<REVISTA TEXTO DIGITAL>

ISSN 1807-9288
- ano 2 n.1 2006 -
http://www.textodigital.ufsc.br

MEMÓRIAS DO TEXTO

MEMORIES OF THE TEXT

Maria Clara Paixão de Sousa
Doutora em Lingüística pela Unicamp
Universidade Estadual de Campinas
mariaclara.ps@gmail.com

0. Apresentação

Este artigo relata os primeiros resultados do desenvolvimento de uma técnica de processamento de textos que permite controlar sucessivas etapas de edição, com o objetivo de aproximar a edição especializada de textos e os desenvolvimentos recentes das tecnologias de processamento em meio digital.

Esta técnica surgiu como decorrência da experiência na construção do *Corpus Histórico do Português Anotado Tycho Brahe*. Desde o início da formação deste corpus eletrônico, em 1998, enfrentamos desafios técnicos e teóricos resultantes da multiplicidade de objetivos potenciais do uso do material trabalhado. A meta inicial da construção desse corpus foi a de preparar textos para serem automaticamente analisados por ferramentas computacionais, com o objetivo de possibilitar buscas especializadas (por classes de palavra e por estrutura sintática). Nessa vertente, desenvolvemos um analisador de classes de palavras, e estamos desenvolvendo um analisador sintático para o português.

A meta da análise lingüística automática tem impactos na preparação dos textos eletrônicos, por exemplo ao impor limitações à variação de grafia e à codificação de caracteres. De outro lado, nos deparamos com a questão dos parâmetros filológicos relevantes a serem respeitados na seleção e classificação dos textos incluídos no corpus.

Inicialmente, o corpus incluía poucas edições de época, e diversas re-edições dos textos originalmente escritos nos 1500, 1600 e 1700. Com o decorrer das pesquisas sobre as mudanças lingüísticas com base nesses textos, a importância do uso de edições originais passou a se revelar para o grupo. De fato, a depender dos temas de pesquisa, pequenas alterações eventualmente trazidas pelas subseqüentes edições podem ter conseqüências importantes na análise lingüística do texto. Idealmente, o corpus seria formado apenas por edições de época.

Na confluência dessas duas vertentes (computacional e filológica) encontramos o seguinte desafio: desenvolver um tratamento dos textos que permitisse o processamento automático sem perder a possibilidade do uso de material filologicamente consistente.

O projeto *Memórias do Texto* (Paixão de Sousa, 2004) partiu desse desafio, e desenvolveu uma técnica de controle de edições cujos primeiros resultados já podem ser apresentados. O novo sistema, por sua vez, traz alguns novos desafios, e abre possibilidades interessantes para o trabalho com o texto digital em um plano mais amplo, conforme pretendo mostrar aqui.

1. O sistema de controle de edições: Aspectos técnicos

O sistema de anotação dos nossos textos teve como primeiro objetivo codificá-los digitalmente de modo a produzir versões adequadas para o tratamento computacional sem perder a qualidade filológica. O projeto se iniciou pelo desenvolvimento de uma técnica de anotação que permitisse preparar textos digitalizados a partir de edições de época para o uso no Corpus. A Figura 1 abaixo mostra um exemplo representativo de nosso material típico de partida, na forma de um fac-simile digitalizado de um texto impresso no século 16, a "*História da Província de Santa Cruz*", de Pero Magalhães de Gandavo:

AO MVITO ILLVSTRE SENHOR
DOM LIONIS PEREIRA,
Epiſtola de Pero de
Magalhães.

ESTE pequeno ſeruiço (muito illuſtre ſenhor) que offereço a V.M. das premicias de meu fraco entendimento, poderá nalgũa maneira conhecer os deſejos que tenho de pagar com minha poſsibilidade algũa parte do muito que ſe deue á inclita fama de voſſo heroyco nome. E iſto aſsi pelo merecimẽto do nobiliſsimo ſangue & clara progenie donde traz ſua origem, como pelos tropheos das grandes victorias,& caſos bem afortunados que lhe hão ſuccedido neſſas partes do Oriente em que Deos o quis fauorecer com tam larga mão, que nam cuido ſer toda minha vida baſtante pera ſatisfazer á menor parte de ſeus louuores. E como todas eſtas razões me ponham em tanta obrigaçam, & eu entenda que outra nenhũa couſa deue ſer mais aceita a peſſoas de altos animos que a liçam das eſcrituras,per cujos meyos ſe alcançam os ſegredos de todas as ſciencias,& os homẽs vém a illuſtrar ſeus nomes & perpetualos na terra com fama immortal, determiney escolher a V. M. entre os mais ſenhores da terra, & dedicarlhe eſta breue hiſtoria. A qual eſpero que folgue de ver có attençam & receberma benignamente debaixo de ſeu empara: aſsi por ſer couſa noua, & eu a eſcreuer como teſtemunha de viſta: como por ſaber quam particular affeiçam V. M. tem ás couſas do ingenho, & que por eſta cauſa lhe nam ſera menos aceito o exercicio das eſcrituras, que o das armas. Poronde com muita razam fauorecido deſta confiança poſſa ſeguramente ſair a luz com eſta pequena empreſa & diuulgala pela terra ſem nenhum receo,tendo por defenſor della a V. M. Cuja muito illuſtre peſſoa noſſo Senhor guarde & acrecẽte ſua vida & eſtado por longos & felicis annos.

Figura 1: Documento Original (fac-simile):

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. <http://purl.pt/121>.

Nossa primeira tarefa, diante deste tipo de material, é transcrever o texto, para em seguida modernizar sua grafia – de modo a possibilitar o posterior processamento pelas ferramentas automáticas de análise lingüística (analisador morfológico e sintático). A idéia central era fazer com que esta manipulação do texto fosse controlada, de modo a garantir a recuperabilidade das formas originais. Para isto foi desenvolvida uma técnica de anotação baseada em linguagem XML (*eXtended Markup Language*), que vem sendo implementada desde fins de 2004, seguindo diretrizes apresentadas em Paixão de Sousa e Trippel (2004, 2006) e documentadas em um Manual livremente disponível em rede (Paixão de Sousa, 2006). Atualmente, temos quatro novos textos nesse formato, um deles inteiramente preparado, e três em processo de preparação.

O princípio central por trás da técnica de anotação de edição é a de codificar no texto estruturas variantes, de modo a possibilitar o controle ou mapeamento das intervenções realizadas nos documentos. Para cada intervenção em um item original, cria-se e anota-se uma estrutura variante composta pelo item original e o item inserido pelo editor. As estruturas variantes e seus componentes são numerados por um *script* identificador automático, que processa o texto depois que todas as intervenções foram anotadas.

A Figura 2 baixo mostra um trecho da transcrição do texto mostrado sob forma de reprodução digital na Figura 1 acima, com algumas estruturas variantes anotadas (os itens originais são anotados como <or>; os itens editados, como <ed>; e a estrutura variante assim formada, como <v>):

```
E
como
todas
<v    id="g_008_v_48" type="mod">
<ed   id="g_008_e_48">estas</ed>
<or   id="g_001_o_48">eſtas</or></v>
razões
me
ponham
em<nl/>
tanta
<v    id="g_008_v_49" type="mod">
<ed   id="g_008_e_49">obrigação</ed>
<or   id="g_001_o_49">obrigaçam</or></v>,
<v    id="g_008_v_50" type="mod">
<ed   id="g_008_e_50">e</ed>
<or   id="g_001_o_50">&amp;amp;</or></v>
eu
entenda
que
outra
<v    id="g_008_v_51" type="mod">
<ed   id="g_008_e_51">nenhuma</ed>
<or   id="g_001_o_51">nenhũa</or></v>
<v    id="g_008_v_52" type="mod">
<ed   id="g_008_e_52">cousa</ed>
<or   id="g_001_o_52">couſa</or></v>
<v    id="g_008_v_53" type="mod">
<ed   id="g_008_e_53">deve</ed>
<or   id="g_001_o_53">deue</or></v>
<v    id="g_008_v_54" type="mod">
<ed   id="g_008_e_54">ser</ed>
<or   id="g_001_o_54">ſer</or></v><nl/>
mais
aceita
a
<v    id="g_008_v_55" type="mod">
<ed   id="g_008_e_55">pessoas</ed>
<or   id="g_001_o_55">peſſoas</or></v>
de
```

```
altos
<v    id="g_008_v_56" type="mod">
<ed   id="g_008_e_56">ânimos</ed>
<or   id="g_001_o_56">animos</or></v>
que
a
<v    id="g_008_v_57" type="mod">
<ed   id="g_008_e_57">lição</ed>
<or   id="g_001_o_57">liçam</or></v>
das
<v    id="g_008_v_58" type="mod">
<ed   id="g_008_e_58">escrituras</ed>
<or   id="g_001_o_58">eſcrituras</or></v>,
```

Figura 2: Anotação de Estruturas Variantes

Os textos assim anotados podem ser oferecidos aos usuários do Corpus sob diferentes formas. Para isso, aplica-se ao documento-fonte anotado em XML programações em XSLT (*eXtended Stylesheet Transformation Language*). As programações são gravadas no servidor que armazena o Corpus, e ativadas por *scripts* pelo usuário para gerar, instantaneamente, versões com a transcrição do texto original ou a edição modernizada, bem como léxicos das intervenções realizadas. As figuras a seguir mostram exemplos destas diferentes versões.

A figura 3 abaixo mostra a primeira versão do texto-base, na qual se preservam os vocábulos e grafias originais transcritos:

AO MVITO ILLVSTRE SENHOR
DOM LIONIS PEREIRA,
Epiſtola de Pero de Magalhães.

N E S T E pequeno ſeruiço
(muito illuſtre ſenhor) que offere-
ço a V.M. das premicias de meu fra
co entendimento, poderá nalgũa
maneira conhecer os deſejos que
tenho de pagar com minha poſsibi
lidade algũa parte do muito queſe
deue á inclita fama de voſſo heroy-
co nome. E iſtoaſsi pelo mereci-
mẽto do nobiliſsimo ſangue & cla
ra progenie donde traz ſua origem,
como pelos tropheos das grandes
victorias , & caſos bem afortunados que lhe hão ſuccedido neſſas par
tes do Oriente em que Deos o quis fauorecer com tam larga mão,
que nam cuido ſer toda minha vida baſtante pera ſatisfazer á menor
parte de ſeus louuores . E como todas eſtas razões me ponham em
tanta obrigaçam , & eu entenda que outra nenhũa couſa deue ſer
mais aceita a peſſoas de altos animos que a liçam das eſcrituras , per
cujos meyos ſe alcançam os ſegredos de todas as ſciencias , & os ho-
mẽs vém a illuſtrar ſeus nomes & perpetualos na terra com fama im
mortal , determiney escolher a V.M. entre os mais ſenhores da ter
ra , & dedicarlhe eſta breue hiſtoria . A qual eſpero que folgue de
ver cõ attençam & receberma benignamente debaixo de ſeu empa-

ro : aſsi por ſer couſa noua , & eu a eſcreuer como teſtemunha de vi-
ſta : como por ſaber quam particular affeiçam V.M. tem ás couſas
do ingenho , & que por eſta causa lhe nam ſera menos aceito o exer
cicio das eſcrituras , que o das armas. Poronde com muita razam
fauorecido deſta confiança poſſa ſeguramente ſair a luz com eſta pe
quena empreſa & divulgala pela terra ſem nenhum receo , ten-
do por defenſor della a V.M. Cuja muito illuſtre peſ-
ſoa noſſo Senhor guarde & acrecẽte ſua
vida & estado por longos &
felicis annos .

Figura 3: Versão Trasncrição do Original

A figura 4 abaixo mostra uma segunda versão do texto-base, na qual se visualiza o texto na forma editada:

AO MUITO ILUSTRE SENHOR
DOM LIONIS PEREIRA,
Epístola de Pero de Magalhães.

[43] NESTE pequeno serviço
(muito ilustre senhor) que ofereço
a Vossa Mercê das primícias de meu fraco
entendimento, poderá nalguma
maneira conhecer os desejos que
tenho de pagar com minha possibilidade
alguma parte do muito que se
deve à ínclita fama de vosso heróico
nome. [44] E isto assim pelo merecimento
do nobilíssimo sangue e clara
progenie donde traz sua origem,
como pelos troféus das grandes
vitórias, e casos bem afortunados que lhe hão sucedido
nessas partes
do Oriente em que Deus o quis favorecer com tão larga mão,
que não cuido ser toda minha vida bastante pera satisfazer à
menor
parte de seus louvores. [45] E como todas estas razões me
ponham em
tanta obrigação, e eu entenda que outra nenhuma cousa deve
ser
mais aceita a pessoas de altos ânimos que a lição das
escrituras, por
cujos meios se alcançam os segredos de todas as ciências, e
os homens
vem a ilustrar seus nomes e perpetuá-los na terra com fama
imortal,
determinei escolher a Vossa Mercê entre os mais senhores da
terra,
e dedicar-lhe esta breve história. [46] A qual espero que
folgue de
ver com atenção e receber-ma benignamente debaixo de seu
amparo:

assim por ser coisa nova, e eu a escrever como testemunha de vista:
como por saber quão particular afeição Vossa Mercê tem às coisas
do engenho, e que por esta causa lhe não será menos aceito o exercício
das escrituras, que o das armas. [47] Por onde com muita razão
favorecido desta confiança possa seguramente sair a luz com esta pequena
empresa e divulgá-la pela terra sem nenhum receio, tendo
por defensor dela a Vossa Mercê Cuja muito ilustre pessoa
nosso Senhor guarde e acrescente sua
vida e estado por longos e
felizes anos.

Figura 4: Versão Texto Editado

A figura 5 abaixo mostra uma terceira versão do texto-base, na qual se listam as variantes (itens originais e suas edições):

Item Editado	Item Original	Identificação
MUITO	MVITO	[g_008_v_1]
ILUSTRE	ILLVSTRE	[g_008_v_2]
Epístola	Epiſtola	[g_008_v_3]
NESTE	N E S T E	[g_008_v_4]
serviço	ſeruiço	[g_008_v_5]
ilustre	illuſtre	[g_008_v_6]
senhor	ſenhor	[g_008_v_7]
ofereço	offere-ço	[g_008_v_8]
Vossa Mercê	V.M.	[g_008_v_9]
primícias	premicias	[g_008_v_10]
fraco	fra-co	[g_008_v_11]
nalguma	nalgũa	[g_008_v_12]
desejos	deſejos	[g_008_v_13]
possibilidade	poſsibilidade	[g_008_v_14]
alguma	algũa	[g_008_v_15]
que ſe	queſe	[g_008_v_16]
que se	que ſe	[g_008_v_17]
deve	deue	[g_008_v_18]
à	á	[g_008_v_19]
ínclita	inclita	[g_008_v_20]
vosso	voſſo	[g_008_v_21]
heróico	heroy-co	[g_008_v_22]
isto assim	iſto aſsi	[g_008_v_23]
iſto aſsi	iſtoaſsi	[g_008_v_24]
merecimento	mereci-mẽto	[g_008_v_25]
nobilíssimo	nobiliſsimo	[g_008_v_26]

Item Editado	Item Original	Identificação
sangue	ſangue	[g_008_v_27]
e	&	[g_008_v_28]
clara	cla-ra	[g_008_v_29]
sua	ſua	[g_008_v_30]
troféus	tropheos	[g_008_v_31]
vitórias	victorias	[g_008_v_32]
e	&	[g_008_v_33]
casos	caſos	[g_008_v_34]
sucedido	ſuccedido	[g_008_v_35]
nessas	neſſas	[g_008_v_36]
partes	par-tes	[g_008_v_37]
Deus	Deos	[g_008_v_38]
favorecer	fauorecer	[g_008_v_39]
tão	tam	[g_008_v_40]
não	nam	[g_008_v_41]

Figura 5: Versão Léxico de Edições

Esse novo sistema atende plenamente alguns dos objetivos lançados pelo nosso projeto. Torna-se possível agora prepararmos textos digitais a partir de impressões realizadas entre os séculos 16 a 19, sem prejuízo da agilidade das ferramentas de análise lingüística automática. O uso da anotação XML possibilitou a codificação completa dos textos tanto no que toca a seus cabeçalhos (para fins de catálogo, classificação e busca) como no que toca a estrutura dos textos (ou seja, a codificação eletrônica das estruturas gráficas, como paginação, paragrafação, etc.). Todos os textos do Corpus foram integrados ao novo sistema e podem ser acessados por meio de um Catálogo produzido com a aplicação da linguagem de busca X-Query à anotação XML (cf. detalhamento técnico em Paixão de Sousa e Trippel, 2006).

Entretanto, uma outra vertente do projeto Memórias do Texto tem início justamente agora, como se expõe a seguir.

2. As Memórias do Texto

Como resultado dessa técnica de anotação, podemos rastrear as intervenções sofridas pelos textos no decorrer de suas sucessivas edições. Com isso, o projeto *Memórias do Texto* volta-se agora para sua segunda vertente, que consiste em uma reflexão sobre a história editorial dos textos do corpus. As perguntas propostas de o início eram as seguintes:

- Em que medida as intervenções sofridas pelos textos no decorrer de suas sucessivas edições transformaram a linguagem dos textos?

- Como o exame das formas originais dos textos pode refletir nos estudos históricos da língua (*por exemplo, interferindo na qualidade dos textos como fonte para o estudo de mudanças sintáticas, morfo-sintáticas, fonológicas, etc.*)?

- Como o exame das intervenções realizadas nos textos em cada momento pode refletir nos estudos históricos da língua (*por exemplo, quanto à construção de um imaginário de linguagem culta ou de normatização, etc.*)?

Para enfrentar essa reflexão, aproveitaremos a técnica de edição controlada já desenvolvida para realizar cotejos

exaustivos de algumas obras selecionadas. Tomando-se um documento de partida que tenha sido trabalhado por mais de um editor, o sistema de codificação e identificação das variantes torna possível gerar versões distintas correspondentes ao produto de cada editor, incluindo listas paralelas onde se mostrem as intervenções de cada um. Para as pesquisas no campo da lingüística, isso abre algumas vertentes interessantes, como a de investigar as diferentes tendências da interferência editorial (por exemplo, no sentido de correção normativa) em diferentes momentos históricos.

O sistema de edição controlada pode permitir ainda outras vertentes para a exploração desse corpus digital - por exemplo, no plano da crítica genética, e da história editorial. Nesse sentido, iniciamos recentemente uma parceria com o grupo de trabalho do projeto "*Caminhos do Romance no Brasil*". Este trabalho conjunto se dará tanto no sentido do aproveitamento das tecnologias de texto do projeto *Memórias do Texto* pela biblioteca eletrônica do projeto *Caminhos*, como pela possibilidade do aprofundamento da pesquisa sobre a história editorial dos textos do corpus eletrônico graças à rica reflexão que tem lugar no projeto *Caminhos*. Acreditamos que a parceria entre as áreas de lingüística computacional e estudos literários pode apresentar desenvolvimentos interessantes para ambos os lados.

Referências

PAIXÃO DE SOUSA, M.C. (2004). *Memórias do Texto: Aspectos tecnológicos na construção de um corpus eletrônico do português*.<http://www.ime.usp.br/~tycho/participants/psousa/memorias/>

PAIXÃO DE SOUSA, M.C. (2006). *Manual de Preparação de Textos para o Corpus Histórico do Português Tycho Brahe.* <http://www.ime.usp.br/~tycho/corpus/manual/prep/manual_completo.html>

PAIXÃO DE SOUSA, M.C. & TRIPPEL, T. (2006). *Metadata and XML standards at work: a corpus repository of Historical Portuguese texts*. V International Conference on Language Resources and Evaluation (LREC 2006), Gênova, maio de 2006. <http://www.ime.usp.br/~tycho/participants/psousa/2006/lrec_psousa_trippel.pdf>

PAIXÃO DE SOUSA, M.C. & TRIPPEL, T. (2004). *Single source processing of historic corpora for diverse uses*. Association for Literary and Linguistic Computing (ALLC) – Annual Conference, 2004, Universidade de Gothenburg, fevereiro de 2004. <http://www.ime.usp.br/~tycho/participants/psousa/2004/allc.pdf>

CORPUS HISTÓRICO DO PORTUGUÊS ANOTADO TYCHO BRAHE. <http://www.ime.usp.br/tycho/corpus>

CAMINHOS DO ROMANCE NO BRASIL – Biblioteca Eletrônica. <http://www.caminhosdoromance.iel.unicamp.br/>.