A Construção do Corpus Anotado do Português Histórico Tycho Brahe: o sistema de anotação morfológica*

Charlotte Galves and Helena Britto+

Unicamp, Caixa Postal 6153, CEP 13082-970 Campinas, Brasil
galvesc@ime.usp.br helenabritto@mpc.net.com.br

Resumo: Inspirado no sistema proposto para o *Penn-Helsinki Parsed Corpus of Middle English* (PPCME) [28], o sistema de etiquetas utilizado no etiquetador morfológico automático usado para a construção do *Corpus Anotado do Português Histórico Tycho Brahe* já foi apresentado em diversos trabalhos ([10], [3], [2], [6]). Neste trabalho, apresentaremos e discutiremos os processos lingüísticos decisórios subjacentes à elaboração do conjunto de etiquetas e suas aplicações.

1 Introdução

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 originalmente em português por falantes nativos do português europeu nascidos entre 1550 e 1850 (cf. http://www.ime.usp.br/~tycho/corpus).

Inserido em um projeto de pesquisa que visa sobretudo a estudar a relação entre mudança prosódica e mudança sintática no período do português comumente denominado *clássico,* 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 rápida e inambigüamente informações categoriais e estruturais pertinentes a análises morfo-sintáticas da língua.[1] Segundo a metodologia proposta no PPCME [28], a etiquetagem morfológica dos textos constitui o primeiro passo deste 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 cinqüenta mil (50.000) palavras cada, deve estar eletronicamente disponível em três formatos:

(i) ortograficamente transcritos: *O fim da minha jornada verá V. Ex.a.*
(ii) morfologicamente etiquetados: *O/D fim/N da/P+D-F minha/PRO$-F jornada/N-F verá/VB-R Vossa/PRO$-F Excelência/NPR ./.*
(iii) sintaticamente anotados (ainda em fase de construção)[2]

O presente trabalho discute a construção e uso do sistema de anotação morfológica do ponto de vista lingüístico, e será organizado como segue. Na seção 2, discutiremos os requisitos para um sistema de etiquetas morfológicas. A seção 3 apresentará o sistema adotado para a anotação do corpus. Na seção 4, proporemos alguns estudos de caso, justificando decisões que foram tomadas na atribuição de uma ou mais etiquetas a certos itens lexicias. Finalmente, na seção conclusiva do trabalho, indicamos o rendimento, do ponto de vista lingüístico, da aplicação das etiquetas proposta a textos de autores nascidos nos séculos XVI e XVII.

2 Requisitos para um sistema de anotação morfológica

Para a elaboração do conjunto de etiquetas e sua utilização, procuramos satisfazer os seguintes requisitos.

2.1 Adequação Descritiva

As etiquetas têm que representar e discriminar adequadamente as categorias necessárias à descrição dos enunciados presentes na língua em geral, e no corpus em particular. Obviamente, essa restrição impede a utilização pura e simples de um etiquetador elaborado para uma certa língua no processamento de uma outra. Por esta razão, por exemplo, a adaptação para o português do sistema elaborado para o inglês médio envolveu a introdução de uma série de etiquetas destinadas a codificar as propriedades flexionais da língua (gênero dos nomes, morfologia verbal diversificada, sistema de pronomes-complemento átonos – doravante clíticos).

2.2 Recuperabilidade da Informação

O objetivo do corpus anotado é permitir aos estudiosos da história do português obterem de maneira rápida e confiável as informações necessárias para desenvolver análises sincrônicas ou diacrônicas de aspectos lexicais, morfológicos e sobretudo sintáticos da língua. O conjunto de etiquetas tem que ser construído e aplicado de modo a permitir recuperar da maneira mais econômica e exaustiva possível essas informações. Por outro lado, é preciso ressaltar a necessidade de não fechar análises na

base da língua moderna, ocultando fenômenos em mudança. Esse ponto será ilustrado abaixo com a questão do particípio (v. seção 4.4). Note-se que isso significa também que se tende para um sistema capaz de abranger o mais uniformemente possível todas as fases da língua, uma vez que não se podem definir a priori fronteiras entre períodos.

2.3 Simplicidade Computacional

O número total de etiquetas diferentes compondo o conjunto deve ser compatível com um tratamento computacional do corpus, nomeadamente com o treinamento de um etiquetador automático aplicado a este. O problema da complexidade computacional ligada à riqueza morfológica do português é amplamente discutido nas referências citadas acima, e solucionado com um sistema de etiquetas com estrutura interna, compostas por uma parte de base, à qual é associada a classe do item lexical, e sub-partes que expressam tanto um sub-grupo dentro de uma mesma classe de palavra (1), quanto traços flexionais carregados pelo item (2).

(1) o/D problema vs. um/D-UM problema
(2) os/D-P belos/ADJ-P campos/N-P

Este sistema de etiquetas em sub-níveis permite dois passos distintos no treinamento do etiquetador morfológico automático, contornando assim a complexidade computacional criada pelo crescimento do número de etiquetas distintas devido à riqueza morfológica. Continua pertinente porém a necessidade de limitar o número tanto das etiquetas-raiz, quanto das sub-etiquetas. Restringimos assim, por exemplo, as sub-etiquetas flexionais de gênero e número às classes de palavras que expressam a concordância nominal, o que exclui os pronomes pessoais.

2.4 Adequação Sintática

A anotação morfológica serve de base para a anotação sintática (*parsing*) do corpus. Ela deve portanto facilitar ao máximo a tarefa do parser automático, otimizando a interação entre a adequação descritiva morfológica e a adequação descritiva sintática. Este requisito nos leva, por exemplo, a atribuir uma etiqueta distinta ao item lexical até, para distinguir o seu uso como preposição (P) ou como operador de focalização (FP). Apesar de esse item ser tradicionalmente classificado como preposição em ambos os usos, eles devem ser distinguidos do ponto de vista sintático, já que só no primeiro há um sintagma caracterizável como preposicionado.

2.5 Decidilibidade

A atribuição de uma ou várias etiquetas a uma palavra tem que ser baseada em regras não ambíguas, cuja regularidade de aplicação facilita o treinamento das ferramentas automáticas, e a verificação manual do resultado das mesmas. Não pode deixar margem a dúvidas susceptíveis de serem resolvidas diferentemente conforme o anotador.[3] Além disso, deve-se ressaltar que em se tratando de textos históricos, a intuição do falante moderno pode ser inadequada (cf. o problema já levantado na seção 2.2) O requisito da decidilibidade traz a necessidade de etiquetas *default*, atribuídas em casos não marcados, ou ambíguos. Enfim, o papel da anotação não é analisar os dados, mas disponibilizá-los de tal modo que possam ser analisados de maneira sistemática.

Claramente, estes requisitos nem sempre podem ser satisfeitos conjuntamente, uma vez que são em parte antagônicos. Para nos atermos a exemplos simples, consideremos por exemplo a adequação descritiva, que leva à multiplicação das etiquetas, vs. a simplicidade computacional que impõe a sua limitação. A decidibilidade, por outro lado, favorece a biunivocidade entre palavra e etiqueta, entrando em choque com a adequação descritiva e a adequação sintática que requerem em muitos casos mais de uma etiqueta associada a um único vocábulo. A melhor escolha consistirá portanto numa otimização, caso a caso, desses requisitos, privilegiando consistentemente o objetivo do corpus.

3 O Sistema de Etiquetas Morfológicas

Norteado pelos requisitos acima, o sistema de anotação morfológica proposto é formado por dois grupos básicos de etiquetas: etiquetas categoriais (utilizadas para a classificação do item lexical segundo a classe de palavra a que pertence) e flexionais (articuladas às categorias por meio de diacríticos (v. (1) e (2)), podendo ser de natureza verbal, designadores de informações modo-temporais, ou não-verbal, indicadoras de traços flexionais de gênero e número).[4]

As motivações lingüísticas para a composição do sistema de anotação ora proposto podem ser assim resumidas.

3.1 Verbos

No que tange a etiquetas categoriais, dois tipos de itens [+V] são contemplados. De um lado, sob o rótulo VB, são codificados os verbos chamados *plenos* – i.e. que atribuem papel temático a seu(s) argumento(s). Por outro, etiquetas categoriais distintas são atribuídas a *ser* (SR), *estar* (ET), *ter* (TR) e *haver* (HV), pois que estas formas parecem oscilar diacronicamente entre um comportamento de verbos plenos e um de simples auxiliares verbais, restritos apenas a portar informação flexional.

A distinção entre formas [+V] finitas e não-finitas, indubitavelmente relevante aos estudos sintáticos[5], é codificada em nosso sistema pela ausência (em formas infinitivas não-flexionadas) vs. presença (nas formas finitas em geral e não-finitas

de infinitivo (visivelmente) flexionado, gerúndio e particípio[6]) de etiquetas flexionais específicas associadas às etiquetas categorias acima listadas.

(3) *Ser/SR* grande/ADJ-G homem/N ,/, (...) bom/ADJ *é/SR-P* para/P o/D mundo/N (*in* Chagas).[7]
(4) quem/WPRO a/P esta/D-F *der/VB-SR* muito/Q tempo/N (*in* Sousa).
(5) nada/Q *quero/VB-P* de/P ninguém/Q mais/ADV-R que/CONJS o/D *encomendarem-me/VB-F+CL* a/P Deus/NPR (*in* Chagas).
(6) E/CONJ *havendo/HV-G* um/D-UM homem/N de/P *ler/VB* (*in* Lobo).

Ainda com relação às etiquetas flexionais associadas às formas lexicais [+V], a opção por codificar apenas informação modo-temporal, sem explicitação de traços número-pessoais, foi motivada, por um lado, para a adequação do sistema de etiquetas ao requisito referente à simplicidade computacional (v. seção 2.3)[8], e por outro lado, por privilegiarmos sobremaneira a codificação de morfologia visível.[9]

3.2 Nomes e pronomes

Nomes e pronomes tônicos. Tradicionalmente, nomes e pronomes compõem classes distintas; por isso, etiquetá-los diferentemente é o esperado. A história do português europeu, entretanto, oferece argumentos, para além dos tradicionais, que fortalecem a necessidade desta distinção. Trata-se do comportamento distinto entre pronome e nomes em sentenças não-dependentes com ordem *XP clítico V* (XP [+referencial]), quando observamos a mudança diacrônica gradual da língua – comportamento este não esperado do ponto de vista gramatical [5].[10]

O presente sistema prevê etiquetas distintas para essas categorias, que são ainda subdivididas entre pronomes tônicos (PRO) e possessivos (PRO$), por um lado, e nomes comuns (N) e próprios (NPR), por outro[11]. Destas, todas exceto PRO podem vir associadas a etiquetas flexionais de gênero e número.

(7) entre/P os/D-P *mistérios/N-P* do/P+D *recato/N* (*in* Chagas).
(8) Amai-o/VB-I+CL *vós/PRO* muito/Q com/P todo/Q *vosso/PRO$* coração/N (*in* Chagas).

Clíticos. Para os clíticos, duas etiquetas são propostas: SE, para o clítico *se* em todos seus contextos; CL, para os demais clíticos (*me*, *te*, *o*, *a*, *lhes*, etc). Ao clítico *se* é atribuída uma etiqueta particular em virtude de seu específico comportamento, não só pelo fato de desempenhar várias funções sintáticas (como partícula reflexiva, apassivizadora ou indeterminadora), como também por mostrar propriedades morfológicas idiossincráticas (não permitindo contrações – *ele mo deu* vs. **ele so deu*), mas sobretudo por ser, por si só, um tópico de investigação sincrônica e diacrônica.

3.3 Determinantes e Pronomes Demonstrativos

A etiqueta D, associada ou não às etiquetas flexionais de gênero e número, é aplicada não só aos elementos tradicionalmente chamados artigos definidos (*o*, *a*, *os*, *as*), mas também aos pronomes demonstrativos flexionáveis (*este*, *esse*, *aquele*, *esta*, *essa*, *aquela*, etc), uma vez que, no decorrer de toda a história do português europeu tais elementos apresentam idêntica distribuição sintática. Por outro lado, aos pronomes demonstrativos não-flexionáveis (*isto*, *isso* e *aquilo*), que apresentam efetivo comportamento pronominal, aplicamos a etiqueta DEM. Finalmente, para os determinantes indefinidos, aplicamos a etiqueta D obrigatoriamente associada à sub-etiqueta -UM (e opcionalmente às flexionais). Desse modo, diferenciamos dos demais, este determinante que, dentre outras, possui a propriedade de poder ser [±referencial].

3.4 O tratamento das Conjunções

Na adaptação para o português do sistema de anotação morfológica adotado para o inglês médio, encontramos problemas devidos a tradições gramaticais diferentes, baseadas, em grande parte, em funcionamentos morfo-sintáticos distintos. Esse problema diz respeito essencialmente à categorização das conjunções. No sistema de anotação do PPCME, a única conjunção de subordinação é *that*, etiquetada /C, e as únicas conjunções de coordenação são *and* e *but*, etiquetadas /CONJ. A tradição gramatical portuguesa nos leva a incluir na classe das conjunções muitos mais itens lexicais, etiquetados no inglês como preposições. Incluímos assim na classe das palavras etiquetadas /CONJ itens como *contudo*, *porém, que* (com interpretação explicativa, frequentemente encontrado nos textos do século 17), etc., e adicionamos ao nosso sistema a etiqueta /CONJS, para conjunções de subordinação, como *embora*, *conforme*, *como*, *que* (com interpretação causal, comparativa ou consecutiva).[12]

3.5 Adjetivos, Advérbios e Quantificadores

Os itens lexicais que, do ponto de vista interpretativo, quantificam sobre entidades ou eventos recebem a etiqueta Q, a qual podem ser associadas etiquetas flexionais. No português moderno, a propriedade quantificacional aplicada a entidades pode ser neutralizada a depender da posição, dentro do sintagma nominal, do item que a expressa. Assim, itens geralmente classificados como quantificadores em posição pré-nominal podem ser interpretados como adjetivos, quando em posição pós-nominal. No presente sistema de anotação, desenvolvido para ser aplicado ao português clássico e moderno, decidimos pela posição de acordo com a qual os quantificadores, contrariamente aos adjetivos e aos verbos, são etiquetados /Q em qualquer contexto,

inclusive quando tais elementos têm distribuição e interpretação de itens adjetivais ou são precedidos de determinante e não seguidos de nome [19].[13] Esse tratamento diferenciado corresponde ao reconhecimento de que se trata de uma classe fechada com propriedades semânticas muito particulares. A sua recuperabilidade enquanto classe, seja qual for o contexto, é assim privilegiado no tratamento dessas palavras.

Finalmente, quanto aos advérbios, aos denominados *de intensidade*, identificados na literatura como quantificadores de evento, é atribuída a etiqueta Q, restando a etiqueta ADV propriamente aos advérbios locativos, de tempo e de modo.

4 Estudos de Caso

4.1 A Estrutura Interna do Sintagma Nominal

Nomes e verbos. É uma característica recorrente da fase do português que consideramos no nosso corpus a sequência determinante-verbo:

(9) o seu *cansar* e o seu *folgar* consiste em mui diferentes empregos do nosso ./. (*in* Sousa)

Como etiquetar os verbos nesses contextos? Do ponto de vista distribucional, encontram-se exatamente na posição de um nome. Devemos então etiquetá-los como nome, tornando assim a tarefa do etiquetador automático mais simples? Num primeiro momento, assumimos essa posição, tomando a etiqueta N como *default* nesses casos. Ou seja, só etiquetariamos como VB os infinitivos precedidos de artigo quando houvesse claras marcas da sua natureza verbal, ou bem marcas flexionais, ou bem a presença de argumentos não preposicionados, como em (10):

(10) E assim a êle se deve, depois de Deus, o *conservar/VB* as fazendas; a êle o *apertar/VB* e *intimidar/VB* o inimigo, sendo a uns freio para o não seguirem, e a outros espora para o perseguirem (*in* Vieira)

Desse ponto de vista, os infinitivos do exemplo (9) têm que ser considerados como Ns. Ora, esta decisão nos apareceu rapidamente como inadequada, por ocultar uma construção muito frequente do português do século 17, codificando-a como um sintagma nominal normal e portanto prejudicando o nosso objetivo de facilitar o estudo da sintaxe do período. Desse último ponto de vista, é VB e não N que deve ser tomado como etiqueta *default*[14].

Encontramos aqui um caso claro de contradição entre a simplicidade computacional, que favoreceria a ocorrência da mesma categoria num mesmo contexto, e a recuperabilidade da informação sintática, no caso a distribuição das formas infinitivas na história da língua. O privilégio que damos à segunda se justifica pelo objetivo do *Corpus Tycho Brahe*.

A estrutura interna dos DPs e a robustez das classes. À primeira vista, essa escolha cria um segundo problema: o da adequação sintática. Com efeito, os sintagmas que contêm as formas infinitivas em (9) e (10) têm distribuição de sintagmas nominais na oração, sendo respectivamente sujeito e objeto direto. O fato de tais formas infinitivas serem etiquetadas como VB parece então dificultar a tarefa do analisador sintático. Porém esse problema se resolve facilmente se, em lugar de marcar sintagmas nominais como NPs, o parser trabalha com a categoria DP, assinalada claramente pela presença do determinante.

Essa decisão mostra-se adequada não só para o problema que acabamos de discutir como para os vários outros casos da língua portuguesa em que o determinante é seguido de categoria diferente de N. Note-se que uma decisão baseada em meras considerações distribucionais nos levaria a etiquetar como N muitas outras palavras que obviamente não pertencem a essa classe. O uso da categoria DP permite facilitar uma codificação unitária das palavras pertencentes a classes fechadas como os advérbios ou os quantificadores, facilitando tanto a recuperabilidade da informação – de novo, trata-se de não ocultar fenômenos sintáticamente relevantes, como a grande diversidade da estrutura interna dos DPs [30] - quanto a decidibilidade, uma vez que se limita drasticamente o número de etiquetas suceptível de ser atribuido a um mesmo item lexical. Veja-se por exemplo o item lexical *mais*, que conforme a sua distribuição, teria que ser considerado advérbio (11) ou nome (12), mas que em função do raciocínio acima deverá receber sempre a mesma etiqueta: ADV-R.

(11) a/D-F melhor/ADJ-R-G pintura/N é/SR-P a/D-F que/WPRO *mais/ADV-R* se/SE parece/VB-P com/P a/D-F obra/N da/P+D-F natureza/N (*in* Lobo)

(12) Tudo/Q isto/DEM digo/VB-P a/P Vossa/PRO$-F Paternidade/NPR como/CONJS a/P quem/WPRO devo/VB-P dar/VB conta/N do/P+D meu/PRO$ espírito/N ,/, e/CONJ ,/, como/CONJS for/SR-SR tempo/N ,/, darei/VB-R do/P+D *mais/ADV-R* da/P+D-F minha/PRO$-F vida/N ./. (*in* Chagas)

Limites da análise unitária: a focalização. A atribuição de uma mesma etiqueta a uma mesma palavra tem contudo limites. Tomemos, por exemplo, o vocábulo *mesmo*. O seu uso mais freqüente é de adjetivo precedido de determinante e seguido de nome, apesar de também ocorrer apenas acompanhado por determinante. Conforme o raciocínio acima, *mesmo* deve ser etiquetado como adjetivo em ambos os casos. Entretanto, em certos contextos, *mesmo* aparece claramente numa posição não acessível a outros adjetivos: antes de determinante ou pronome (*mesmo o rapaz/ele*)[15] ou depois de pronome pessoal ou demonstrativo (*ele/isto mesmo*). Nesses casos, trata-se muito claramente de focalização do sintagma que segue ou do pronome que antecede. Recorremos então à etiqueta FP, que, além de ser descritivamente adequada, deve facilitar a tarefa de análise sintática subsequente, ao excluir a possível análise dessas sequências como mini-orações adjetivais, única construção em que um verdadeiro adjetivo poderia ocorrer neste mesmo contexto (*bonito, o rapaz!* ou *Acho isto interessante*). Quando *mesmo* aparece em posição pré-nominal, a interpretação de focalização é eventualmente disponível, mas em muitos casos fica sujeita a variação

na interpretação. Nesses casos, assumimos portanto o valor *default* da posição, que é ADJ. Observe-se que as considerações de ordem distribucional, que afetam a simplicidade computacional e de adequação sintática vão nesse caso no mesmo sentido, e permitem chegar a uma regra simples na atribuição das duas etiquetas ao mesmo item lexical.

4.2 Etiquetagem da Forma *Que*

A forma *que* traz muitas vezes ambiguidade entre *que* conjunção explicativa, etiquetada /CONJ, e *que* relativo, etiquetado /WPRO. Em caso de dúvida, o valor *default* é /WPRO, porque facilita o estudo das estratégias de relativização. A terceira função de *que* é de puro complementizador (/C), introduzindo o complemento de um verbo ou de um nome, e também de uma preposição e locuções como *já que*, *ainda que,* etc. Um problema interessante surge com as orações relativas com pronome-lembrete. Do ponto de vista da gramática gerativa, não há propriamente relativização, já que não há movimento, e *que* é simplesmente um complementizador. Contudo, não nos parece correto atribuir-lhe a etiqueta /C por duas razões: por um lado, não haveria base para uma distinção sintática entre orações relativas e complementos de nome; por outro lado, dificultaria a recuperação de dados relativos às construções relativas.

4.3 Questões de morfologia flexional: formas verbais em -ra infinitivos e particípios

Neste item, trataremos da etiquetagem de formas verbais baseada na morfologia explícita, em particular no que diz respeito à terminação verbal -ra e às formas participiais.

Terminação Verbal em -ra. Quanto às formas verbais terminadas em -ra, observa-se que estas têm distribuição sintática bastante distinta quando períodos diferentes da língua são comparados. A título de exemplificação, se no português moderno é francamente preferencial a presença do subjuntivo passado em sentenças condicionais (*se houvesse*), no português clássico formas em -ra são sistematicamente encontradas em tal contexto (*se houvera*). Além disso, sentenças-matriz do português moderno nas quais são encontradas formas do futuro do pretérito apresentam no português clássico formas em -ra. Tais distinções certamente constituem um profícuo tema para investigação diacrônica. Visando a privilegiar a recuperabilidade de tais formas, independentemente do contexto de sua ocorrência ou seu valor condicional, mais-que-perfeito ou hipotético, tais itens são, segundo nosso sistema de anotação, sistematicamente etiquetados como /RA.

Imperativo. Para o tratamento do imperativo, incorporamos em nosso sistema de anotação uma observação já amplamente divulgadas na literatura sintática sobre línguas românicas – o fato de que apenas formas com marcação morfológica especificamente imperativa (2as pessoas do singular e plural) apresentam comportamento sintático particular. Assim sendo, apenas a estas aplicamos a sub-etiqueta -I, indicadora de imperativo (*aceitai/VB-I vós* vs. *aceite/VB-SP Vossa Mercê*).

5 Conclusões

Aplicado a textos de autores nascidos nos séculos XVI (*Luis de Sousa*) e XVII (*António Vieira*[16] e *António das Chagas*)[17] – perfazendo um total de 150.000 palavras –, o sistema de anotação ora proposto mostrou-se lingüísticamente consistente, no sentido de não parecer haver distorções de informação lingüística em qualquer dos itens lexicais em questão. Note-se que esta observação está fortemente baseada no processo de correção manual dos arquivos etiquetados automaticamente, e não propriamente numa avaliação de efetiva produtividade do etiquetador automático.[18] Note-se ainda que, eventualmente, o fato de os arquivos etiquetados automaticamente terem sido corrigidos inicialmente por mais de um lingüista, ou de os textos neles contidos serem de gêneros diferentes poderia ser um fator complicador à obtenção da consistência e sistematicidade mencionada. Entretanto, um vez confrontado o trabalho dos diferentes lingüistas frente aos textos citados, o resultado alcançado indica que o sistema de anotação proposto é suficientemente robusto e sistemtático.[19] O futuro retreinamento do etiquetador automático com base nos textos manualmente corrigidos e subseqüente submissão a este dos demais textos previstos no *Corpus Tycho Brahe* inidicará com mais precisão quão robusto e sistemático o sistema proposto deve efetivamente ser considerado.

Referências Bibliográficas

1. Borer, H.: The Syntax of Pronominal Clitics. (Syntax and Semantics, Vol. 19). Orlando, Fla., Academic Press (1986)
2. Britto, H. & Finger, M.: Constructing a Parsed Corpus of Historical Portuguese. Proceedings of the International Humanities Computing Conference ACH-ALLC'99 (1999) 234-235 (http://www.iath.virginia.edu/ach-allc.99/proceedings/britto.html)
3. Britto, H., Galves C., Ribeiro, I., Augusto, M. & Scher, A.: Morphological Annotation System for Automatic Tagging of Electronic Textual Corpora: from English to Romance Languages. Proceedings of the 6th International Symposium of Social Communication (1999) 582-589
4. Burtler, T., Fisher, S., Hockey, S., Coulombe, G., Clements, P., Brown, S., Grundy, I., Carte, K., Harvey, K. & Wood, J.: Cam a Team Tag Consistently? Experiences on the Orlando Project. Proceedings of the International Humanities Computing Conference ACH-ALLC'99 (1999) 234-235 (http://www.iath.virginia.edu/ach-allc.99/proceedings/ burtler.html)
5. Cardinaletti, A. & Starke, M.: The Typology of Structural Deficiency. Manuscrito não-publicado. Univ. Veneza/Univ. Genebra (1994)
6. Chacur, D. & Finger, M.: Etiquetagem do Português Clássico Baseado em Corpus. Artigo submetido ao IV PROPOR. Évora, Portugal
7. Chomsky, N.: Lectures on Government and Binding. Foris, Dordrecht (1981)
8. Chomsky, N.: Barriers. The MIT Press, Cambridge, Massachusetts (1986)
9. Cinque, G.: Types of A' Dependencies. The MIT Press, Cambridge, Massachusetts (1990)

10. Finger, M.: Tagging a Morphologically Rich Language. Proceedings of the 1st Workshop on Text, Speech and Dialogue TDS'98 (1998) 39-44
11. Galves, C.: Clitic Placement in Europen Portuguese: Evidence for a Non-homogeneous Theory of Enclisis. Workshop sobre o Português. Associação Portuguesa de Lingüística, Lisboa.
12. Jaeggli, O.: Topics in Romance Syntax. Foris, Dordrecht (1982)
13. Jaeggli, O. & Safir, K.: The Null Subject Parameter. Foris, Dordrecht (1989)
14. Kayne, R.: Null Subject and Clitic Climbing. In: Jaeggli & Safir, *op.cit.* (1989)
15. Kayne, R.: Romance Clitics, Verb Movement, and PRO. Linguistic Inquiry 22 (1991) 647-686
16. Koster, Jan & May, R: On the Constituency of Infinitives. Language 58-1 (1982) 117-143
17. Lakoff, R.T.: Abstract Syntax and Latin Complementation. The MIT Press, Cambridge, Massachusetts (1968)
18. Martins, A.M.: Clíticos na História do Português. Tese de Doutorado. Univ. Lisboa (1994)
19. Mateus, M.H.M., Brito, A.M., Duarte, I.S. & Faria, I.H. Gramática da Língua Portuguesa. Livraria Almedina, Coimbra (1983)
20. McCloskey, J.: Inflection and Conjuction in Modern Irish. Natural Language and Liguistic Theory 4 (1986) 245-282.
21. Pollock, J-Y.: Verb Movement, Universal Grammar, and the Structure of IP. Linguistic Inquiry 20 (1989) 365-424
22. Raposo, E.: Romance Infinitival Clauses and Case Theory. In: Neidle, C. & Nunez-Cedeno, R. (eds): Studies in Romance Languages. Foris, Dordrecht (1987) 237-249
23. Raposo, E.: Teoria da Gramática: a Faculdade da Linguagem. Caminho, Lisboa (1992)
24. Rizzi, L.: Null Objects in Italian and the Theory of *pro*. Linguistic Inquiry 17 (1986) 501-557.
25. Rizzi, L.: Relativized Minimality. The MIT Press, Cambridge, Massachusetts (1990)
26. Rizzi, L.: A Parametric Approach to Comparative Syntax: Properties of the Pronominal System. In: Haegeman, L.: The New Comparative Syntax. Longman, Londres (1997)
27. Stowell, T.: Origins of Phrase Structure. Tese de Doutorado. MIT (1981)
28. Taylor, A. & Kroch, A.: The Penn-Helsinki Parsed Corpus of Middle English II. Manuscrito não publicado. Univ. Pensilvânia (1998)
29. Torres-Morais, M.A.: Do Português Clássico ao Português Europeu Moderno: Um Estudo da Cliticização e do Movimento do Verbo. Tese de Doutorado. UNICAMP (1995)
30. Zamparelli, R.: Layers in DP: the basic idea (http://ww.cogsci.ed.ac.uk/~roberto/ layers/basic.html) (1996)

---

* O projeto de pesquisa *Padrões Rítmicos, Fixação de Parâmetro e Mudança Lingüística* (coord. Charlotte Galves), no qual o presente trabalho se insere, é financiado pelo Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) - Proc. 98/03382-0.

+ A pesquisadora, pós-doutoranda, conta com auxílio financeiro da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) - Proc. 98/12074-3.

[2] Atualmente, dispomos de dez (10) arquivos ortograficamente transcritos e três (03) arquivos morfologicamente etiquetados (cf. http://www.ime.usp.br/~tycho/corpus).

[3] Conforme discutido em [4], aplicar etiquetas consistentemente é um dos princípios metodológicos fundamentais na elaboração de grandes *corpora* anotados.

[4] Para uma visão completa das etiquetas do *Corpus Tycho Brahe* e suas aplicações, v. http://www.ime.usp.br/corpus/manual/etiq_english.html.

[5] Para as várias distinções sintáticas entre sentenças finitas vs. não-finitas (com ou sem clíticos), v. [17], [7], [27], [16], [1], [8], [20], [22], [14], [21], [9], [15], [23], [11], dentre outros.

[6] Como sugerido por um dos pareceristas do *IV PROPOR*, e também como já proposto anteriormente [3], o presente sistema de anotação prevê etiquetas distintas para particípios com função adjetival e passiva (-AN) vs. verbal (-PP). Observe-se que, frente a eventuais ambiguidades entre essas formas, utilizamos como *default* a etiqueta -PP.

[7] Para a referência completa dos textos donde os exemplos presentes neste trabalho foram extraídos, cf. http://www.ime.usp.br/~tycho/corpus/list_txt/list.html.

[8] Embora relevante do ponto de vista lingüístico, a não-marcação de traços de pessoa num corpus que trata da história do português europeu não é problemática, uma vez que, durante toda sua história, a língua nunca perdeu as propriedades sintáticas (ordem VS generalizada e presença de sujeitos nulos referenciais) que parecem ser motivadas por tais traços (cf. [7], [12], [24], [13], [25], [26]).

[9] No que diz respeito ao infinitivo, por exemplo, contrariamente à gramática tradicional, não fazemos distinção de tratamento entre uma forma de 3a pessoa e uma forma não flexionada.

[10] No decorrer do século XIX (período em que já ocorrera a mudança gramatical que levou à atual agramaticalidade de sentenças como (a) e (b) (cf. [18]; [29]), ainda se observam nos textos vários exemplos como (a), mas dificilmente (b):

(a) Ella te leva também uns rebuçados (*in* Garrett, A. (1799-1854) *apud* [29])
(b) e a costureira o demorou (in Garrett, A. (1799-1854) *apud* [29])

[11] A divisão entre nomes comuns e próprios é também motivada por seus distintos comportamentos sintáticos, já observados anteoriormente [7].

[12] Anthony Kroch (comunicação pessoal) argumenta que a preposição está visível na formação de palavras como *contudo*, *porém*, *porque*, *embora*, *conforme*, etc. Apesar disto, seguindo a tradição gramatical, mantivemos a distinção entre conjunções coordenativas vs. subordinativas. Note-se, entretanto, que a possibilidade de substituir CONJS por P – e, assim, suprimir uma etiqueta – pode ser sempre considerada.

[13] Note-se que o próprio corpus ora em construção deverá permitir um estudo sistemático acerca da distribuição de quantificadores e adjetivos na história do português, que até o momento não foi realizado.

[14] Essa decisão não deixa de ser aparentemente problemática também. Por exemplo, como etiquetar palavras que no português moderno são claramente nominalizadas como *jantar*, *poder*, *ser* (no sentido de ser humano)? A leitura dos textos responde claramente a essa pergunta. Essas palavras têm no português setecentista exatamente o mesmo valor que têm hoje. Veja as seguintes frases:

(a) e/CONJ pois/CONJ lhe/CL descobri/VB-D o/D nome/N ,/, é/SR-P necessário/ADJ ,/, senhor/N Leonardo/NPR ,/, que/C lhe/CL deis/VB-SP agora/ADV o/D ser/N ./. (*in* Rodrigues Lobo)
(b) e/CONJ o/D jantar/N e/CONJ cea/N ia/VB-D todos/Q-P os/D-P dias/N-P da/P+D-F cozinha/N do/P+D Arcebispo/NPR ./. (*in* Frei Luis de Sousa)

Em (a), a palavra *ser* tem claramente o sentido filosófico de "essência", recorrente nos textos considerados, e *jantar* tem seu sentido moderno em (b).

[15] Note-se que esta observação de caráter distribucional se estende também aos nomes próprios.

[16] No que diz respeito a António Vieira, o texto em questão diz respeito à correspondência pessoal do autor.

[17] Para a referência dos textos de tais autores, cf. nota. 7.

[18] Para uma avaliação computacional da aplicação de tais etiquetas, v. Chacur & Finger (1999).

[19] Para verificar tais arquivos etiquetados, v. http://www.ime.usp.br/~tycho/corpus.