NAMIUTI, Cristiane. O corpus anotado do português histórico: um avanço para as pesquisas em lingüística histórica do português. *Revista Virtual de Estudos da Linguagem – ReVEL.* V. 2, n. 3, agosto de 2004. ISSN 1678-8931 [www.revel.inf.br].

O CORPUS ANOTADO DO PORTUGUÊS HISTÓRICO: UM AVANÇO PARA AS PESQUISAS EM LINGÜÍSTICA HISTÓRICA DO PORTUGUÊS

Cristiane Namiuti[1]

cristianenamiuti@yahoo.com

Introdução

Inspirado no corpus parseado do Inglês Médio (*Penn-Helsinki Parsed Corpus of Middle English* (PPCME)) o projeto temático "Padrões Rítmicos Fixação de Parâmetros e Mudança Lingüística"[2] teve como um de seus objetivos principais a construção de um corpus anotado do português histórico.

Elaborado nos moldes do PPCME, o *Corpus Anotado do Português Histórico Tycho Brahe* consiste em um corpus eletrônico anotado morfológica e sintaticamente, com livre acesso pela internet, composto por textos em prosa, escritos em português por falantes nativos do português europeu, nascidos entre 1550 e 1850 (cf. http://www.ime.usp.br/~tycho/corpus). O objetivo do *Corpus Tycho Brahe* é disponibilizar publicamente dados históricos do português europeu anotados de tal maneira que os estudiosos de sua história possam recuperar de maneira rápida informações categoriais e estruturais pertinentes à análise morfo-sintática da língua.

[1] Doutoranda do Instituto de estudos da linguagem (IEL/UNICAMP) com projeto de pesquisa financiado pela fapesp (04/01557-0) e orientado por Charlotte Galves.

1. O Corpus Anotado do Português Histórico

A construção do Corpus Tycho Brahe *(CTB)* envolveu a elaboração de ferramentas computacionais de anotação, das quais as principais são o *etiquetador automático* desenvolvido por Marcelo Finger do IME-USP, e o *analisador automático* para o Português obtido a partir do treinamento de um analisador universal desenvolvido na Universidade de Pensilvânia por Dan Bickel. O treinamento desse analisador foi possível graças ao desenvolvimento de um sistema de anotação sintática, nos moldes do sistema proposto por Taylor e Kroch (1998), para o Inglês Médio, e da anotação manual de 50.000 palavras de um texto por Helena Britto, pós-doutoranda do projeto.

Segundo a metodologia proposta no PPCME, a etiquetagem morfológica dos textos constitui o primeiro passo do processo de anotação, servindo de base para a codificação sintática subseqüente. Contudo, é importante ressaltar que os textos automaticamente etiquetados são disponibilizados independentemente, uma vez que já contêm por si informações relevantes para estudos da língua. Por isso, cada um dos textos, representados por aproximadamente cinqüenta mil (50.000) palavras cada, deve estar eletronicamente disponível em três formatos:

i. ortograficamente transcritos: *Senhor: Ofereço a Vossa Majestade as Reflexões sobre a vaidade dos homens*

ii. morfologicamente etiquetados: *Senhor/NPR :/. Ofereço/VB-P a/P Vossa/PRO$-F Majestade/NPR as/D-F-Reflexões/NPR-P sobre/P a/D-F vaidade/N dos/P+D-P homens/N-P*

iii. sintaticamente anotados:
```
(IP-MAT (NP-SBJ *pro*)
        (NP-VOC (NPR Senhor))
        (. :)
        (VB-P Ofereço)
                (PP (P a)
                (NP (PRO$-F Vossa)
```

[2]Projeto desenvolvido no Instituto de Estudos da Linguagem (IEL-UNICAMP), financiado pela fapesp (98/3382-0) e coordenado pelos professores: Charlotte Galves e Bernadette Abaurre (IEL-UNICAMP), e António Galves (IME-USP).

(NPR Majestade)))
(NP-ACC (D-F-P as)
(NPR-P Reflexões)
(PP (P sobre)
(NP (D-F a)
(N vaidade)
(PP (P+D-P dos)
(NP (N-P homens))))))

Atualmente o CTB contém 41 textos ortograficamente transcritos, 23 morfologicamente etiquetados e somente 1 sintaticamente anotado. Entretanto, o fato de não se ter ainda os textos anotados sintaticamente não nos impede de fazermos uma pesquisa lingüística ágil e abrangente tanto no que diz respeito à quantidade de dados quanto no espaço temporal que se pretende percorrer na história do português europeu, pois o sistema de marcação morfológica do CTB (cf. http:www.ime.usp/~tycho/manual/tags.html) traz informações precisas que nos possibilitam fazer buscas automáticas e/ou semi-automáticas e a recuperação instantânea dos dados lingüísticos.

Portanto a etiquetagem morfológica dos textos históricos já é um grande avanço nas pesquisas de lingüística diacrônica.

2. A busca automática

A busca dos dados para a pesquisa se dá a partir do texto etiquetado através de expressões regulares (er) e comandos da linguagem de programação "perl" no ambiente "Linux".
(cf. http://www.ime.usp.br/~tycho/corpus/utilities)

A er é a chave dos comandos e "scripts" de busca, é ela que indicará o dado a ser buscado nos textos do CTB. Contudo são as etiquetas morfológicas as bases das ers que utilizamos para buscar nossos dados.

Já buscamos automaticamente 26.620 dados com clíticos: 3.586 dados com o clítico pós verbal (Vcl, dados de ênclise), 21.386 dados com o clítico pré verbal(clV, dados de próclise) e 1.648 dados em que existe um elemento entre o clítico e o verbo

(clXV, dados de interpolação).

2.1 A expressão regular, O algoritmo dos comandos de busca.

Sintaxe da *er:* as seqüências dos elementos a serem localizados nos textos devem vir indicadas utilizando as etiquetas morfológicas e os símbolos dos comandos descritos abaixo entre **(/ /);**:

***** - significa zero ou mais vezes. Ex: *bicicleta**s**** - o *s* final da palavra pode não ocorrer ou ocorrer mais de uma vez: *bicicleta, bicicleta**sss***.

+ - uma ou mais vezes. Ex: *Pssi**u**+* - o *u* final pode aparecer uma ou mais vezes, ou seja - *pssiu* ou *pssiuuuu ...*

.* - qualquer caracter 0 ou mais vezes. Ex: *gov.** - isto pode prever *gov, governo, governador, governadora, governadores ...*

? – zero ou uma vez. Ex: *governador(es)**?***- isto prevê que o que está entre parênteses pode ocorrer uma vez ou não ocorrer: *governador* ou *governado**res***

[^] - "diferente de". Ex: ***[^(gover)]****nador* - palavras que terminam com *nador* mas não pode ser *governador*, prevê por exemplo *senador*.

| - "ou". Ex: ***(gov|sen)****nador* – prevê *governador* e *senador*

**** - símbolo de escape (utilizado para diferenciar um símbolo de um caracter do texto, ex: nas etiquetas temos por exemplo "P+CL" - para podermos esclarecer que este "+" não indica "*uma ou mais vezes*" e sim um "caracter" do texto, utilizamos o símbolo de escape "P\+CL".

^ - indica a posição de início absoluto de linha. Ex: ***^Governador*** – prevê as ocorrências da palavra *Governador* em início absoluto de frase.

( ) – deve ser empregado quando for necessário agrupar caracteres que serão governados pelo mesmo símbolo de comando. Como quando usamos o símbolo **|**

"ou" e o símbolo **?** *"zero ou uma vez"* . Exs: governadores***? –*** prevê *governadore* e *governadores*; já *governador**(es)?*** – prevê *governador* e *governador**es***.

( ) - também guarda as informações contida dentro dele para recuperar mais adiante na *er*.

\ \ - para recuperar a informação guardada nos parênteses utiliza-se estas barras referindo o número dos parênteses que se quer recuperar... Ex.: (/(liberdade**) \1**\/) – a informação contida dentro dos parênteses foi recuperada em "\1\" e traduz a seqüência: *"liberdade liberdade"...*

Comandos de busca *perl* como *perl –ne "print if (**/er/**);" foo1.txt > foo2.txt* buscam as informações indicadas pela ***er*** em arquivos de entrada (foo1.txt) e guarda-as em arquivos de saída (foo2.txt).

Um dos assuntos de interesse nos textos do corpus Tycho Brahe é a variação no uso de próclise e ênclise verbal dos pronomes clíticos. Os pronomes clíticos no CTB são marcados pelas etiquetas **/CL** e **/SE** enquanto que os verbos são marcados com etiquetas específicas de acordo com o tipo (ser - /SR; *estar* – /ET, *ter* -/TR; *haver* - /HV e os outros verbos - /VB) e com o tempo e o modo (/P – presente do indicativo; /SP – presente do subjuntivo; /D passado do indicativo; /SD passado do subjuntivo; /R futuro do indicativo; /SR futuro do subjuntivo; /RA – formas em "ra": passado perfeito). As duas *ers* a seguir indicam as seqüências "clítico-verbo" e "verbo-clítico":

1) (/\/(CL|SE [^\/]* \/(VB|TR|SR|HV|ET)\-(S**?**|P|S**?**D|S**?**R|RA)/): Que se traduz: a etiqueta /CL ou a etiqueta /SE (que marcam qualquer *clítico* ou o clítico *se*) seguida de vários caracteres diferentes de **/** (o que se interpreta como uma palavra) seguidos de uma etiqueta de verbo finito.

2) (/\/(VB|TR|SR|HV|ET)\-(S**?**|P|S**?**D|S**?**R|RA)\+(CL|SE)/): Indica as etiquetas de verbo finito com ênclise: */VB-P+CL*, */VB-SP+SE* e assim por diante.

O resultado final desta etapa automática são arquivos com as ocorrências de próclises e ênclises a verbos finitos nos textos em questão, como nos exemplos:

1. [mco-clV][3] *Dado: lhes/CL restituîa/VB-D*: E/CONJ lançando/VB-G as/D-F-P contas/N-P ao/P+D que/WPRO lhes/CL bastava/VB-D para/P a/D-F jornada/N ,/, isso/DEM **lhes/CL restituîa/VB-D** ,/, com/P Nunca/ADV-NEG Deos/NPR queira/VB-SP que/C vossas/PRO$-F-P mercês/NPR-P lhes/CL falte/VB-SP o/D necessario/ADJ para/P seu/PRO$ caminho/N ,/, e/CONJ com/P o/D mais/ADV-R ficava/VB-D ./.

2. [mco-Vcl][4] *Dado: puzeraõ-lhes/VB-D+CL*: Os/D-P Romanos/NPR-P na/P+D-F paz/N ,/, que/WPRO fizeraõ/VB-D com/P os/D-P Carthaginezes/NPR-P ,/, puzeraõ-lhes/VB-D+CL por/P condição/N ,/, que/C lhes/CL entregassem/VB-SD a/D-F armada/N ,/, que/WPRO tinhaõ/TR-D :/. **puzeraõ-lhe/VB-D+CL** o/D fogo/N ,/, e/CONJ ficaraõ/VB-D todos/Q-P quietos/ADJ-P ./.

2.3 Os Programas de busca do CTB com base nas "ers" e nos comandos "perl":

Para a pesquisa sobre os clíticos elaborei os seguintes *scripts* com a assessoria de Miguel Galves:

a) getclitics-finite.pl: contem 6 expressões regulares que localizam separadamente as próclises com /CL e com /SE, as ênclises com /CL e /SE em início absoluto de sentença e as ênclises com /CL ou /SE com algum constituinte antecedendo o verbo finito.

3. [clV-mel] *Dado: as/CL publicará/VB-R*: Deus/NPR **as/CL publicará/VB-R** ./.

[3] [clV-mel] = uma sentença proclitica com um clítico /CL no texto de F. M. de Melo.

[4] [0Vse-mel] = uma sentença enclítica com o clítico /SE e com o verbo em posição inicial absoluta de sentença no texto de F. M. de Melo.

4. [0Vse-mel] Dado: *Achaque-se/VB-SP+SE:* **Achaque-se/VB-SP+SE** embora/ADV à/P+D-F melancolia/N êste/D argumento/N ,/,

    b) getse-infinitival.pl: contem 3 expressões regulares que localizam separadamente as próclises com /SE, as ênclises com /SE em início absoluto de sentença e as ênclises /SE com algum constituinte antecedendo o verbo infinitivo.

5. [cou-VBse] E/CONJ navegando/VB-G por/P antre/P elas/PRO ,/, dali/P+ADV a/P dezoito/NUM dias/N-P chegaram/VB-D a/P uma/D-UM-F Ilha/NPR verde/ADJ-G ,/, de/P que/WPRO lhe/CL saíram/VB-D alguns/Q-P paráos/N-P com/P gente/N da/P+D-F terra/N baça/ADJ-F ,/, como/CONJS a/D-F de/P Maluco/NPR ;/. e/CONJ chegando/VB-G junto/ADV de/P uma/D-UM-F das/P+D-F-P náos/N-P da/P+D-F conserva/N ,/, lhe/CL falou/VB-D um/D-UM dos/P+D-P paráos/N-P em/P Portuguez/NPR ,/, e/CONJ lhes/CL disse/VB-D :/. Bons/ADJ-P dias/N-P ,/, matalotes/N-P ,/, e/CONJ voltaram/VB-D logo/ADV ,/, porque/CONJ viram/VB-D **despedir-se/VB+SE** da/P+D-F náo/N Capitanea/NPR o/D esquife/N pera/P os/CL ir/VB chamar/VB ,/, e/CONJ daqui/P+ADV se/SE ficaram/VB-D estas/D-F-P Ilhas/NPR-P chamando/VB-G as/D-F-P dos/P+D-P Matalotes/NPR-P ,/, que/WPRO estão/ET-P em/P dez/NUM gráos/N-P ./.

    c) getinterpolagion.pl: contem 3 expressões regulares que localizam separadamente as ocorrências de interpolação da negação, de constituintes complementos e advérbios, e também a interpolação de verbo infinitivo.

6. [cou-clnegV] Dado: o/CL não/NEG fazer/: <P_09> Aqui/ADV surgio/VB-D a/D-F Armada/NPR ,/, e/CONJ se/SE deteve/VB-D trinta/NUM e/CONJ dous/NUM dias/N-P ,/, e/CONJ deixáram/VB-D de/P a/CL povoar/VB por/P não/NEG ser/SR a/D-F terra/N boa/ADJ-F ,/, e/CONJ porque/CONJ levava/VB-D o/D Villa-Lobos/NPR determinado/VB-AN de/P **o/CL não/NEG** fazer/VB mais/ADV-R de/P doze/NUM gráos/N-P ./.

7. [cou-clxV] Dado: lhe/CL êles/PRO: E/CONJ não/NEG faltou/VB-D quem/WPRO murmurasse/VB-SD de/P António/NPR de/P Almeida/NPR ,/, havendo/HV-G que/C vinha/VB-D peitado/VB-AN dos/P+D-P Castelhanos/NPR-P ,/, porque/CONJ trazia/VB-D peças/N-P ,/, e/CONJ brincos/N-P ,/, que/WPRO **lhe/CL êles/PRO** deram/VB-D ./.

    d) getnegpro.pl: contem uma expressão regular que localiza as próclises antecedidas por negação.

    e) getposs.pl: contem 6 expressões regulares que buscam pronomes possessivos

antecedidos ou não de artigo ou/e preposição.

8. [cou-Dposs] Dado: o/D seu/PRO$: Destas/P+D-F-P cousas/N-P deo/VB-D conta/N ao/P+D Xá/NPR Ismael/NPR,/, dizendo-lhe/VB-G+CL como/WADV estavam/ET-D dispostas/VB-AN-F-P pera/P com/P mais/ADV-R facilidade/N tornar/VB a/P ganhar/VB **o/D seu/PRO$** ./.

9. [cou-Pposs] Dado: de/P sua/PRO$: Martim/NPR Afonso/NPR soube/VB-D **de/P sua/PRO$-F** ida/N ,/, e/CONJ o/CL saiu/VB-D a/P receber/VB fóra/ADV ,/, mostrando-se-lhe/VB-P+SE+CL Dom/NPR Estevão/NPR carregado/VB-NA ,/, e/CONJ de/P poucos/Q-P cumprimentos/N-P :/. <P_06> e/CONJ ali/ADV lhe/CL fez/VB-D entrega/N da/P+D-F India/NPR ,/, perante/P Fernão/NPR Rodrigues/NPR de/P Castelo-Branco/NPR ,/, Veador/NPR da/P+D-F Fazenda/NPR ,/, e/CONJ de/P João/NPR da/P+D-F Costa/NPR ,/, Secretário/NPR ,/, que/WPRO disso/P+DEM fez/VB-D seu/PRO$ termo/N ordinário/ADJ ./.

10. [cou-P-Dposs] Dado: pera/P o/D seu/PRO$: E/CONJ porque/CONJS receou/VB-D que/C os/D-P inimigos/N-P fôssem/VB-SD após/P êle/PRO ,/, mandou/VB-D diante/ADV sua/PRO$-F mulher/N ,/, e/CONJ ele/PRO foi/VB-D passando/VB-G por/P tôdas/Q-F-P as/D-F-P Cidades/NPR-P ,/, que/WPRO tinha/TR-D tomadas/VB-AN-F-P ,/, levando/VB-G as/D-F-P guarnições/N-P que/WPRO nelas/P+PRO tinha/TR-D posto/VB-PP ,/, e/CONJ foi-se/VB-D+SE caminhando/VB-G apressadamente/ADV **pera/P o/D seu/PRO$** Reino/NPR ./.

f) tagout.pl: contem uma expressão regular que retira as etiquetas morfológicas das palavras.

Input: [clV-mel] Dado: as/CL publicará/VB-R: Deus/NPR as/CL publicará/VB-R ./.
Output: [clV-mel] Dado: as publicará: Deus as publicará .

Os programas criam automaticamente seus arquivos de saída com os dados separados prontos para a análise lingüística.

3. Sumário

O avanço tecnológico da pesquisa de corpus permite aos lingüistas a precisão e a agilidade tão almejada nas ciências da linguagem. Pois com o uso de ferramentas automáticas poder-se-á descrever e analisar grandes corpora em uma velocidade bastante razoável para a pesquisa lingüística. Além de proporcionar a vantagem dos dados já estarem eletronicamente disponível e

de fácil acesso.

REFERÊNCIAS BIBLIOGRÁFICAS

1. BRITO Helena, GALVES, Charlotte, RIBEIRO, Ilza, AUGUSTO, Marina e SCHER, Ana Paula. (1997). Morphological annotation system for automated tagging of eletronic textual corpora: from English to Romance Languages. *http://www.ime.usp.br/~tycho/participants/compl_list.html*
2. FINGER, Marcelo.(1999). Técnicas de otimização da precisão empregadas no etiquetador Tycho Brahe. *http://www.ime.usp.br/~tycho/participants/compl_list.html*
3. GALVES, Charlotte e BRITTO, Helena. (1998). A construção do Corpus Anotado do Português Histórico Tycho Brahe: o sistema de anotação morfológica. *http://www.ime.usp.br/~tycho/participants/compl_list.html.*
4. GALVES, Charlotte, BRITTO, Helena e FINGER, Marcelo. (1998). Computational and linguistic aspects of the construction of The Tycho Brahe Parsed Corpus of Historical Portuguese. *http://www.ime.usp.br/~tycho/participants/compl_list.html*
5. GALVES, Charlotte, BRITTO, Helena e PAIXÃO DE SOUSA, Maria Clara. (2003). Clitic Placement in European Portuguese: Results from the Tycho Brahe Corpus. *http://www.ime.usp.br/~tycho/participants/compl_list.html*
6. GALVES, Charlotte. (1998). Clitic Placement in the History of Portuguese and the Syntax-phonology interface. *http://www.ime.usp.br/~tycho/participants/compl_list.html*