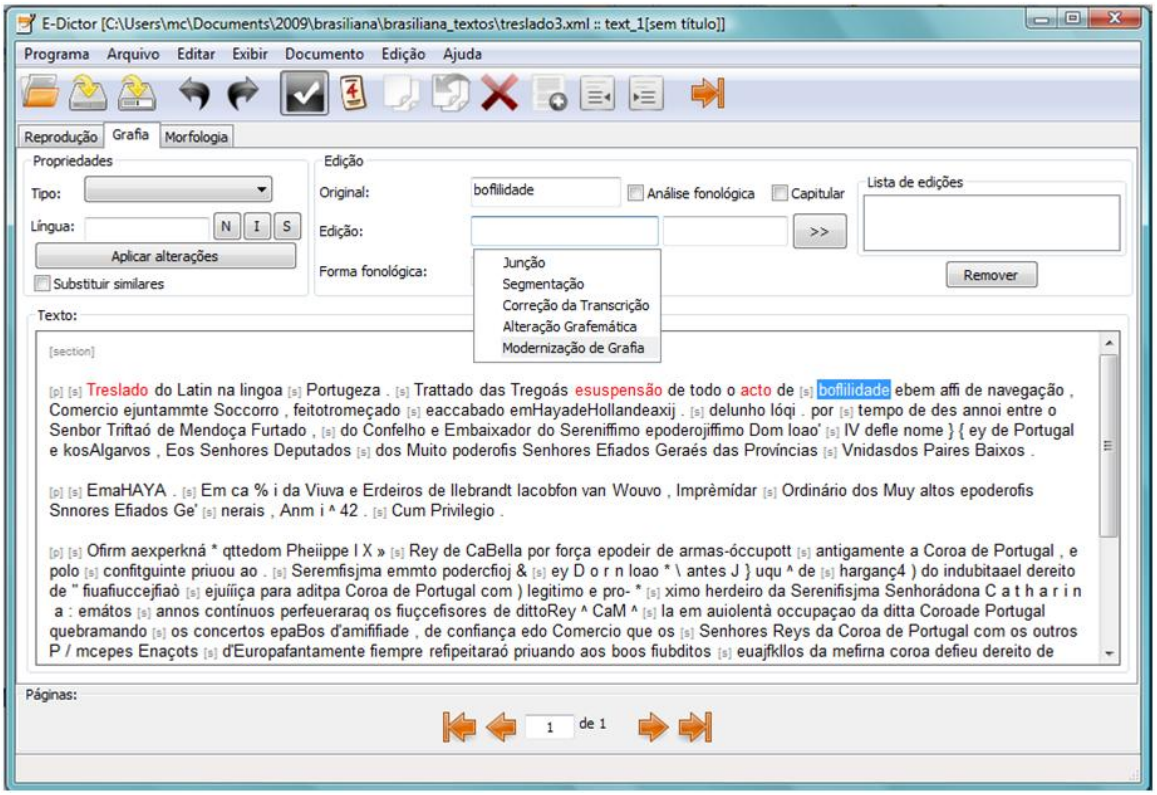
séculos 16 a 19 (num total de 2.400.000 palavras), o *Corpus Anotado do Português Histórico Tycho Brahe,* construído no âmbito do projeto *Padrões Rítmicos, Fixação de Parâmetros e Mudança Lingüística*[7].

O tratamento editorial dos textos nesses moldes apresenta duas vantagens principais: primeiro, gera formatos passíveis de leitura automática com finalidade de busca por conteúdo, favorecendo a pesquisa acadêmica em geral a partir do acervo (em particular, com o aproveitamento dos textos para outros programas de anotação, tais como a anotação de categorias morfológicas e sintáticas, para a análise lingüística automática – tendo sido esta sua finalidade original). Além disso, o trabalho de edição gera um sub-produto de interesse para um público mais amplo: os textos em versão modernizada, de leitura facilitada – o que ampliaria, potencialmente, o público leitor de uma Biblioteca. O processo apresenta, entretanto, a desvantagem de demandar um grande investimento de tempo e recursos humanos. Nesse sentido, ainda no âmbito do projeto Tycho Brahe, concebemos uma ferramenta semi-automática para apoiar o trabalho de edição eletrônica, o E-Dictor. No presente projeto, utilizamos essa ferramenta, com grande benefício para os progressos do trabalho - bem como ao próprio desenvolvimento do software. A seção abaixo detalha este processo.

3.3.1 A Edição com a ferramenta E-Dictor

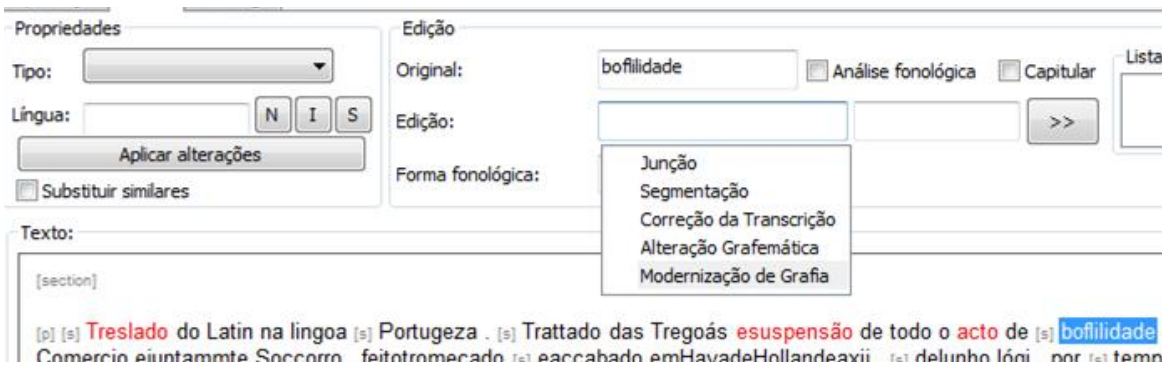
O E-Dictor é um software de anotação concebido como ferramenta auxiliar de anotação eletrônica (Paixão de Sousa, Kepler & Faria, 2010; Paixão de Sousa & Kepler, 2007), atualmente em uso pela equipe de edição da BBD. A grande motivação para o desenvolvimento do E-Dictor foi a experiência de anos de codificação manual em XML, que mostrou acarretar os seguintes problemas: (i) dificuldades no treinamento de codificadores;(ii) erros estruturais de codificação (digitação, por exemplo) que passavam despercebidos; (iii) muito tempo gasto na codificação e em revisões, em função dos problemas acima. O E-Dictor facilitou sobremaneira o treinamento dos codificadores: sua tarefa agora é basicamente a de compreender o processo de edição e aprender a usar a ferramenta, sem necessidade de compreender a lógica por trás do XML. A estrutura da ferramenta torna impossíveis os erros de codificação; e representa uma redução de no mínimo 50% no tempo de edição. O E-Dictor, através de sua interface, visa a evitar o contato direto entre o editor (usuário) e a estrutura XML subjacente. Para isso, a interface prima pela exibição do conteúdo textual, deixando as marcas de estrutura em segundo plano, embora visíveis (estas são importantes, afinal o editor precisará ter acesso às quebras de linha, página, marcas de fim de sentença, parágrafo, etc.). A interface principal de edição do E-Dictor está ilustrada na figura a seguir.

[7] Projeto Temático Fapesp, <http://www.tycho.iel.unicamp.br/~tycho>



E-Dictor - Interface de Edição 1

Para fins deste relatório, será pertinente apenas descrevermos brevemente as funcionalidades de edição gerais da ferramenta (um maior detalhamento está exposto em Paixão de Sousa, Kepler e Faria, 2010). A figura a seguir amplia a interface disponível (aba "Grafia") para esta tarefa:



E-Dictor - Interface de Edição 2

O item "bofilidade", ressaltado em fundo azul, é a palavra sendo editada na figura. É sobre ela que podemos fazer uma série de modificações, comentadas a seguir:

- **Painel "Propriedades":** aqui, podemos especificar o "Tipo" do símbolo (de acordo com as definições de subtipos previstas nas preferências da aplicação), a "Língua" ou idioma (se for estrangeiro), opções de formatação (negrito, itálico e sublinhado) e temos o botão "Aplicar alterações" com a opção para "Substituir similares", comentados mais à frente.

- **Painel "Edição":** aqui, podemos marcar algumas propriedades do símbolo, bem como inserir os níveis de edição (de acordo com os níveis informados nas preferências da aplicação). Vamos aos seus elementos:

- **O campo "Original"** exibe a forma original do símbolo, como transcrita do texto-fonte. Normalmente, a forma original não deve ser alterada, mas se preciso, pode ser feito. Repare que o E-Dictor exibe sempre o nível máximo de edição na área de texto, não o texto original. À frente deste campo, temos as propriedades "Análise fonológica" (que diz ao E-Dictor para exportar a forma original para análise fonológica) e "Capitular" (que informa que no texto original esta palavra inicia com capitular).
- **O campo "Edição"** permite escolher um nível de edição, cujo conteúdo será especificado no campo imediatamente à frente. Após informar o conteúdo, é preciso clicar no botão ">>" para incluir o nível de edição na "Lista de Edições".
- **A "Lista de Edições":** lista as edições incluídas para o símbolo, permitindo sua alteração ou exclusão, através do botão "Remover".

O resultado de quaisquer destas alterações será armazenado em um arquivo de base em linguagem XML, com um formato-fonte como o exemplo a seguir:

```
<?xml version="1.0" encoding="utf-8" ?>
- <document>
+ <head id="isidoro_pag_11">...- <body>
- <text t="full" words="170" id="text_1">
- <sc id="sc_1">
- <p id="p_1">
- <s id="s_1">
  - <w id="s_1#0"><o>Loanda</o> <e t="mod">Luanda</e> </w>
  - <w id="s_1#1"><o>,</o> </w>
  - <w id="s_1#2"><o>eſtava</o> <e t="graf">estava</e> </w>
  - <w id="s_1#3"><o>com</o> </w>
  - <w id="s_1#4"><o>o</o> </w>
  - <w id="s_1#5"><o>meſmo</o> <e t="graf">mesmo</e> </w>
  - <w id="s_1#6"><o>emprego</o> </w>
  - <w id="s_1#7"><o>,</o> </w>
  - <w id="s_1#8"><o>ſe</o> <e t="graf">se</e> </w>
  - <w id="s_1#9"><o>alvoraçaraõ</o> <e t="mod">alvoroçaram</e> </w>
  - <w id="s_1#10"><o>os</o> </w>
  - <w id="s_1#11"><o>ânimos</o> </w>
  - <w id="s_1#12"><o>deſtes</o> <e t="graf">destes</e> </w>
  - <w id="s_1#13"><o>povos</o> </w>
  - <w id="s_1#14"><o>,</o> </w>
  - <w id="s_1#15"><o>na</o> </w>
  - <w id="s_1#16"><o>eſperança</o> <e t="graf">esperança</e> </w>
  - <w id="s_1#17"><o>de</o> </w>
  - <w id="s_1#18"><o>conſeguirem</o> <e t="graf">conseguirem</e> </w>
  - <w id="s_1#19"><o>hum</o> <e t="mod">um</e> </w>
  - <w id="s_1#20"><o>Prelado</o> </w>
  - <w id="s_1#21"><o>,</o> </w>
  - <w id="s_1#22"><o>cheyo</o> <e t="mod">cheio</e> </w>
  - <w id="s_1#23"><o>de</o> </w>
```

Fundamentalmente, o que o código-base forma é um banco de palavras, que pode ser transformado em um texto humanamente legível se assim o desejarmos, e no formato que desejarmos. Por conta desta característica, consideramos este sistema como ideal para codificar os textos neste projeto. Ao par da finalidade original da edição em camadas do E-Dictor (tal seja: codificar variações de grafia), acrescentamos uma nova camada de edição: **a codificação dos erros de reconhecimento automático.** O resultado disso é que os arquivos de base neste caso representam um banco de dados de variações de grafia e de erros de reconhecimento. Isso poderá ser útil no desenvolvimento de novos programas de reconhecimento de caracteres voltados para os textos portugueses antigos (conforme nos asseguram os matemáticos e engenheiros ligados ao projeto Brasiliana Digital). Um exemplo desta codificação de erros de reconhecimento seria o seguinte:

```
<w id="s_4#5">
  <o>referc</o>
  <e t="ocr">refere</e>
</w>
```

Neste caso, a palavra "**refere**" foi erroneamente reconhecida como "**referc**" pelo programa de reconhecimento (um erro comum, de troca de "**e**" por "**c**"). O editor do texto identificou o erro, e inseriu, no E-Dictor, a forma correta - "refere". O programa anotou, no código XML, as duas formas: a forma original (errada) e a forma inserida pelo editor (correta). Isso significa que as duas formas permanecerão ligadas no arquivo (são versões do item número s_4#5, ou seja, da palavra número 4 da sentença número 5 deste texto), e portanto o erro do programa de reconhecimento fica registrado. Uma lista dos erros mais frequentes poderá ser usada, futuramente, para o treinamento de um programa de reconhecimento voltado aos textos desta época.

Uma última funcionalidade da ferramenta E-Dictor confere mais uma utilidade aos textos ali editados. Com base no documento anotado em XML, diferentes versões (ou visualizações) de um mesmo texto podem ser geradas (ou seja: produzidas, automaticamente, dentro da própria ferramenta E-Dictor). Originalmente, isso foi pensado para possibilitar a geração de versões diplomáticas, semi-diplomáticas ou modernizadas dos textos. Agora, com a aplicação da ferramenta para a correção do reconhecimento, podemos gerar também edições diplomáticas dos textos, com o reconhecimento de caracteres corrigido (o que equivaleria a uma transcrição conservadora e correta realizada por um editor humano). Os resultados completos desta etapa são avaliados abaixo.

3.3.3 Resumo dos resultados desta etapa

O trabalho de edição eletrônica dos textos tomou a segunda metade do período de trabalhos no projeto. Foram produzidos assim três resultados principais: primeiro, e mais evidente, seis textos com a sequência de caracteres 100% precisa (que serão futuramente incorporados ao Acervo, favorecendo o resultado das buscas dos usuários). Segundo, a edição filológica dos textos, que agora poderão ser utilizados em três versões: original, modernizada, e morfologicamente etiquetada (para fins de pesquisa linguística).Terceiro, um resultado menos evidente: a lista das palavras corrigidas nesta etapa (produzida como resultado automático do uso do E-Dictor) poderá agora ser utilizada como base para uma segunda etapa de treinamento do software de reconhecimento. De fato: o software Abbyy Finereader 9.0 possui, além do módulo de treinamento, um módulo de dicionário editável, de modo que é possível adicionar itens lexicais a um dos idiomas conhecidos ou mesmo criar um idioma inteiramente novo. Isso significa que os resultados da formação de um dicionário de grafias antigas (subproduto automático da técnica de intervenção editorial com o E-dictor) poderão ser adicionados ao Abbyy, aumentando a capacidade de processamento obtida com o treinamento de caracteres descrito anteriormente.

Os produtos mais concretos desta etapa - os textos em todas as suas possíveis versões - são apresentados no anexo eletrônico a este relatório, em http://lampiao.brasiliana.usp.br/lingua/EnsinarComPesquisa2010Relatorio.

4. Balanço Geral e Perspectivas

No que remete aos resultados imediatos do projeto, a pesquisa conduzida ao longo de 2010 nos levaram a uma importante conclusão geral: a intervenção editorial e o desenvolvimento de programas automáticos de reconhecimento da grafemática antiga e de grafias em variação são abordagens complementares para o desafio da busca por conteúdo em arquivos digitalizados a partir de textos antigos., e devem ser conduzidas paralelamente. De um lado, a possibilidade de conseguirmos treinar um software automático de reconhecimento para processar com eficiência os textos portugueses antigos seria ideal; entretanto, esse ideal, ao que indicam nossos experimentos, pertence ainda ao longo prazo. De outro lado, a técnica da intervenção editorial permite resultados de 100% de precisão em um prazo relativamente curto. Além disso, a edição traz subprodutos muito interessantes, tais como os textos modernizados (quanto à grafemática e quanto às grafias), as listas de palavras, e a análise morfológica. Assim, concluímos que o trabalho de correção dos resultados de OCR via E-Dictor é vantajoso neste momento, tanto com vistas a um resultado palpável mais imediato, como com vistas ao desenvolvimento de softwares de OCR a longo prazo. Essa perspectiva motivou a proposta de um novo projeto, apresentado ao programa Ensinar com Pesquisa 2011, e aprovado em fevereiro deste ano.

Quanto aos resultados menos palpáveis, mas não menos importantes, podemos apontar os seguintes aspectos gerais. Conforme afirmava o projeto inicial, esta proposta de pesquisa pretendia incluir uma dimensão de formação para os alunos envolvidos - em especial, tendo em vista sua inserção num programa fomentado pela Pró-reitoria de Graduação. Na ocasião da proposta inicial, salientou-se que a metodologia de pesquisa pela qual optamos, ao conjugar os campos da filologia e lingüística histórica com o campo da ciência da computação, representaria a abertura de uma linha de pesquisa tradicional da Universidade para a realidade atual da interdisciplinaridade. Para o aluno de letras, tratava-se de uma oportunidade de conhecer campos tecnológicos inovadores, o que só terriam a contribuir para sua formação pessoal e ampliação de seus horizontes futuros de inserção no mundo da pesquisa acadêmica e no mercado de trabalho. Esse caráter interdisciplinar fundamental refletiria-se ainda no ambiente de trabalho que se oferece aos futuros bolsistas do projeto, o *Laboratório da Brasiliana* USP, que reúne pesquisadores das áreas da engenharia, da matemática, da história, da acervologia, e outros, propiciando um espaço de diálogo e interação extremamente favorável a uma experiência universitária contemporânea e estimulante. Pensava-se, assim, estar possibilitando aos alunos uma primeira experiência de pesquisa estimulante e formadora.

Ao final dos trabalhos, concluo, como coordenadora, que esta tarefa foi ao menos parcialmente cumprida. Em sua ampla maioria, os alunos bolsistas envolvidos neste projeto mostraram-se extremamente dedicados ao longo da pesquisa. Quanto às tarefas específicas a serem cumpridas, terminaram-nas no prazo estabelecido, mostrando-se colaborativos em relação aos colegas, compareceram às reuniões de acompanhamento (cuja frequência, na maior parte do ano, foi mensal - cf,

documentação nos anexos eletrônicos), e permaneceram sempre acessíveis para conversas via recursos virtuais. Ressalte-se, ainda, que esses alunos tomaram parte ativa no cotidiano do Laboratório de Pesquisas Brasiliana Digital: para além de suas tarefas imediatas de pesquisa, atenderam aos seminários ali oferecidos mensalmente ao longo de 2010, e colaboraram na organização do Seminário Mindlin 2010 - "O Futuro das Bibliotecas". Penso também ser possível afirmar que esses alunos irão prosseguir com pesquisas neste campo: duas das bolsistas (Leila Rosa de Oliveira e Márcia Aparecida dos Santos Mendes) voltaram a se inscrever na versão deste ano do projeto, e outros dois (Fabiana Marcondes Ferraz e Jáderson Johnattan Porto), embora já não sejam bolsistas, continuam participando das atividades livres de pesquisa do laboratório, e matricularam-se em disciplinas optativas no Instituto de Estudos Brasileiros da USP, relacionadas à área da história da edição em Portugal. Naturalmente, como é esperado em um projeto envolvendo diversos alunos, o grau de engajamento e dedicação de cada um (tanto nas tarefas específicas como neste âmbito mais geral de colaboração com um grupo maior de pesquisadores) não foi homogêneo - como, de resto, se pode depreender dos relatórios individuais elaborados por cada bolsista, que estão disponíveis nos anexos eletrônicos. Assim, numa avaliação geral deste âmbito fundamental da formação discente, o projeto conclui sua primeira etapa com a constatação de um progresso importante quanto ao desenvolvimento acadêmico dos bolsistas.

Referências Bibliográficas

Aluísio, S. (2007). Córpus Históricos, Recursos Léxicos e Ferramentas para a tarefa de criação de dicionários. I Escola Brasileira de Lingüística Computacional USP, Setembro de 2007. <http://moodle.icmc.usp.br/ebralc>

Antunes, C. (2009) Prefácio à Edição eletrônica de "Nova Lusitânia, História da Guerra Brasílica", de F.B. Freire. http://www.brasiliana.usp.br/nova_lusitania

Biderman, M.T. (2005). "Dicionário Histórico do Português do Brasil (sécs XVI, XVII e XVIII)". Projeto CNPq – Milênio.

Candido Jr, A. (2008). "Criação de um ambiente para o processamento de corpus de Português Histórico". Dissertação de Mestrado. Instituto de Ciências da Computação e Matemática Computacional, Universidade de São Paulo.

Castilho, A.T. de (1998) "Para a história do português brasileiro". São Paulo:Humanitas. Vol I: Primeiras idéias.

Galves, A. & Galves, C. (1995). "A case study of prosody driven language change ". Unicamp, Mimeo; <http://www.ime.usp.br/~galves/artigos/clpep.pdf>

Kato, M.A. & Roberts, I. (orgs.) (1993) "Português brasileiro: uma viagem Diacrônica". Campinas: Editora da Unicamp.

Kepler, F.N. (2005) Um Etiquetador Morfo-Sintático Baseado em Cadeias de Markov de Tamanho Variável. Dissertação de Mestrado, Instituto de Matemática e Estatística, Universidade de São Paulo..

Kepler, F.N. (em curso) Parser Sintático Baseado em Cadeias de Markov de Tamanho Variável. Tese de Doutorado em Andamento, Instituto de Matemática e Estatística, Universidade de São Paulo.

Mattos e Silva, R.V. (1988) Fluxo e refluxo: uma retrospectiva da lingüística histórica no Brasil. D.E.L.T.A., 4.1: 85-113. São Paulo.

Megale, H. & Cambraia, C.N. (1999). Filologia Portuguesa no Brasil. D.E.L.T.A, vol. 15, número especial:1:22. São Paulo.

Mindlin, J. (2005). Destaques da biblioteca indisciplinada de Guita e José Mindlin. São Paulo, Edusp/Fapesp; Rio de Janeiro, Fundação Biblioteca Nacional.

Paixão de Sousa, M.C. (2004). Memórias do Texto: Aspectos Tecnológicos na Construção de um corpus histórico do português. Projeto de Pós-doutorado. Departamento de Lingüística, IEL, Unicamp; Fundação de Amparo à Pesquisa do Estado de São Paulo (04/03642-4).

Paixão de Sousa, M.C. (2005). A Anotação da variação de grafia no Corpus Histórico do Português Tycho Brahe: Frentes abertas para estudos do léxico. Apresentação na mesa-redonda Linguística computacional e Léxico, V Encontro de Corpora. Universidade Federal de São Carlos (UFSCar). São Carlos, novembro.

Paixão de Sousa, M.C. (2006[a]) Lingüística Histórica. Em "Introdução às Ciências das Linguagem: Língua, Sociedade e Conhecimento". José Horta Nunes e Claudia Pfeiffer (Orgs.). Campinas, Pontes: 2006.

Paixão de Sousa, M.C. (2006[b]). Memórias do Texto. Revista Texto Digital. n. 2. Universidade Fedral de Santa Catarina. <http://www.textodigital.ufsc.br/num02/paixao.htm>

Paixão de Sousa, M.C. (2006[c]). Edições Críticas Eletrônicas: Fundamentos e Diretrizes. <http://www.ime.usp.br/~tycho/participants/psousa/memorias/ece>

Paixão de Sousa, M.C. (2007[a]) Digital Text: Conceptual and methodological frontiers. Em: Amelia Sanz e Dolores Romero (Orgs.): "Literatures in the Digital Era: Theory and Praxis". Cambridge, Cambridge Scholars Press.

Paixão de Sousa, M.C. (2007[b]). Lingüística de Corpus e História da Língua Portuguesa: Propostas, Resultados e Desafios, Resumo de Coordenação de Mesa Redonda. V Congresso Internacional da Associação Brasileira de Lingüística – ABRALIN. Belo Horizonte, 2 de março de 2007.

Paixão de Sousa, M.C. & Kepler, F. N. (2007). E-Dictor: Uma ferramenta para a anotação de edição especializada em XML. VII Encontro de Lingüística de Corpus (São Paulo, USP).

Paixão de Sousa, M.C.; Kepler, F.N.; Faria, P.P.F. (2010). E-Dictor: Novas perspectivas na codificação e edição

de corpora de textos históricos. In: Tania Shepherd; Tony Berber Sardinha; Marcia Veirano Pinto. (Org.). Caminhos da linguística de corpus. Campinas: Mercado de Letras, 2010.Puntoni, P. (2007). Para uma Biblioteca Brasiliana Digital. Projeto de Pesquisa sediado na BBM/USP.

Paixão de Sousa, MC., Kepler, F.N. & Faria, P (2009). E-Dictor 1.0. <http://oncoto.dyndns.org:44880/projects/edictor>

Sanchez, A. (1995). Definicion e historia de los corpus. In: A. SANCHEZ et al (org.). CUMBRE – Corpus Linguistico de Espanol Contemporaneo. Madrid: SGEL.

Trippel, T. & Paixão de Sousa, M.C. (2006). "Metadata and XML standards at work: a corpus repository of Historical Portuguese texts". Papers from the V International Conference on Language Resources and Evaluation (LREC 2006).

W3C (2008 [a]). "Extensible Markup Language". <http://www.w3.org/XML>, 10.12.2008

W3C (2008 [b]). "The Extensible Stylesheet Language Family". <http://www.w3.org/Style/XSL/>, 10.12.2008

II. Anexos: Documentação

A documentação completa dos resultados da pesquisa

está armazenada no seguinte endereço eletrônico:

http://lampiao.brasiliana.usp.br/lingua/EnsinarComPesquisa2010Relatorio

Endereços dos Relatórios Individuais

Fabiana Marcondes Ferraz

Relatório individual de Fabiana Marcondes Ferraz,
bolsista do projeto Edições Filológicas na Brasiliana Digital,
anexo do Relatório Geral, 2010

Anexo Eletrônico: http://lampiao.brasiliana.usp.br/lingua/node/104

Jáderson Johnattan Porto

Relatório individual de Jáderson Johnattan Porto,
bolsista do projeto Edições Filológicas na Brasiliana Digital,
anexo do Relatório Geral, 2010

Anexo Eletrônico: http://lampiao.brasiliana.usp.br/lingua/node/105

Leila Rosa de Oliveira

Relatório individual de Leila Rosa de Oliveira,
bolsista do projeto Edições Filológicas na Brasiliana Digital,
anexo do Relatório Geral, 2010

Anexo Eletrônico: http://lampiao.brasiliana.usp.br/lingua/node/108

Márcia Aparecida dos Santos Mendes

Relatório individual de Márcia Aparecida dos Santos Mendes,
bolsista do projeto Edições Filológicas na Brasiliana Digital,
anexo do Relatório Geral, 2010

Anexo Eletrônico: http://lampiao.brasiliana.usp.br/lingua/node/106

Thelma Tavares Dias

Relatório individual de Thelma Tavares Dias,
bolsista do projeto Edições Filológicas na Brasiliana Digital,
anexo do Relatório Geral, 2010

Anexo Eletrônico: http://lampiao.brasiliana.usp.br/lingua/node/107