Universidade de São Paulo
Pró-Reitoria de Graduação
Programa Ensinar com Pesquisa
2012

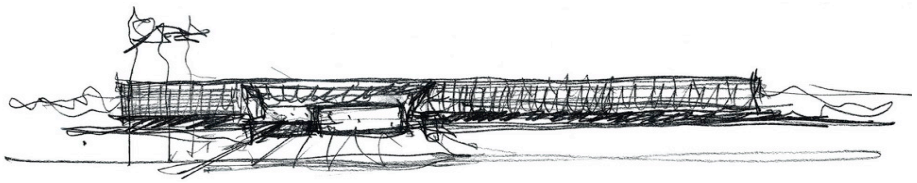
Projeto

Edições Filológicas na Brasiliana Digital: Construção de um corpus de informações ligadas

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

1. Introdução

Esta proposta de pesquisa insere-se no contexto maior dos projetos *Brasiliana USP* [1] 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[2] 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 2010. A esta nova condição física, favorecedora da consulta pública, soma-se a construção da *Biblioteca Brasiliana Digital* (BBD), que elevará ainda mais o alcance do conteúdo dos dois acervos para fins de pesquisa geral e acadêmica, sob os moldes propostos no projeto "*Por uma Biblioteca Brasiliana Digital*" (Puntoni, 2007)[3].



Maquete do Novo Edifício da Biblioteca Brasiliana USP

Desde fins de 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. O grupo inicial, coordenado pela proponente do presente projeto, foi formado por um pesquisador doutorando da área de lingüística computacional; uma pesquisadora mestranda da área de lingüística histórica; e uma pesquisadora graduanda em letras, bolsista de Iniciação Científica do CNPq . 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 presente proposta de pesquisa pretende 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. A pesquisa aproveitará e ampliará os resultados colhidos por projetos anteriores, realizados em 2010 e 2011.

---

1 cf. <http://www.brasiliana.usp.br>

2 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>

3 Projeto financiado pela *Fundação de Amparo à Pesquisa do Estado de São Paulo*, 07/597833R (Auxílio Regular à Pesquisa), do qual a proponente do presente Projeto participa como pesquisadora colaboradora.

2. Objetivos

Os objetivos deste projeto partem da 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 a etapa atual temos duas metas pontuais:

Metas para o ano de 2012:

- Refinamento dos resultados do treinamento de sofwtare de reconhecimento de caracteres para tratamento de textos em português impressos nos séculos XVI e XVII;
- Ampliação 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;
- Criação de um corpus anotado com viés semântico, com base no qual será construido um banco de dados de entidades nomeadas, com vistas à formação de um corpus de informações ligadas.

3. Justificativas

A iniciativa do grupo *Língua Brasiliana* se justifica tanto pela importância do acervo *Brasiliana USP* para o estudo da Língua Portuguesa, como pelas contribuições que os estudos lingüísticos e filológicos podem trazer para o desenvolvimento da pesquisa em torno do acervo.

O acervo *Brasiliana USP* representa um tesouro prestes a ser trazido a público, e encerra um universo documental único para os estudos voltados à história da língua portuguesa. Com o seu expressivo conjunto de livros e manuscritos (cerca de 40.000 volumes), a Brasiliana de Mindlin é considerada a mais importante coleção do gênero formada por um particular, comportando obras de literatura brasileira e portuguesa, relatos de viajantes, manuscritos históricos e literários (originais e provas tipográficas), periódicos, livros científicos e didáticos (cf. Mindlin, 2005). Entre as inúmeras preciosidades, destacam-se: uma das mais completas coleções de obras do século XVI ao XIX escritas por viajantes pelo interior do Brasil; exemplares de primeiras edições de diversos literatos do século XIX e XX; manuscritos raros, como um dos poucos exemplares conhecidos da *Notícia do Brasil* de Gabriel Soares de Souza de 1580; os primeiros exemplares da imprensa régia no Brasil no início do século XIX; e coleções raríssimas de revistas científicas do século XIX e XX. A idéia de se iniciar uma linha de pesquisa em lingüística histórica ligada à *Brasiliana* partiu em primeiro lugar da observação desse grande potencial latente da Biblioteca como documentação lingüística. Tomamos por missão impulsionar a revelação desse potencial, catalogando e preparando o acervo para futuras pesquisas sediadas na USP e em outras instituições. A presente proposta justifica-se assim, antes de tudo, pela contribuição que o acervo *Brasiliana* representa para os estudos sobre a língua portuguesa no Brasil, frente à riqueza e raridade do material ali reunido.

Observamos ainda que a iniciativa de digitalização destes documentos reforça ainda mais o valor do acervo para os estudos históricos da língua, sobretudo do ponto de vista da vertente mais atual da Lingüística Histórica, voltada para a exploração do potencial de pesquisa representado pelo adensamento da circulação de textos antigos no meio eletrônico, aproveitando e sistematizando este material como universo empírico de estudo (cf. Paixão de Sousa 2007[b]). É fato que a

pesquisa histórica sobre a língua no Brasil vem alcançando importância central nas últimas décadas, com a retomada do interesse pelo olhar diacrônico e a renovação da relevância teórica dos estudos da mudança lingüística em diferentes quadros (entre outros, com Mattos e Silva 1988, Galves & Galves 1995, Kato & Roberts 1996, Castilho 1998). Testemunha disso é a vigorosa atuação de diferentes grupos de pesquisa centrados nesta área nas diferentes instituições universitárias brasileiras, notadamente na própria Universidade de São Paulo, no âmbito do Departamento de Letras Clássicas e Vernáculas (instituição-chave neste Projeto, onde sua coordenadora atua como docente). Esse processo trouxe, como conseqüência, a intensificação do trabalho com textos antigos no Brasil (cf. Megale & Cambraia, 1999). E para boa parte das pesquisas realizadas a partir da década de 1990, a junção dos estudos diacrônicos com a prática de edição de textos passa a conferir centralidade para um terceiro campo: a Lingüística de Corpus, compreendida como o trabalho com o dado de língua em meio eletrônico. No plano internacional e brasileiro, esse crescimento vem sendo revelado pela construção de grandes corpora anotados, destacando-se, no âmbito da língua portuguesa, os corpora contemporâneos *Lácio-Web*[4] e *Corpus Dialectal do Português*[5]; e os históricos *Corpus Informatizado do Português Medieval*[6], e *Corpus Histórico do Português Tycho Brahe*[7]. O presente projeto é integrado por especialistas diretamente envolvidos nesses avanços recentes, o que permitirá a realização de pesquisas de ponta nesse sentido.

Essa intensificação do trabalho com textos antigos no meio eletrônico faz reviver hoje algumas das questões caras à filologia e à crítica textual em todos os tempos. De fato, os estudos históricos realizados com base em textos antigos dependem, antes de tudo, da garantia da fidelidade às formas originais dos textos; entretanto, para os estudos com base em corpora eletrônicos, esse pressuposto fundamental precisa conviver com requerimentos impostos pela vertente computacional e lingüística – tais sejam: a necessidade de quantidade, agilidade e automação no trabalho estatístico de seleção de dados. A conjunção dessas vertentes é o principal desafio do trabalho de edição especializada e análise lingüística de textos antigos no meio eletrônico hoje (como apontei em Paixão de Sousa, 2005, 2007[b]). O presente projeto ecoa, assim, tendências recentes da pesquisa histórica sobre a língua no Brasil: parte da concepção da Lingüística Histórica como área fundamentalmente multidisciplinar, alimentada pelos campos da História e da Lingüística (formando o campo de reflexão onde se articulam diferentes concepções de língua, e diferentes concepções de história, (cf. Paixão de Sousa 2006[a]); e orienta seu horizonte metodológico pela aliança entre conceitos e metodologias da filologia tradicional e avanços tecnológicos da área de processamento artificial da linguagem escrita (cf. Paixão de Sousa 2007[a],[b]). Buscamos, nesse contexto, uma abordagem global do texto, refletida na integração entre diferentes planos de análise, e possibilitada sobretudo pela aliança entre a filologia tradicional e a linguística computacional. A exploração dessa fronteira de pesquisa fundamenta-se em experiências anteriores que mostraram alguns caminhos promissores para respondermos aos desafios colocados para a construção de um acervo digital de textos antigos, como se verá adiante.

De outro lado, os estudos sobre a história da língua nos moldes propostos neste projeto contribuirão na formação da biblioteca digital e no desenvolvimento de pesquisas em torno do acervo a partir de outros campos. A instrumentação dos textos para fins de pesquisa lingüística pode constituir um apoio significativo para outras áreas acadêmicas interessadas na sua exploração, aprofundando a qualidade da difusão dos textos - e ao mesmo tempo, a aplicação de novas tecnologias de difusão digital pode ampliar o alcance do acervo junto ao público geral. O horizonte de metas mais geral do *Projeto*

4 Cf. <http://www.nilc.icmc.usp.br/lacioweb/>
5 Cf. <http://www.clul.ul.pt/sectores/variacao/cordialsin/projecto_cordialsin_descricao.php>
6 Cf. <http://cipm.fcsh.unl.pt/>
7 Cf. <http://www.tycho.iel.unicamp.br/~tycho/corpus/>

Brasiliana inclui o impacto da difusão e democratização do conhecimento guardado no acervo no público leitor geral e acadêmico: a digitalização do acervo é um empreendimento de fôlego, que pretende mais que simplesmente reproduzir eletronicamente o conteúdo físico da Biblioteca. Idealiza-se de fato a ampliação do alcance do acervo como parte de uma ampla reflexão sobre os impactos culturais e sociais da formação das bibliotecas digitais, concebendo-se a vocação da preservação da cultura a partir de um conceito crítico de difusão do conhecimento.

O presente projeto traz para o *Projeto Brasiliana* um campo de pesquisa avançada, representado pela inclusão da pesquisa em Linguagem Humana e Tecnologia na construção da *Biblioteca Digital,* abrindo o horizonte para o fortalecimento das relações entre as humanidades e o uso das tecnologias informáticas. Nesta proposta, a pesquisa em lingüística computacional aparece como ponto de partida e de suporte para a preparação das obras para fins de análise lingüística, instrumentando os textos para reconhecimento e extração de estruturas. Essa instrumentação gera como subprodutos as interpretações editoriais e a possibilidade de processamento computacional dos textos com variação de grafia, facilitando a leitura e as buscas por conteúdo nos textos mais antigos. Tais resultados têm importantes impactos na ampliação da difusão dos textos, tanto por favorecem a pesquisa acadêmica em diversas áreas, como por possibilitarem a democratização do público leitor geral. Acreditamos, por isso, que a preparação dos textos para fins de pesquisa lingüística contribuirá, a um tempo, para a meta de produzir material para pesquisas acadêmicas de ponta, e para a meta de democratizar a difusão dos conteúdos do acervo, delineadas na idealização do *Projeto Brasiliana.*

Importa salientar, por fim, o caráter formador da proposta ora apresentada para o aluno de graduação em letras. A metodologia de pesquisa pela qual optamos, ao conjugar os campos da filologia e lingüística histórica com o campo da ciência da computação, representa a abertura de uma linha de pesquisa tradicional da Universidade para a realidade atual da interdisciplinaridade. Para o aluno de letras, trata-se de uma oportunidade de conhecer campos tecnológicos inovadores, o que só terá 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 reflete-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. Esse aspecto da abertura ao diálogo e à interdisciplinaridade não significa, entretanto, a proposta de um trabalho panorâmico – ao contrário: o aluno de letras, nesta proposta, fará parte de um gruo de especialistas – especialistas na história da língua – que pretendem contribuir de modo sólido para o desenvolvimento do projeto coletivo da Biblioteca Digital. Pensamos, assim, estar possibilitando aos alunos uma primeira experiência de pesquisa estimulante e formadora.

4. Metodologia

4.1 Processos e objetivos na construção de uma biblioteca digital

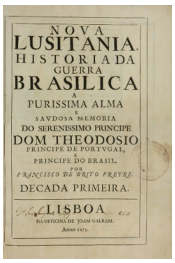
Uma biblioteca digital pode ser definida como um conjunto de documentos sistematizados para acesso mediado no meio digital, preparados para reconhecimento e extração automática de informação. Esta proposta de pesquisa parte da constatação dos desafios técnicos que se apresentam à construção de uma biblioteca digital a partir de textos antigos transportados de suportes em papel. Para compreendê-los, iremos partir do seguinte diagrama ilustrativo dos processos envolvidos na "digitalização" – isto é: na **captação de informação não-digital e geração de informação digital**:

1. Seleção no Acervo Físico

2. "Digitalização"

(i) Geração de Arquivos de Imagens

Scanner



Jpeg, png, tiff

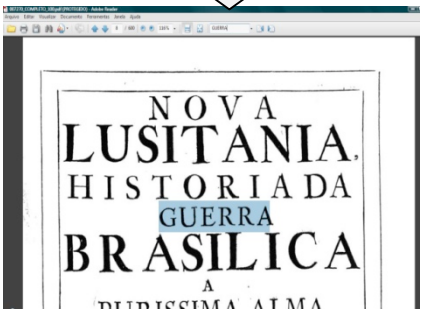
(ii) Geração de Documentos de Acesso (textos)

(a) Pós-processamento de imagens
(b) Geração de arquivos de texto
(c) Tratamento para extração de Conteúdo (*Catalogação*, *Revisão*, etc)

Txt, doc, xml, html.., pdf...

3. Interface

Extração de Informação (Conteúdos)

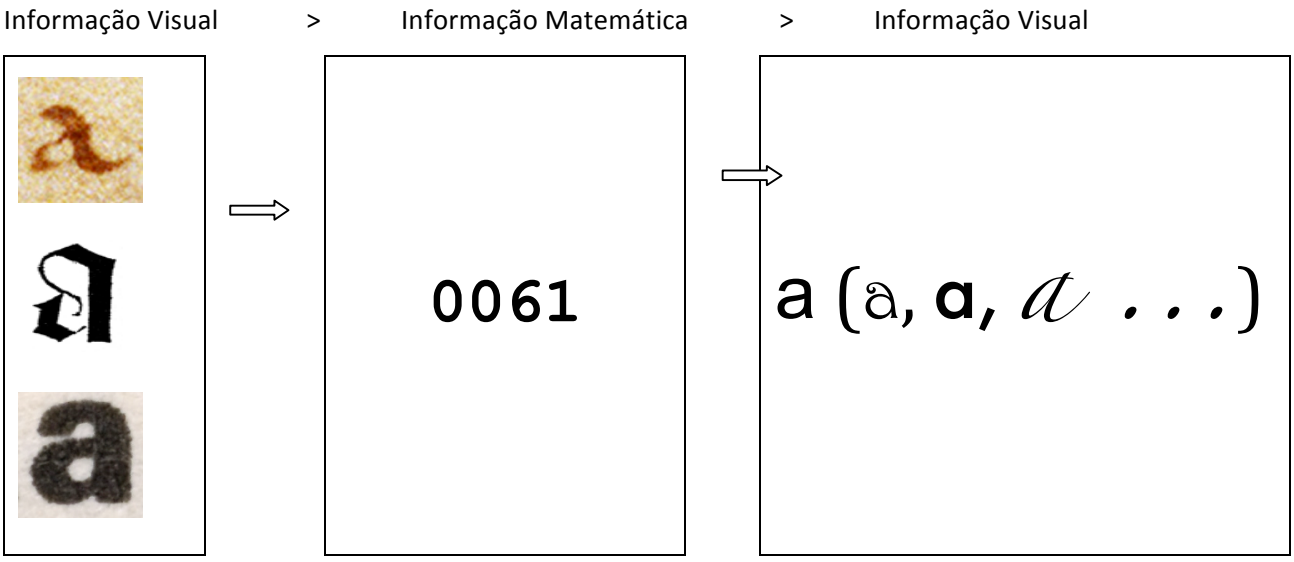


Portal Web com documentos PDF, Html...

O ideário da formação da BBD prevê a meta de oferecer acesso a reproduções as mais fiéis possíveis aos originais, e a meta de fazer do acervo um objeto da pesquisa geral e acadêmica. Entretanto, 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*. Como veremos a seguir, 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.

4.2 Desafios na construção de uma biblioteca digital de textos antigos

4.2.1 Primeiro desafio: o reconhecimento de caracteres

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.

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. Nas figuras abaixo alguns desses elementos são

ilustrados: desenhos de tipos em desuso, e caracteres em desuso (ligaduras, etc):

Caracteres da Escrita Gótica Rotunda - Séculos XV e XVI [1]

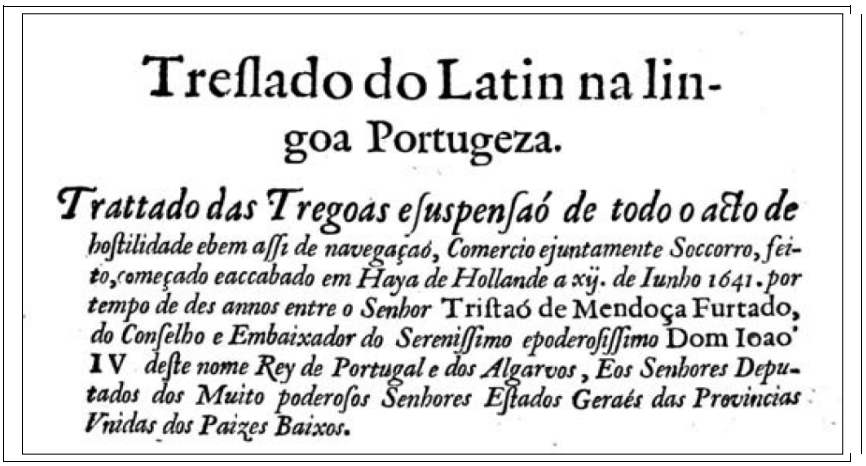
Exemplos de ligaduras da escrita humanística [2]

As principais dificuldades técnicas atualmente enfrentadas no processamento dos textos mais antigos estão exemplificadas no Quadro 4 a seguir, que mostra os resultados da aplicação de um programa de reconhecimento óptico de caracteres de última geração a um texto português do século XVII, o "*Trattado das tregoas e suspensao do todo o acto de hostilidade* …", impresso em 1641 (autor desconhecido)[8]. No Quadro, podem-se observar os diversos erros de reconhecimento resultantes da aplicação do programa.

[8] "Trattado das tregoas e suspensao de 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 (…)". BBD, <http://hdl.handle.net/1918/01936100>

Quadro 4: *Teste preliminar de performance dos resultados de processamento por OCR*

(a) Documento impresso, 1641 (Fac-simile)



Treſlado do Latin na lin-
goa Portugeza.

Trattado das Tregoas eſuspenſaó de todo o acto de hoſtilidade ebem aſſi de navegaçaó, Comercio ejuntamente Soccorro, 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 Ioao' IV *deſte nome Rey de Portugal e dos Algarvos, Eos Senhores Deputados dos Muito poderoſos Senhores Eſtados Geraés das Provincias Vnidas dos Paizes Baixos.*

(b) Saída do OCR: (Abbyy FineReader 8.0)

```
Treflado do Latin na lin-                                                    1
goa Portugeza.
Trattado das Tregoás efuspenfaó de todo o aBo de
boflilidade ebem affi de navegaçaó, Comercio ejuntammte Soccorro,fei-
totromeçado eaccabado emHayadeHollandeaxij. de lunho lóqi.por               5
tempo de des annoi entre o Senbor Triftaó de Mendoça Furtado,
do Confelho e Embaixador do Sereniffimo epoderojiffimo Dom Ioao'
defle nome }{ey de Portugal e kosAlgarvos, Eos Senhores Depu¬
tados dos Muito poderofis Senhores Efiados Geraés das Províncias
Vnidasdos Paires Baixos.                                                    10
```

(c) Saída Corrigida:

```
Treflado do Latin na lin-                                                    1
goa Portugueza.
Trattado das Tregóas eſuspenſaó de todo o acto de
hoſtilidade ebem affi de navegação, Comercio ejuntamente Soccorro, fei-
to, começado eaccabado emHaya de Hollande a xij. de Iunho  1641. por       5
tempo de des annos entre o Senhor Triftaó de Mendoça Furtado,
do Conſelho e Embaixador do Sereniffimo epoderofiffimo Dom Ioao'
IV deſte nome Rey de Portugal e dos Algaravos, Eos Senhores Depu-
tados dos Muito poderoſos Senhores Eſtados Geraés das Provincias
Vnidas dos Paízes Baixos.                                                   10
```

(d) Erros de Leitura:

1.	Treflado	*por*	Treſlado	linha 1
2.	efuspenfaó	*por*	eſuspenſaó	linha 3
3.	aBo	*por*	acto	linha 3
4.	boflilidade	*por*	hoſtilidade	linha 4
5.	affi	*por*	affi	linha 4
6.	ejuntammte	*por*	ejuntamente	linha 4
7.	fei-/totromeçado	*por*	fei-/to,começado	linha 4/5
8.	lunho	*por*	Iunho	linha 5
9.	lóqi.	*por*	1641	linha 5
10.	annoi	*por*	annos	linha 6
11.	Triftaó	*por*	Triſtaó	linha 6
12.	Confelho	*por*	Conſelho	linha 7
13.	Sereniffimo	*por*	Sereniſſimo	linha 7
14.	epoderojiffimo	*por*	epoderofiſſimo	linha 7
15.	defle	*por*	deſte	linha 8
16.	}{ey	*por*	Rey	linha 8
17.	kosAlgarvos	*por*	dos Algaravos	linha 8
18.	poderofis	*por*	poderoſos	linha 10
19.	Efiados	*por*	Eſtados	linha 10
20.	Paires	*por*	Paízes	linha 11

Neste caso, a cada linha, o programa errou duas palavras (total de 20 erros em 11 linhas) – o que representa uma média representativa dos textos desta época, segundo as pesquisas preliminares que estamos realizando junto aos arquivos de teste. A baixa performance obtida neste teste remete em primeiro lugar ao problema da grafemática. Note-se como no texto usado como exemplo no Quadro, o tipo [ſ] ("s longo") é interpretado consistentemente como [f]. Isso se explica uma vez que o programa funciona por reconhecimento de traços, e encontra, em [f], um desenho semelhante àquele elemento desconhecido (cf. [Treſlado] (d1), [Conſelho] (d12) etc). .

4.1.2 Segundo desafio: As variações de grafia

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. No mesmo trecho exemplificado no quadro acima, com 82 palavras segmentadas, 22 apresentam grafias distintas das atuais (isso, desconsiderando-se diferenças grafemáticas, como I por J, ſ por s, etc; e desconsiderando-se também problemas de segmentação, como [ejuntamente]):

> Treſlado do Latin na lin-
> goa Portugueza.
> Trattado das Tregóas eſuspenſaó de todo o acto de
> hoſtilidade ebem alſi de navegaçaó, Comercio ejuntamente Soccorro, fei-
> to, começado eaccabado emHaya 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 Ioao'
> IV deſte nome Rey de Portugal e dos Algaravos, Eos Senhores Depu-
> tados dos Muito poderoſos Senhores Eſtados Geraés das Provincias
> Vnidas dos Paízes Baixos.

As grafias antigas trazem dois problemas principais para os acervos digitais. Em primeiro lugar, note-se que, ainda que o reconhecimento de caracteres fosse perfeito (como no quadro acima), este texto ainda apresentaria dificuldades de leitura consideráveis para o leitor moderno não especializado. Este mesmo leitor 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). No texto de exemplo, itens como [Latin], [Portugeza], [Trattado], [Tregoás], [navegaçaó] tiveram seus caracteres perfeitamente reconhecidos pelo programa – mas, ainda assim, não fornecem subsídios adequados para buscas. Assim, buscas pelas palavras [Tratado]/ [navegação] não encontram como correspondentes [Trattado]/[navegaçaó], mas sim resultam em "*falha de busca*". Da mesma forma, uma busca por [Conselho] não resultaria no item [Conſelho], que poderia existir num arquivo com OCR corrigido. 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. Na seção a seguir expõem-se nossas propostas iniciais para abordar essas duas dificuldades no processamento de textos antigos portugueses.

4.3 Propostas para o enfrentamento desses desafios

Os acervos digitais de textos mais antigos têm tomado diferentes caminhos para contornar os problemas de processamento aqui exemplificados. Alguns deles optam pela produção de uma coleção de textos digitados em caracteres modernos e com ortografia atual – ou seja, tornados processáveis para buscas [9]. Outras, preocupadas com a fidelidade aos documentos originais, oferecem reproduções de seu acervo em imagens: fidedignas, mas cegas a buscas[10].

Na BBD, estamos implementando uma metodologia experimental que procura conciliar a agilidade de processamento e a preocupação com a fidedignidade, aproveitando os avanços advindos da aliança entre a filologia tradicional e a linguística computacional. Ao longo dos anos de 2010 e 2011, nosso grupo de pesquisas experimentou uma metodologia que combina o aperfeiçoamento de uma ferramenta automática de reconhecimento de caracteres a um método semi-automatico de correção e edição de textos, permitindo o enfrentamento do problema da grafemática antiga (cf. 4.2.1) e o problema da variação de grafia (4.2.2) ao mesmo tempo. Com o presente projeto, pretendemos refinar os resultados desses primeiros anos, e expandir o escopo da anotação aplicada aos textos.

4.2.1 Tratamento do problema da grafemática antiga

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"[11], e a experiência de parceria entre o projeto METAe[12] e a empresa Abbyy, resultando na construção do *Abbyy FineReader XIX*, um OCR capaz de reconhecer caracteres góticos do século XIX[13]. 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 momento, estamos lançando 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, conforme se detalha também em Paixão de Sousa (2009)[14]. Ao longo de 2010 realizamos testes sistemáticos, utilizando o módulo de "Treinamento de padrão" oferecido pelo programa (cf. 4.3).

---

9 *Projeto Vercial*, <http://alfarrabio.di.uminho.pt/vercial>; *Biblioteca Virtual Miguel de Cervantes*, <http://www.cervantesvirtual.com>; *Projeto Gutenberg*, <http://www.gutenberg.org>; *Victorian Web*, <http://www.victorianweb.org>.

10 *Biblioteca Nacional de Lisboa*, <http://bnd.bn.pt/>; *Biblioteca Nacional do Rio de Janeiro*, <http://www.bn.br>; *Caminhos do Romance no Brasil*, <http://www.iel.unicamp.br/caminhos>; *Oxford Digital Library*, <www.odl.ox.ac.uk>.

11 Cf. <http://www.digitalmedievalist.org/>

12 Cf. <http://meta-e.aib.uni-linz.ac.at/>

13 Cf. <http://www.frakturschrift.com/>

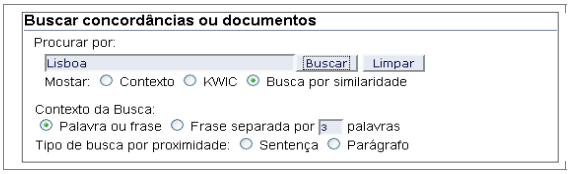
14 Cf. <https://docs.google.com/leaf?id=0B1y1cdrdSiX0ZDI5MWI0ZTMtZDI5NC00OTAyLTg2OWItOGJjMmUxMWRlMGYx&sort=name&layout=list&num=50>

4.2.2 Tratamento do problema da variação de grafias

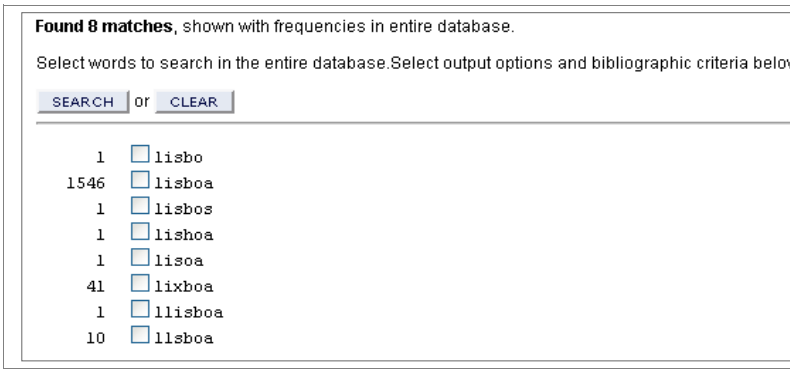
O segundo desafio no processamento computacional de textos antigos apontado mais acima foi a questão da variação de grafias. Nossas experiências anteriores de pesquisa indicam duas abordagens possíveis para solucionar esse desafio: a intervenção editorial, e a aplicação de programas automáticos de reconhecimento 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 reconhecimento e 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 reconhecimento, 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), ilustrada no Quadro 6. 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 quinhentistas impressos. De fato, o sistema se fundamenta na aplicação em textos escanerizados (submetidos, previamente, à conferência humana), que são normalizados quanto às variações grafemáticas mais imprevisíveis.

Quadro 6*: Janela do sistema de buscas do DHPB, com a ferramenta de equivalência de grafias*

(a) *Entrada de busca no mecanismo avançado do DHPB*



(b) *Resultado de busca no mecanismo avançado do DHPB*



Já a técnica da intervenção editorial consiste, fundamentalmente, na estratégia tradicional da edição filológica. O diferencial, neste caso, pode ser o uso de um ferramental eletrônico para o trabalho, nos moldes do projeto *Memórias do Texto*[15] (Paixão de Sousa 2004, 2005, 2006[b], 2006[c], 2007[b]; Trippel & Paixão de Sousa 2006). O Quadro 5 a seguir ilustra esse sistema, no qual as variações grafemáticas e de grafia são codificadas e normalizadas, preparando os textos para o tratamento posterior por ferramentas automáticas de leitura, e ao mesmo tempo preservando a integridade das informações originais do texto, pela possibilidade de gerar versões em camadas de edição. A partir do texto original, ou de uma digitalização (cf. Quadro 5, (a)[16]), 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 (c) ou modernizadas (d), acessíveis para a leitura de um público amplo e, ao mesmo tempo, para a leitura por máquinas[17].

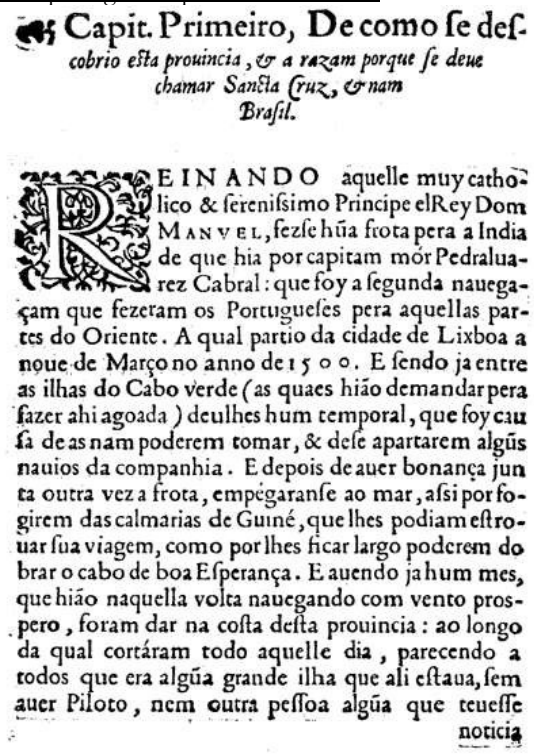
[15] Projeto de pós-doutorado (Fapesp, 04/03462-4), <http://www.ime.usp.br/~tycho/participants/psousa/memorias>

[16] *História da prouincia Sãcta Cruz que vulgarme[n]te chamamos Brasil / feita por Pero Magalhães de Gandauo.* Em Lisboa : na officina de António Gonsaluez: vendense em casa de Ioão Lopez, 1576. - 48 f. : 1 est. ; 4º (18 cm) - Assin: A-F//8. - Anselmo 709. - Faria - BN Rio de Janeiro. Biblioteca Nacional de Lisboa: <http://purl.pt/121>

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

Quadro 5: *Ilustração do sistema de Edição Eletrônica – Corpus Tycho Brahe*

(a) Texto português impresso no século 16



Capit. Primeiro, De como ſe deſcobrio eſta prouincia, & a razam porque ſe deue chamar Sancta Cruz, & nam Braſil.

REINANDO aquelle muy catholico & ſereniſsimo Principe elRey Dom MANVEL, fezſe hũa frota pera a India de que hia por capitam mór Pedraluarez Cabral: que foy a ſegunda nauegaçam que fezeram os Portugueſes pera aquellas partes do Oriente. A qual partio da cidade de Lixboa a noue de Março no anno de 1500. E ſendo ja entre as ilhas do Cabo verde (as quaes hião demandar pera fazer ahi agoada) deulhes hum temporal, que foy cauſa de as nam poderem tomar, & deſe apartarem algũs nauios da companhia. E depois de auer bonança junta outra vez a frota, empẽgaranſe ao mar, aſsi por fogirem das calmarias de Guiné, que lhes podiam eſtrouar ſua viagem, como por lhes ficar largo poderem dobrar o cabo de boa Eſperança. E auendo ja hum mes, que hião naquella volta nauegando com vento proſpero, foram dar na coſta deſta prouincia: ao longo da qual cortáram todo aquelle dia, parecendo a todos que era algũa grande ilha que ali eſtaua, ſem auer Piloto, nem outra peſſoa algũa que teueſſe noticia

(b) Arquivo XML dos anotadores [excerto]

```
<sec t="ch" author="PMG"> <sec t="title">
<ed_mark id="g_008_v_373" t="mod">
<ed id="g_008_e_373">Capítulo</ed>
<or id="g_008_o_373">Capit.</or> </ed_mark> Primeiro, De como
<ed_mark id="g_008_v_374" t="mod">
<ed id="g_008_e_374">se</ed>
<or id="g_008_o_374">fe</or> </ed_mark>
<ed_mark id="g_008_v_375" t="mod">
<ed id="g_008_e_375">descobriu<sec t="l"/> </ed>
<or id="g_008_o_375">def-<sec t="l"/>cobrio</or> </ed_mark>
esta
<ed_mark id="g_008_v_376" t="mod">
<ed id="g_008_e_376">província</ed>
<or id="g_008_o_376">prouincia</or> </ed_mark>,
<ed_mark id="g_008_v_377" t="mod">
<ed id="g_008_e_377">e</ed>
<or id="g_008_o_377">&amp;</or> </ed_mark> a
<ed_mark id="g_008_v_378" t="mod">
<ed id="g_008_e_378">razão</ed>
<or id="g_008_o_378">razam</or> </ed_mark> porque
<ed_mark id="g_008_v_379" t="mod">
<ed id="g_008_e_379">se</ed>
<or id="g_008_o_379">fe</or> </ed_mark>
<ed_mark id="g_008_v_380" t="mod">
<ed id="g_008_e_380">deve</ed>
<or id="g_008_o_380">deue</or> </ed_mark> <sec t="l"/> chamar
<ed_mark id="g_008_v_381" t="mod">
<ed id="g_008_e_381">Santa</ed>
<or id="g_008_o_381">Sancta</or> </ed_mark> Cruz, e
<ed_mark id="g_008_v_382" t="mod">
<ed id="g_008_e_382">não</ed>
<or id="g_008_o_382">nam</or> </ed_mark> <sec t="l"/>
<ed_mark id="g_008_v_383" t="mod">
<ed id="g_008_e_383">Brasil</ed>
<or id="g_008_o_383">Braſil</or> </ed_mark>.
```

(c) Edição transcrição conservadora

Capit. Primeiro, De como ſe deſcobrio esta prouincia , & a razam porque ſe deue chamar Sancta Cruz, e nam Braſil .

REINANDO aquelle muy catholico & ſereniſsimo Principe elRey Dom MANVEL, fezſe hũa frota pera a India de que hia por capitam mór Pedraluarez Cabral: que foy a ſegunda nauegaçam que fezeram os Portugueſes pera aquellas partes do Oriente. A qual partio da cidade de Lixboa a noue de Março no anno de 1500. E ſendo ja entre as ilhas do Cabo verde (as quaes hião demandar pera fazer ahi agoada) deulhes hum temporal, que foy cauſa de as nam poderem tomar, & deſe apartarem algũs nauios da companhia. E depois de auer bonança junta outra vez a frota, empegaranſe ao mar, aſsi por fogirem das calmarias de Guiné, que lhes podiam eſtrouar sua viagem, como por lhes ficar largo poderem dobrar o cabo de boa Eſperança . E auendo jahum mes , quehião naquella volta nauegando com vento proſpero , foram dar na coſta deſta prouincia : ao longo da qual cortáram todo aquelle dia, parecendo a todos que era algũa grande ilha que ali eſtaua , ſem auer Piloto, nem outra peſſoa algũa que teueſſe

noticia

(d) Edição Modernizada

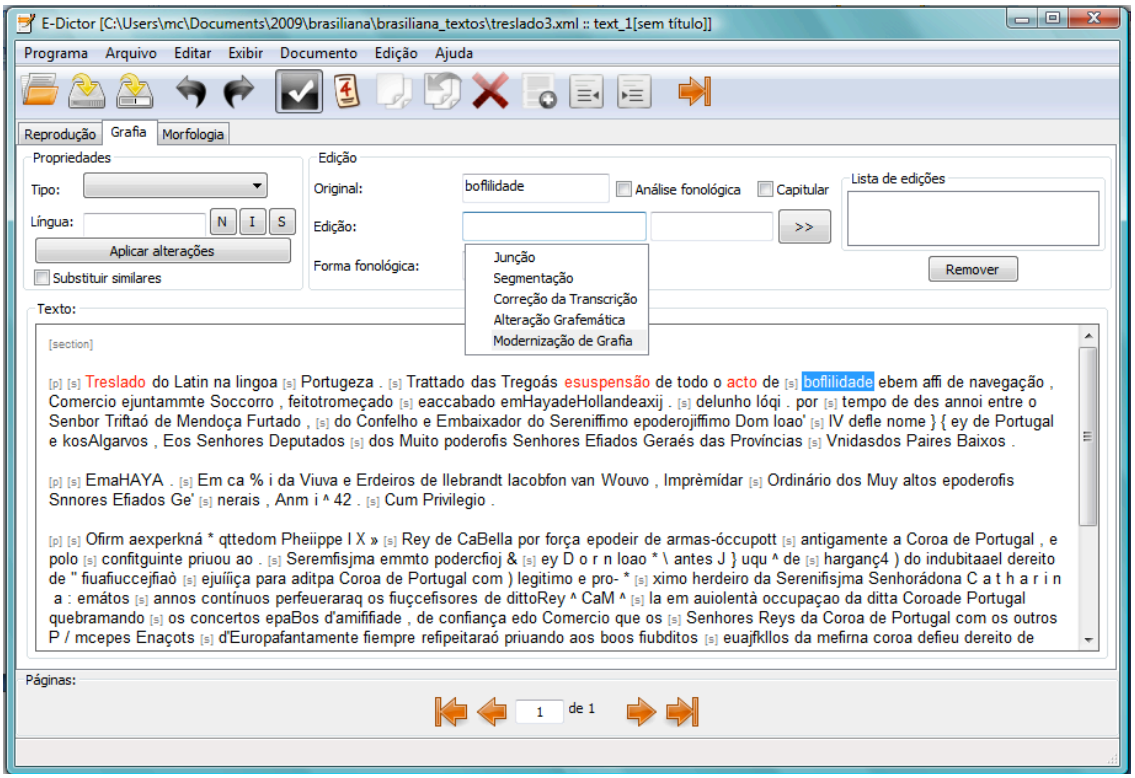
Capítulo Primeiro,
De como se descobriu esta província , e a razão porque se deve chamar Santa Cruz, e não Brasil.

Reinando aquele muito católico e sereníssimo Príncipe el-Rei Dom Manuel, fez-se uma frota para a Índia de que ia por capitão mór Pedro Álvares Cabral: que foi a segunda navegação que fizeram os Portugueses para aquelas partes do Oriente. A qual partiu da cidade de Lisboa a nove de Março no ano de 1500. E sendo já entre as ilhas do Cabo Verde (as quais iam demandar para fazer aí aguada) deu-lhes um temporal, que foi causa de as não poderem tomar, e de se apartarem alguns navios da companhia. E depois de haver bonança junta outra vez a frota, empegaram-se ao mar, assim por fugirem das calmarias de Guiné, que lhes podiam estorvar sua viagem, como por lhes ficar largo poderem dobrar o cabo de Boa Esperança. E havendo já um mês , que iam naquela volta navegando com vento próspero, foram dar na costa desta província: ao longo da qual cortaram todo aquele dia, parecendo a todos que era alguma grande ilha que ali estava, sem haver Piloto, nem outra pessoa alguma que tivesse

[*notícia*]

Esse 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 foi aplicada a uma coleção de textos portugueses dos 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*[18]. 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 da Biblioteca. O processo apresenta, entretanto, a desvantagem de demandar um grande investimento de tempo e recursos humanos. Ainda no âmbito do projeto Tycho Brahe, desenvolvemos ainda uma ferramenta semi-automática para apoiar o trabalho de edição eletrônica, o E-Dictor.



O E-Dictor é um software de anotação concebido como ferramenta auxiliar de anotação eletrônica (Paixão de Sousa, Kepler & Faria, 2009; 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.

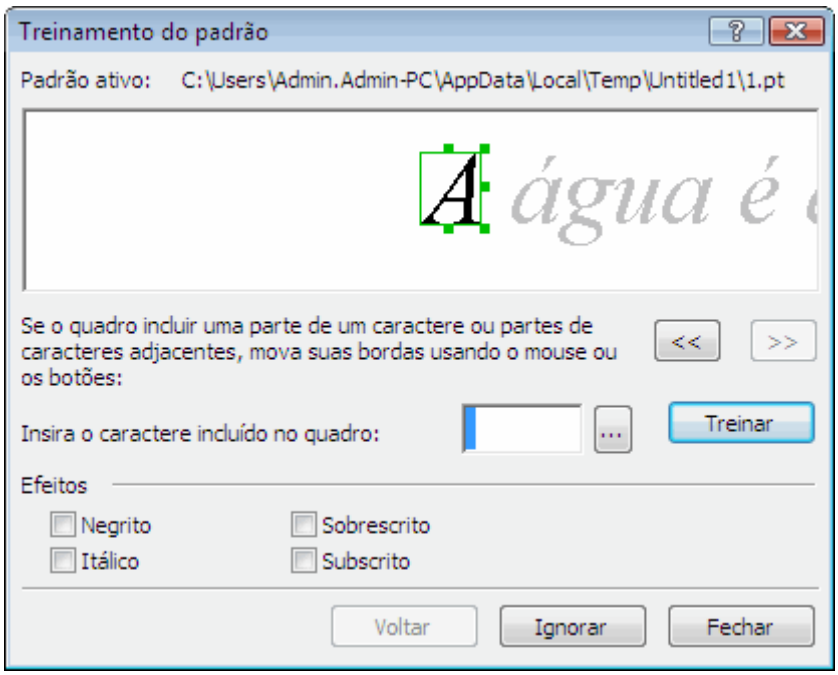
[18] Projeto Temático Fapesp, <http://www.tycho.iel.unicamp.br/~tycho>

4.4 Resultados preliminares do trabalho com o reconhecimento automático e a variação de grafia

Os dois primeiros anos de experimentos no laboratório Brasiliana Digital, conduzidos graças ao fomento do programa Ensinar com Pesquisas - 2010 e 2011, trouxeram resultados importantes que configuram a base do atual projeto. Abaixo descrevem-se as duas etapas principais desses experimentos e resumem-se seus resultados.

A) Correção dos resultados do OCR no módulo "treinamento" do Abbyy Fine Reader 9.0

O objetivo desta etapa era apurar o padrão de acerto do programa de reconhecimento automático de caracteres, dentro das possibilidades do próprio software, que oferece um modulo de "treinamento" customizável.



Neste módulo, é possível visualizar, caractere a caractere, os erros de leitura do programa, e corrigi-los um a um (em um ambiente de interface bastante amigável, como mostra a figura ao lado).

A característica mais interessante deste módulo é 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.

Ao longo do primeiro semestre de 2010, tratamos três obras com este método (num total de 34.280 palavras); as obras foram selecionadas por sua representatividade material (pela tipografia tipicamente seiscentista e setecentista) e linguística (no que remete à grafia e ao vocabulário da época):

(i) [Autor Desconhecido], 1642- Trattado das tregoas e suspensao do todo o acto de hostilidade ....
http://www.brasiliana.usp.br/bbd/handle/1918/01936100#page/1/mode/1up

(ii) Britto Freire, Francisco de , 1675 - Nova Lusitania, historia da guerra Brasilica [...]
http://www.brasiliana.usp.br/bbd/handle/1918/00727000#page/1/mode/1up

(ii) Cunha, Luís Antonio Rosado da, 1747 - Relação da entrada que fez o excellentissimo, e reverendissimo senhor D. Fr. Antonio do Desterro Malheyro ....
http://www.brasiliana.usp.br/bbd/handle/1918/03908100#page/1/mode/1up

Chegamos a uma melhora sensível entre a primeira leitura automática (sem treinamento) e a leitura final (com treinamento): a taxa geral de acertos passou de **59%** para **86%**. Note-se, entretanto, que não chegamos a um patamar de leitura excelente (compare-se a taxa obtida, 86%, com a taxa para textos modernos, 95%). Os melhores resultados do treinamento ainda precisam ser aprimorados, para além dos limites que o software apresentou. Para solucionar isso

passamos para a etapa (B), descrita abaixo.

B) Correção dos resultados do OCR via edição humana, com auxílio do E-Dictor

De posse das melhores leituras produzidas pelo Abbyy 9.0, procedemos à sua correção por meio da ferramenta E-Dictor. Além dos textos já citados, incluímos nesta etapa o resultado da leitura automática das seguintes obras (escolhidas pelos mesmos critérios já citados; neste caso, são obras ainda não lançados no site):

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

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

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

Essa etapa produziu três resultados: primeiro, e mais evidente, os seis textos acima citados com a sequência de caracteres 100% precisa (que serão 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) - cf. Quadro 5. mais acima .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 experimentos conduzidos 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.

Em 2011, de posse desses primeiros resultados, escolhemos um novo grupo de textos, e aplicamos a eles a anotação desenvolvida no ano anterior. Além disso, experimentamos, nesses textos, um trabalho mais fino de anotação, voltado para a identificação e especificação de entidades nomeadas. Comentaremos esse viés dos experimentos na seção a seguir.

4.5 Trabalho Preliminar com as Entidades Nomeadas

. "Entidades nomeadas", numa definição simples, são entidades concretas ou abstratas que possuem um nome próprio; a relevância de sua identificação em um conjunto eletronicamente trabalhado de textos é a possibilidade que esta identificação abre para a realização de buscas semânticas. No universo atual da linguística computacional, existem inúmeros projetos voltados para este tipo de trabalho (), inclusive no que respeita corpora em língua portuguesa (). Para textos mais antigos em português, entretanto, nossos experimentos no âmbito do presente projeto são inéditos.

A relevância do trabalho com as entidades nomeadas nos textos do acervo Brasiliana USP está relacionada ao valor históricos do acervo, que compreende um conjunto extremamente valioso de obras historiográficas sobre o Brasil. Esse conjunto de obras tem um imenso potencial de pesquisa em diferentes áreas (muito particularmente, a historiografia), e a formação de um sistema de buscas inteligentes, ou buscas semânticas, seria um empreendimento de grande impacto científico. Por outro lado, esses textos mais antigos, conforme já ressaltamos, encerram desafios importantes para o processamento automático. No presente projeto, queremos aproveitar os avanços técnicos alcançados nos projetos anteriores quanto ao enfrentamento desse desafio, e ampliar o potencial de processamento automático dos textos aplicando a anotação semântica.

Para se ter uma idéia do tipo de trabalho que se pode fazer nesse campo, comentamos brevemente alguns experimentos pontuais em andamento neste ano de 2011: às crônicas e panfletos editados em 2010, estamos acrescentando a anotação de duas classes de entidades nomeadas - nomes de pessoas e de lugares. Além disso, escolhemos trabalhar, também, neste ano, com o núcleo documental da Tipografia do Arco do Cego, grupo de compreende obras de grande valor científico - em particular, tratados oitocentistas de zoologia e botânica; nessas obras, estamos anotando a nomenclatura científica. O objetivo dessa anotação é propiciar, futuramente, uma busca inteligente que possibilite ao pesquisador encontrar nos textos do acervo personagens históricos, localidades e espécies naturais - e não apenas encontrá-los, como relacioná-los entre si. O potencial dessa anotação para a pesquisa acadêmica é extenso, e explorar esse potencial é uma das metas do presente projeto.

4.6 Perspectivas

A perspectiva de explorar o potencial de pesquisa abertos pelos primeiros experimentos com anotação de entidades nomeadas no Acervo Brasiliana motiva a proposta do presente projeto. O trabalho proposto aqui aos alunos de graduação consiste na elaboração de edições filológicas eletrônicas, seguindo os fundamentos do que foi apresentado anteriormente, com o uso do E-Dictor. Esse trabalho irá colaborar para o desenvolvimento do dicionário de grafias antigas (já iniciado pelos projetos anteriores), e irá formar um banco de dados de entidades nomeadas, nos moldes dos experimentos preliminares de 2011. O banco de dados assim formado contituirá a base para a formação de um sistema de buscas semânticas no Acervo - noutros termos: um corpus de informações ligadas.

5. Planejamento

5.1 Plano de Trabalho

5.1.1 Capacitação técnica e teórica

O trabalho envolverá dois âmbitos de capacitação: a capacitação técnica para o uso das ferramentas computacionais, e a capacitação teórica no campo da filologia e crítica textual. No primeiro âmbito, a capacitação se dará nos dois primeiros meses de bolsa. No segundo âmbito, a capacitação será continuada, compreendendo do primeiro ao décimo mês. Essa será uma atividade de grupo, reunindo a coordenadora e todos os bolsistas, em reuniões quinzenais para discussão de uma bibliografia pré-selecionada relevante.

5.1.2 Trabalho de edição e aplicação das ferramentas

O trabalho de edição dos textos, com a aplicação das ferramentas computacionais, envolve dois momentos, conforme foi exposto na seção 4: o treinamento dos padrões de reconhecimento do software de OCR e o uso da ferramenta E-Dictor para edição dos textos escanerizados (saliente-se que cada bolsista terá a oportunuidade de trabalhar nos dois sistemas); e a aplicação da anotação de entidades nomeadas aos textos em edição.

5.1.3 Avaliação e apresentação de resultados

Além dos encontros quinzenais em torno das leituras dos textos, o grupo terá oportunidade constante de interação, graças à concentração dos trabalhos no Laboratório da Brasiliana. Estão programadas ainda duas jornadas de avaliação e exame dos resultados do trabalho - uma ao final da fase de treinamento do OCR, outra ao final da fase de edição. As jornadas terão uma semana de duração, tempo ao longo do qual o grupo de bolsistas e a coordenadora se reunirão intensivamente para elaborar relatórios internos dos progressos dos trabalhos. Nessas oportunidades, estará em jogo sobretudo o desempenho das ferramentas como auxiliares do trabalho filológico, os eventuais problemas apresentados e possibilidades de desenvolvimento técnicos futuros. Ao final do ano de trabalhos, está planejado um seminário, no qual os bolsistas poderão apresentar os resultados do grupo publicamente, no ambiente do Laboratório da Brasiliana.

5.2 Cronograma

	março	abril	maio	junho	julho	agosto	setembro	outubro	novembro	dezembro	janeiro	fevereiro
Capacitação inicial para o uso das ferramentas	X	X										
Aplicação de OCR de treinamento sobre os textos			X	X	X	X						
Edição e anotação de textos com o programa E-dictor							X	X	X	X		
Jornada de avaliação e exame dos resultados iniciais						X						
Jornada de avaliação e exame dos resultados finais											X	
Elaboração do relatório final										X	X	
Seminário de apresentação dos resultados finais												X
Grupo de leitura (Filologia e História da Língua Portuguesa)	X	X	X	X	X	X	X	X	X	X		

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>

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. (2009). O Processamento automático de textos antigos: Desafios e Experiências. Workshop de Linguística de Corpus do Projeto Para a História do Português Brasileiro (PHPB)..

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

Puntoni, P. (2007). Para uma Biblioteca Brasiliana Digital. Projeto de Pesquisa sediado na BBM/USP.

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