ISSN: 2175-5493

IX COLÓQUIO DO MUSEU PEDAGÓGICO
5 a 7 de outubro de 2011

ORTOGRAFIA OITOCENTISTA: MANUSCRITOS DO SÉCULO XIX DO CORPUS DOVIC ANALISADOS POR UM SOFTWARE DE ETIQUETAGEM MORFOLÓGICA

Fabio Gomes Novais*
(UESB)

Jorge Viana Santos**
(UESB)

Cristiane Namiuti Temponi***
(UESB)

RESUMO
A circulação de textos em ambientes digitais tem se desenvolvido, o que tem reforçado a relevante busca de aperfeiçoamento de ferramentas computacionais de análise. Uma destas é o programa de edição E-Dictor, que possui internamente um etiquetador digital, que classifica morfologicamente palavras considerando seu contexto. Dentro do projeto Memória conquistense propôs-se avaliar o desempenho desse programa quando aplicado à edição de textos originalmente manuscritos como os que integram o corpus DOViC. Nesse sentido, esse trabalho tem como objetivo contribuir na construção do corpus DOViC, investigando o desempenho da anotação morfológica automática consoante aos aspectos da grafia.

PALAVRAS-CHAVE: Etiquetador, Morfologia, Texto digital.

* Graduando em Letras – UESB; bolsista de iniciação científica CNPq. E-mail: fahbio_gommes@hotmail.com
** Orientador. Professor do Departamento de Estudos Lingüísticos e Literários – UESB; Doutor em Lingüística pela UNICAMP. E-mail: jorge-viana@uol.com.br
*** Co-orientadora.Professora do Departamento de Estudos Lingüísticos e Literários – UESB; Doutora em Lingüística pela UNICAMP. E-mail: cristianenamiuti@gmail.com.

ISSN: 2175-5493

IX COLÓQUIO DO MUSEU PEDAGÓGICO
5 a 7 de outubro de 2011

INTRODUÇÃO

Há uma ampla massa de textos antigos impressos ou não a serem editados, ou seja, transformados em textos legíveis e computacionalmente manipuláveis, a fim de constituírem corpora que possibilite a pesquisa científica.

Nesse sentido é que se tem se desenvolvido, dentre outros, projetos como O Corpus Histórico do Português Tycho Brahe, que consiste em "um corpus eletrônico anotado, composto de textos em português escritos por autores nascidos entre 1380 e 1845" (CORPUS TYCHO BRAHE). Esse projeto Tycho Brahe, no Brasil, vem tendo seu corpus ampliado, no âmbito do projeto Padrões Rítmicos, Fixação de Parâmetros e Mudança Lingüística, proposto por Galves (2007), que tem como principal objetivo "modelar a relação entre prosódia e sintaxe no processo que levou do Português Clássico ao Europeu Moderno".

Paralelamente, outro projeto desenvolve o corpus anotado Tycho Brahe: trata-se do projeto Memórias do Texto: aspectos tecnológicos na construção de um corpus histórico do português, desenvolvido na USP por Paixão de Sousa (2009), cujo objetivo é "desenvolver tecnologias de processamento de texto com aplicação na Lingüística de Corpus".

É nesse contexto que se desenvolveu inicialmente por Paixão de Sousa, Pablo Faria e Fabio Kepler uma ferramenta digital denominada E-Dictor, assim definida tecnicamente:

> E-Dictor é uma ferramenta de transcrição e codificação de corpora de textos em formato XML, de modo que possam ser editados e usados de diversas maneiras (como por exemplo, para análises lingüísticas, morfológicas, sintáticas, etc.). O XML define etiquetas e informações para incorporar as edições, bem como informações de layout de textos (títulos, subtítulos, quebras de pagina/linha/coluna, cabeçalhos, nota de rodapé, etc.). (PAIXÃO DE SOUSA, 2009, s/p).

Essa ferramenta, diz a autora Paixão de Souza, não é um processador de texto comum. Trata-se de um software de edição, cujo objetivo fundamental é

> [...] converter em formato XML uma porção de texto simples, permitindo ao usuário editar os elementos da estrutura resultante. O E-Dictor não oferece facilidades de layout como bordas, margens, mudança de fontes, e assim por diante. Ele possibilita ao usuário construir uma estrutura em XML sobre um texto para posteriormente usá-lo em outras tecnologias de gerenciamento (eg. XSLT) e apresentá-lo com uma interface mais amigável. (PAIXÃO DE SOUSA, 2009, s/p).

Na sua versão inicial ele funcionava prioritariamente visando gerar versões diversas de um texto a partir da linguagem em XML. Posteriormente esse programa passou a incorporar um tagger, isto é um mecanismo computacional capaz de, por exemplo, classificar morfologicamente palavras de um corpus de modo a permitir buscas estruturadas em linguagem de bancos de dados.

Tendo como suporte teórico-metodológico esses dois projetos supracitados, o de Galves (2003), e o de Paixão de Sousa (2009), desenvolve-se na UESB, coordenado por Santos e Namiuti (2009), o projeto DOViC (Memória conquistense: recuperação de documentos oitocentistas na implementação de um corpus digital), que visa contribuir com os estudos sobre a história gramatical do português brasileiro alimentando um banco de dados com informações textuais de um período importante desta língua – século XIX – ao mesmo tempo em que preserva a memória de uma cidade do interior baiano, Vitória da Conquista – BA. Assim sua proposta consiste em investir na criação de um banco de dados com referência e tipologia de manuscritos, o que resultará em um corpus digital.

Dentro desse projeto maior, propôs-se avaliar o desempenho do programa computacional E-Dictor quando aplicado a edição de textos originalmente

manuscritos como os que integram o corpus DOViC (cartas de alforria, escrituras, dentre outros). Nesse sentido, o presente subprojeto de pesquisa tem como objetivo geral contribuir na construção do corpus DOViC, investigando os aspectos de sua ortografia; e como objetivo específico analisar índices de erro e acerto do tagger do programa E-Dictor quando aplicado em etiquetagem morfológica de textos manuscritos editados.

Procedimentos Técnicos

Para realizar esta análise foi montado um corpus de dez cartas de alforria extraído do corpus DOViC. Num primeiro momento, após a seleção, fez-se a leitura e decifração dos manuscritos, resultando em textos transcritos paleograficamente. Em seguida, tais textos foram submetidos ao módulo de edição do E-Dictor. Entre as edições aplicadas podem-se citar as seguintes: modernização (atualização da palavra aos atuais padrões), junção (união de partes grafadas separadamente), segmentação (separação de palavras escritas juntas), grafia (implementação de pequenas correções gráficas) e expansão (completar a parte que falta no origina).

Em uma primeira etapa, visando comparar o funcionamento do etiquetador morfológico quando aplicados em textos com e sem edição, procedemos do seguinte modo: a) primeiramente submetemos ao etiquetador o conjunto de textos com as edições supracitadas; b) em seguida, submetemos ao etiquetador o mesmo conjunto de textos sem edições.

E, num terceiro momento, com os dados obtidos foi possível elaborar tabelas e gráficos, no intuito de fazer um levantamento do índice de erro e acerto.

Ortografia e Textos Eletrônicos

Conforme Souza (2002, p. 57), "A fala é algo tão inerente ao ser humano quanto o ato de respirar". Esse desenvolvimento levou a outro de grande importância para a história da sociedade. Visto que, com o tempo surgiu à necessidade de registrar graficamente os fatos, outrora passados oralmente, da vida em sociedade. Conforme afirma Souza (2002) por volta de 3200 a.C., no Oriente Médio surgiu entre o povo sumério um conjunto de sinais gráficos, cujas combinações complexas revolucionaram para sempre a humanidade. Esse fenômeno denominado "escrita" constituiu algo tão impressionante que diversos povos atribuíram o seu aparecimento aos deuses e heróis lendários.

Essa descoberta, a partir de então, espalhou-se entre outras civilizações e a história passou ser caracterizada em função da escrita. Durante o decorrer do tempo a escrita passou por uma evolução até chegar ao sistema alfabético, tal como conhecemos hoje, afirma Souza (2002).

Ainda conforme Souza (2002), essa evolução deu-se porque, com o tempo os ideogramas não conseguiam suprir as possibilidades da língua oral. Novas formas de representar a oralidade foram desenvolvidas. No entanto, quando acharam que seus problemas estavam resolvidos surgiu o inconveniente problema da variação lingüística.

Foram feitos esforços para estabelecer um sistema ortográfico. Mas a imposição de um novo sistema não é rápido e aleatório, afirma Souza (2002), o que justifica a diversidade de grafias no decorrer das épocas.

Tais fatores têm despertado o interesse de diversos pesquisadores para a análise de textos antigos que materializam, ou seja, trazem escritas essas variações lingüísticas.

Surge, então, recentemente um interesse na Lingüística, em diversas áreas (a exemplo da Lingüística de Corpus, Lingüística Computacional, Lingüística Histórica, Sintaxe, Semântica), no sentido de tanto coletar quanto analisar corpora compostos por textos antigos, sejam impressos ou manuscritos.

Além da elaboração de softwares para tratamento eletrônico de dados, essas novas tendências lingüísticas vêm-se aprimorando e desenvolvido novos meios de disponibilizar esses textos em ambientes digitais, visto que têm surgido novas formas de armazenamento de dados digitais. Foi o que mostrou Paixão de Sousa (2009), que em seu ensaio Conceito material de "Texto Digital" refletiu sobre a singularidade do texto digital: "Este se diferencia das demais formas de texto pela inclusão de etapas de processamento artificial da linguagem em sua cadeia de difusão" (Paixão de Sousa, 2009, s/p).

Quanto a isso, a autora propôs uma série de reflexões devido à singularidade do texto digital. Primeiro explorou o conceito de '"texto' como uma ponte no espaço-tempo: um registro de enunciados produzidos num ponto do espaço e do tempo, que podem ser recebidos num ponto diferente do espaço e do tempo" (Paixão de Sousa, 2009, s/p).

Para a realização de tais tarefas de processamento digitais se requer o desenvolvimento de novas tecnologias que posso subsidiar tais intuitos, o que sai do simples plano simbólico, como afirmou Paixão de Sousa (2009, s/p):

> De outro lado está a questão das tecnologias envolvidas na difusão da informação simbólica representada pela escrita. Aqui saímos do plano estritamente material das tecnologias inventadas pelo homem para estabelecer as correspondências simbólicas dentro de cada sistema e propagá-las no tempo e no espaço.

Recorrendo a essas ferramentas digitais, recentes pesquisadores vêm trabalhando na construção de corpora digitais. Um exemplo disso é a construção

do corpus Tycho Brahe que motivou o desenvolvimento de um etiquetador digital. Para ficar mais claro o que isso significa, valeremos da definição de Finger (2000, p. 1) para etiquetagem: "Etiquetar um texto quer dizer classificar cada palavra em contexto a uma categoria morfossintática".

Como exemplo, tomemos uma sentença citada por Finger (2000):

Jejua o enfermo para recuperar a saúde

Tendo essa forma de entrada (input), a saída do processo (output) de etiquetagem seria a mesma sentença no formato [palavra]/[etiqueta]:

Jejua/VB-P o/D enfermo/N para/P recuperar/VB a/D-F saúde/N

Esse programa foi baseado na ferramenta criada em um método já existente, como afirma Finger (2000, p. 2):

> O etiquetador Tycho Brahe é baseado no método desenvolvido por Eric Brill, mas fez-se necessário adaptar este método e introduzir modificações. O método de Brill [Bri95] para etiquetagem morfossintática de palavras é um método baseado em aprendizado computacional. O programa apredendor gera uma série de regras contextuais que serão usadas na etiquetagem.

Tal programa tem um treinamento manual para seu aperfeiçoamento. Desse modo constantemente requer-se renovações para adaptá-lo a novo corpus.

Com base nesse programa pode-se obter um vasto número de informações em um curto período de tempo o que facilita tarefas básicas na coleta de dados estatísticos para comporem suas pesquisas, visto que este apresenta um crescente aumento em dados.

No entanto tal ferramenta apresenta uma dificuldade no que diz respeito a textos não editados[365]. Os índices de acertos em tais textos se reduzem consideravelmente. Quanto a sua precisão com os textos do corpus Tycho Brahe, Finger (2000) apresentou os seguintes dados[366]:

> Os resultados iniciais obtidos para o Corpus Tycho Brahe indicavam que o etiquetador possuía uma precisão de 78,28%. Na data da escrita deste artigo, o etiquetador encontra-se com 95,43% de precisão. Isto corresponde a uma diminuição de 4,55 vezes no número de erros cometidos pelo etiquetador (FINGER, 2000, p. 2).

Como já foi demonstrado em outros estudos, em textos não editados o número de acerto na etiquetagem pelo tagger tende a sofrer uma redução. Isso se apresenta como um obstáculo a ser ultrapassado pelos pesquisadores que se utilizam desta ferramenta visto que, conforme Menegatti (2002), afirmou que os textos portugueses só começaram a ter uma norma ortográfica padrão a partir do século XVIII, o que indica a presença de um vasto número de textos com uma grande diversidade de grafias. Com isso, vê-se a necessidade de um empenho na etiquetação eficaz e na posterior digitalização de tais em meios acessíveis a todos que desejam usá-los em suas pesquisas.

Não obstante, tais aparentes problemas tem sido encarados apenas como o incentivo na busca de novas tecnologias que possam solucioná-las. Por isso, tem-se procurado expandir o uso do tagger em outros corpora para que ele possa ser treinado e por fim apresentar níveis consideráveis de acertos em textos sem edição.

[365] Editados tecnicamente pelo programa E-dictor.
[366] Os quais servirão de base para comparações posteriores.

Resultados e Discussão: Desempenho do Tagger

Uma vez realizadas todas as etapas descritas em 2, que envolve submeter os dados ao tagger, ou seja, ao etiquetador digital, foram encontrados os seguintes resultados, quanto ao índice de erro e acerto da ferramenta acima descrita: na amostra sem edição verificou-se um total de 51% de acerto e 49% de erro, por outro lado, na amostra com edição registrou-se um total de 91% de acerto e 9% de erro.

No que tange aos índices de erro e acerto do etiquetador em cada edição aplicada, comparando-os com a amostra que não foi editada, constataram-se os seguintes resultados: na edição grafia (Exemplo: municipio SE[367] e município CE, que na amostra sem edição tagger classificou como nome próprio) encontrou-se 88% de acerto e 12% de erro na amostra editada e 58% de acerto e 42% de erro na amostra sem edição; quando se aplicou a edição junção (Exemplo: oito centos SE e oitocentos CE, em que o tagger, na amostra sem edição, não a considerou como uma única palavra), constatou-se 100% de acerto e 0% de erro na amostra com edição, por outro lado, 100% de erro e 0% de acerto na amostra que não foi editada; na edição modernização (Exemplo: elle SE e ele CE, que o etiquetador, na amostra sem edição classificou como nome próprio), registrou-se 95% de acerto e 5% de erro na amostra com edição, porém 53% de acerto e 47% de erro na amostra sem edição; na edição segmentação (Exemplo: deliberdade SE e de liberdade CE, a ferramenta, na amostra não editada considerou como sendo apenas uma única palavra), registrou-se 100% de acerto e 0% de erro, na amostra com edição, no entanto 0% de acerto e 100% de erro na amostra sem edição; na edição expansão (Exemplo: test SE e testemunho CE, o etiquetador, na amostra que não foi editada, classificou como nome próprio), constatou-se 100% de acerto na

[367] A sigla SE designa às palavras que não foram editadas tecnicamente; quanto à sigla CE designa às que foram tecnicamente editadas pelo *tagger.*

amostra com edição e 100% de erro na amostra sem edição; por fim, quando foi aplicada uma dupla edição junção e modernização (Exemplo: Tabel leam SE e Tabelião CE, em que o tagger considerou ambas como nome proprio), tanto a amostra com edição como a amostra sem edição, registrou-se um percentual de 100% de erro e 0% de acerto.

CONCLUSÕES

Posteriormente ao procedimento metodológico de análise, e considerando o resultados discutidos acima, pode-se detectar, ainda preliminarmente, em quais circunstancias o tagger do programa E-Dictor, quando aplicado em etiquetagem morfológica de textos manuscritos editados, como é o caso do documentos do corpus DOViC, apresenta maior dificuldade em acertar a classificação morfológica, e em quais edições ele apresentou maior percentual de erro acerto.

Portanto, tal a pesquisa aqui relatada, ainda que inicial, pode – desde já e com futuros desenvolvimentos - contribuir duplamente. De um lado, traz resultados que podem ser usados no aperfeiçoamento da ferramenta digital apresentada. E de outro, podem contribuir na eficácia dos procedimentos de análise eletrônica do corpus DOViC, a exemplo dos que envolvem aspectos de sua ortografia.

REFERÊNCIAS

CORPUS TYCHO BRAHE. http://www.tycho.iel.unicamp.br/~tycho/corpus/index.html.

FINGER, M**. Técnicas de Otimização da Precisão Empregadas no Etiquetador Tycho Brahe.** São Paulo, 2000.

GALVES, C. **Padrões rítmicos fixação de parâmetros e mudança linguística**. Campinas: FAPESP-UNICAMP, 1997. Projeto de pesquisa.

MENEGATTI, T. A**. Regras Lingüísticas para tratamento Computacional da Variação de grafia e abreviaturas do Corpus Tycho Brahe.** São Paulo: UNICAMP, 2002.

NAMIUTI, C. **"Aspectos da história gramatical do português"**. Interpolação, negação e mudança" (Tese de doutorado). Campinas, SP: [s.n.], 2008.

NOVAIS, F. G., SANTOS, J. V**. Análise do desempenho do etiquetador lexical do programa E-Dictor aplicado documentos manuscritos brasileiros do Século XIX,** 2011. (Apresentação de Trabalho no I Congresso de Estudos do Léxico).

NOVAIS, F. G., NAMIUTI, C., SANTOS, J. V. **A ortografia nas cartas de alforria do corpus DOViC:** Análise de desempenho de ferramenta de anotação morfológica automática aplicada à documentos do corpus DOViC**,** 2010**. (**Apresentação de Trabalho no XIV Seminário de Iniciação Científica da UESB).

NOVAIS, F. G., NAMIUTI, C., SANTOS, J. V. **Análise de desempenho de ferramenta de etiquetação morfológica automática aplicada a textos do corpus DOViC,** 2010 **(**Apresentação de Trabalho no V Seminário de Pesquisa em Estudos da Lingua(gem) – SPELL).

PAIXÃO DE SOUSA, M. C. **Conceito material de texto digital:** Um ensaio. Texto Digital**.** Rio de Janeiro: UERJ, 2009, p. 6.

SANTOS, J. V. **Liberdade na escravidão:** uma abordagem semântica do conceito de liberdade em cartas de alforria. (Tese de Doutorado em Lingüística). Campinas: Instituto de Estudos da Linguagem da UNICAMP, 2008.

SOUZA, Nazarete de. **Estudo de alguns aspectos da ortografia da carta de Pero Vaz de caminha.** (Dissertação de mestrado). Campinas: Instituto de Estudos da Linguagem, UNICAMP, 2002.