Revista RBBA

Revista Binacional Brasil Argentina

O CORPUS TYCHO BRAHE UM CORPUS SINTATICAMENTE ANOTADO DO PORTUGUÊS HISTÓRICO

EL CORPUS TYCHO BRAHE: UN CORPUS SINTÁCTICAMENTE ANOTADO DEL PORTUGUÉS HISTÓRICO

Charlotte Galves
Universidade Estadual de Campinas – UNICAMP/CNPq
galvesc@unicamp.br

Resumo

Este artigo apresenta a metodologia de trabalho de construção e uso do *Corpus Sintaticamente Anotado do Português Histórico Tycho Brahe*. Descreve a ferramenta de edição eletrônica *eDictor*, bem como o sistema de etiquetagem de palavras e de anotação sintática aplicado aos textos. Exemplifica o funcionamento e uso da linguagem de busca *Corpus Search*, que procura construções em arquivos sintaticamente anotados, a partir de perguntas de pesquisa. Mostra os avanços da pesquisa sobre a história do português europeu permitidos pela grande quantidade de dados anotados disponíveis no *Corpus Tycho Brahe* para o período dos séculos 16 a 19. Na conclusão, evoca-se a extensão do trabalho ao português brasileiro, a favor da convergência entre a metodologia proposta e o trabalho intensivo de edição de documentos históricos de diversas procedências socio-culturais e geográficas.

Palavras-chave: Corpus Tycho Brahe. Anotação sintática. História do português. Português brasileiro.

Resumen

Este artículo presenta la metodología de trabajo de construcción y uso del Corpus Sintácticamente Anotado del Português Histórico Tycho Brahe. Describe la herramienta de edición electrónica eDictor, así como el sistema de etiquetado de palabras y anotación sintáctica aplicada a los textos. Ejemplifica el funcionamiento y el uso del lenguaje de búsqueda Corpus Search, que busca construcciones en archivos sintácticamente anotados, a partir de preguntas de investigación. Muestra los avances de la investigación sobre la historia del portugués

europeo permitidos por la gran cantidad de datos anotados disponibles en el Corpus Tycho Brahe para el período comprendido entre los siglos XVI y XIX. En la conclusión, se evoca la extensión del trabajo al portugués brasileño, a favor de la convergencia entre la metodología propuesta y el intenso trabajo de edición de documentos históricos de diversas procedencias socioculturales y geográficas.
Palabras clave: Corpus Tycho Brahe. Anotación sintáctica. Historia del portugués. Portugués brasileño.

1- Introdução

Um corpus é uma coleção de textos compilados para um determinado fim. O fim do Corpus Tycho Brahe (doravante CTB) é fornecer dados que nos permitam escrever em detalhes a história do português, em particular no que diz respeito à sua sintaxe. Inicialmente concebido para trazer informações sobre a dinâmica da língua em Portugal entre o século 16 e o século 19, sua vocação é agora estender seu escopo à história do português brasileiro. O CTB tem a particularidade de ser anotado morfossintaticamente, o que significa que cada palavra vem acompanhada de uma etiqueta que expressa propriedades morfológicas dessa palavra, e cada frase vem com sua estrutura sintática. Isso permite fazer buscas não só por itens lexicais, mas também por classe de palavras e por estrutura sintática. Na primeira parte deste artigo, apresentarei as ferramentas de formatação e anotação usadas para a construção do CTB. Na segunda parte, mostrarei como, graças a informação morfossintática acrescentada aos textos pela anotação, grandes quantidades de dados podem ser exploradas de maneira automática e confiável. Enfim, a terceira parte trará alguns exemplos do conhecimento novo que foi possível obter graças ao CTB e outros corpora históricos do português construídos nos mesmos moldes. Terminarei com algumas perspectivas para o futuro.

2- As ferramentas de construção do *Corpus Tycho Brahe*

A primeira preocupação do linguista ao trabalhar com textos antigos é a preservação das suas características originais. Apesar da necessidade de padronizar os textos para poder rodar neles ferramentas computacionais de anotação e busca, é imprescindível manter o acesso ao texto original. A ferramenta de edição eletrônica "eDictor" foi construída com esse objetivo em mente. As Figura 1 e 2 mostram detalhes da interface de eDictor.

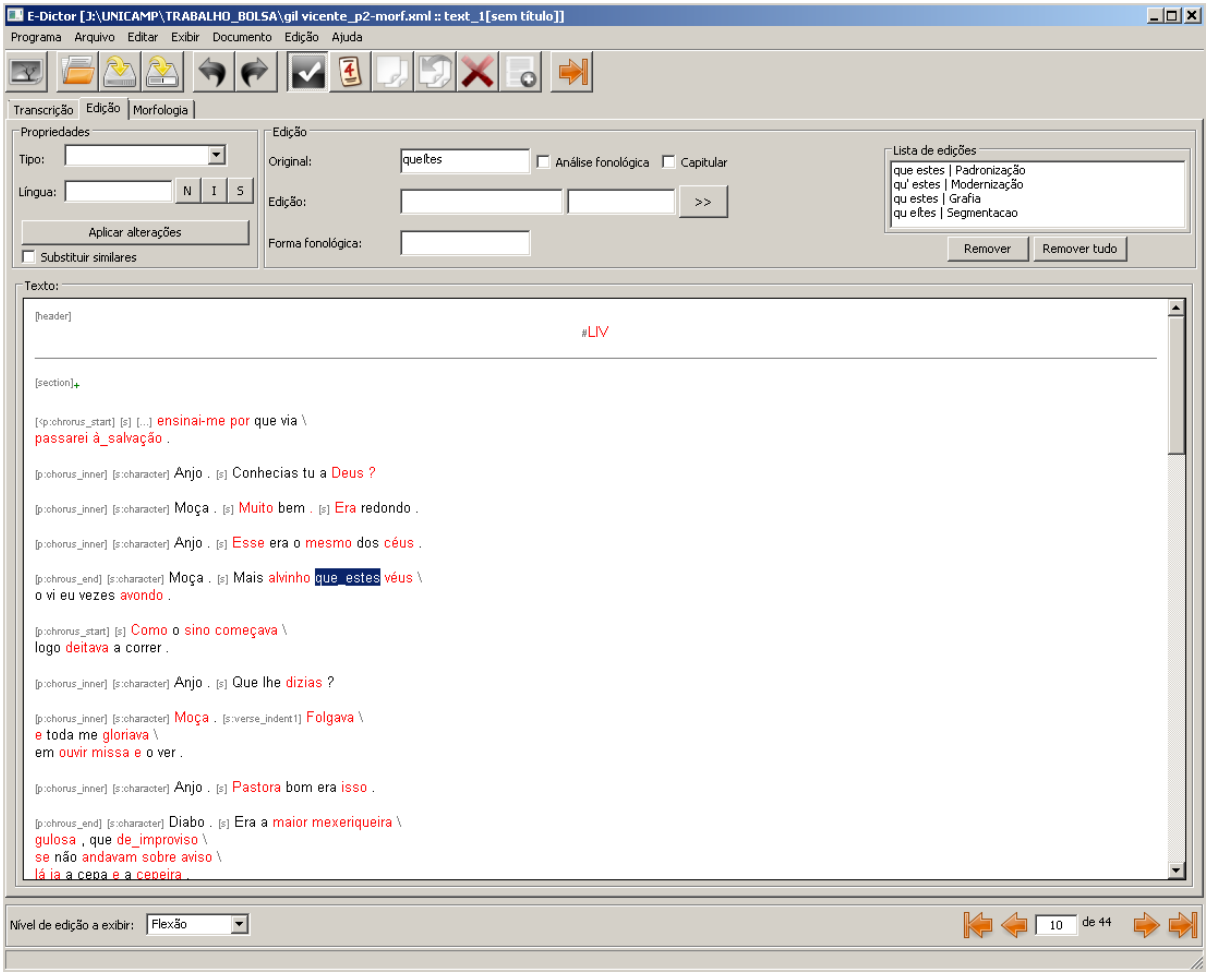

Figura 1- Tela de edição de eDictor
Fonte: Elaboração própria

Na Figura 1, que exibe um trecho do autor quinhentista Gil Vicente, vê-se o procedimento de edição da ferramenta na lista de edições no canto superior direito. A partir da forma *queſtes* (com o ſ da tipografia quinhentista), aplicam-se sucessivamente as operações de "segmentação", que restitui duas palavras independentes (*qu eſtes)*, de "grafia", que substitui o tipo antigo ſ pelo moderno (*qu estes)*, de "modernização", que cria a sequência *qu'estes*, e enfim de padronização, produzindo a forma padrão hoje *que estes*. Como se vê na Figura 2, esses processos são ordenados. Eles podem ser modificados pelos usuários em função das suas necessidades, num arquivo de preferências. Na interface de leitura, as palavras que aparecem em vermelho são palavras que foram modificadas, os traços entre palavras como em *de_sexta-feira* são a marca do processo de segmentação. Uma funcionalidade essencial desse sistema é que o documento pode ser exibido e exportado em qualquer nível de edição, ficando o texto gerado para leitura ou processamento mais ou menos distante do original. No canto

inferior esquerdo da figura 2 – agora com um texto de jornal do séc. 18, vê-se que o "nível de edição a exibir" é o último da lista, intitulado "flexão", aquele em que, entre outros aspectos, desvios de concordância são corrigidos. Note-se que esse nível não costuma ser exportado para as fases de anotação, uma vez que os desvios flexionais são fenômenos morfossintáticos que interessam a análise histórica e devem ser mantidos na descrição e análise.

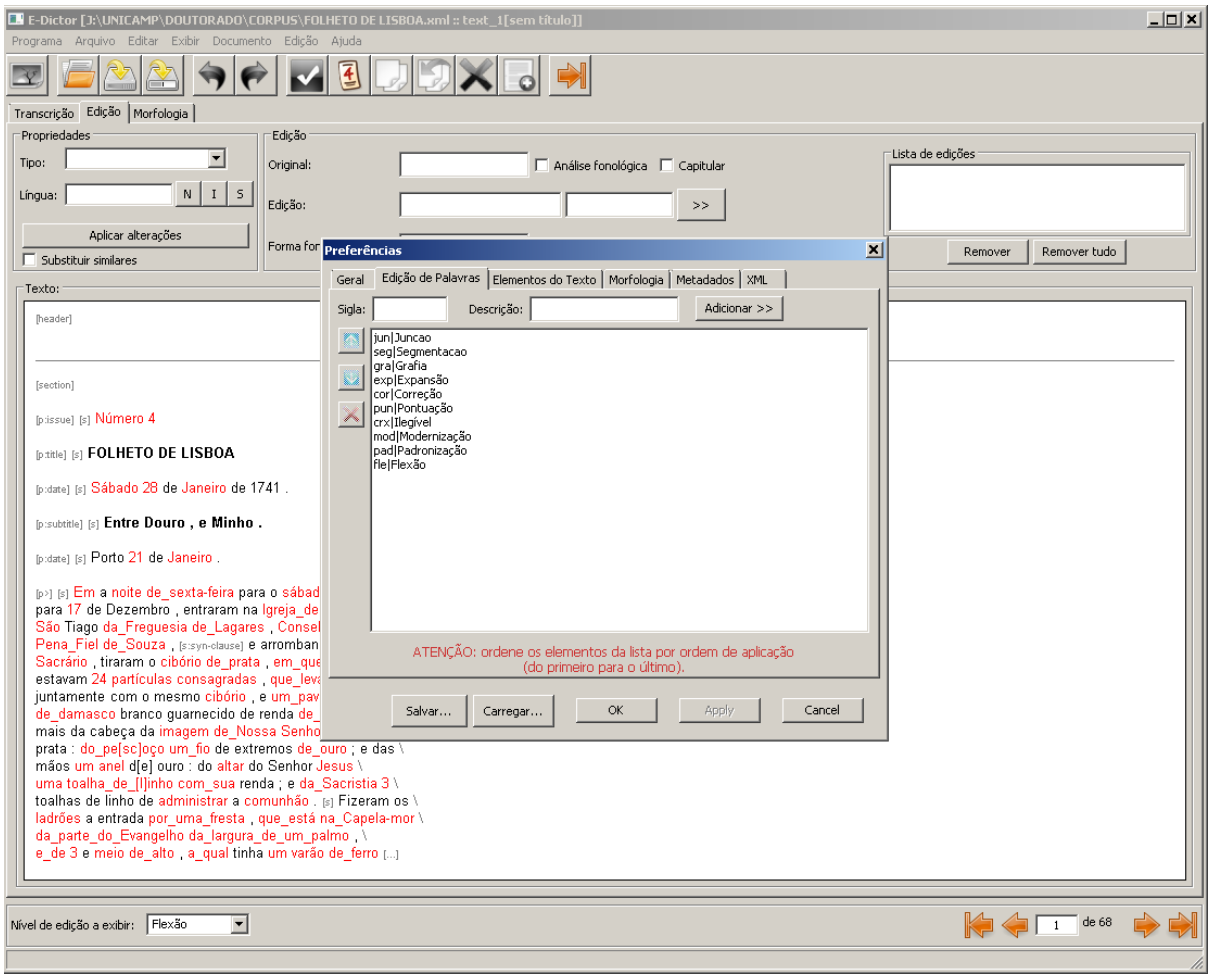

Figura 2- Tela de eDictor com janela de preferências para edição de palavras
Fonte: Elaboração própria

Como se vê nas abas da janela, a marcação do texto pode também ser efetuada a nível dos elementos textuais, com definição de tipos de parágrafos como "abertura", "título", "saudação", etc... Nas preferências da morfologia estão as etiquetas de palavras sobre as quais voltaremos logo abaixo, e nas preferências dos metadados estão listadas classes relacionadas a todas as possíveis informações extra-textuais, como data, autoria, fonte, etc... além das

informações sobre o processamento dos arquivos. Esses metadados estão na base da geração das fichas catalográficas de cada texto, bem como do catálogo como um todo (cf. Figura 3).

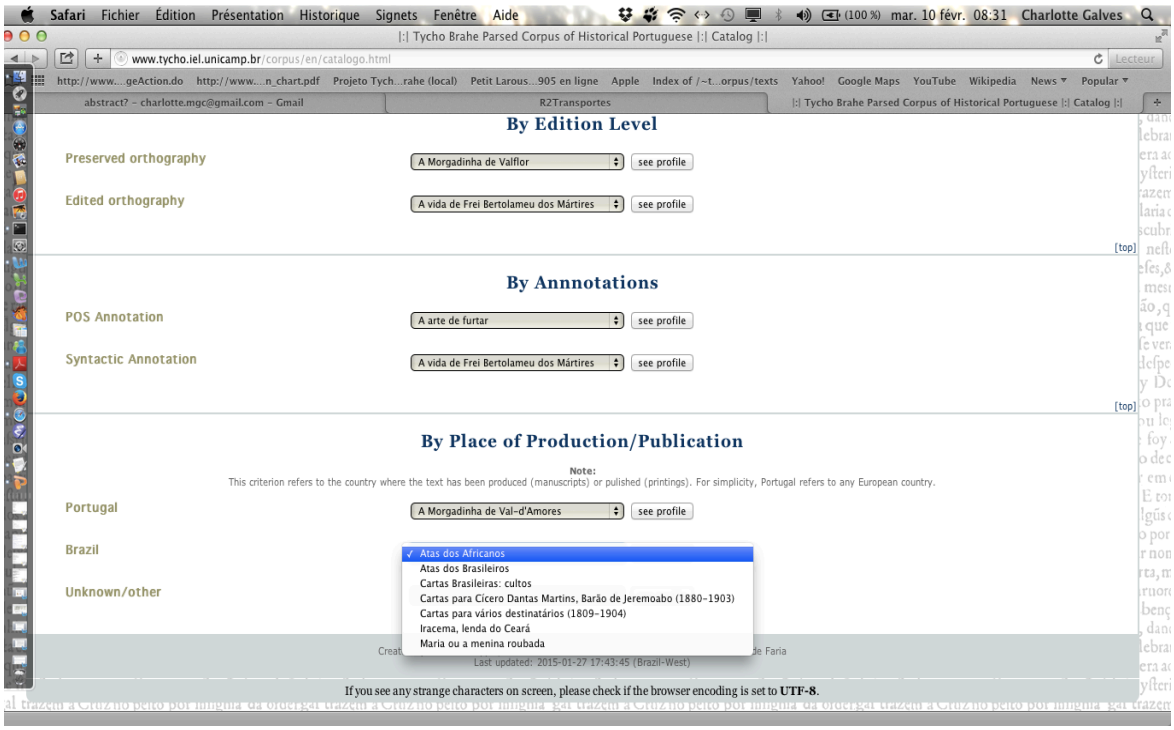

Figura 3- Parte do Catálogo de listas ordenadas do Corpus Tycho Brahe gerado por eDictor
Fonte: Elaboração própria

Ao clicar na aba "Morfologia" na terceira linha do canto superior esquerdo da tela do eDictor (cf. Figuras 1, 2 e 4), tem-se acesso à função "etiquetar", pela qual as palavras recebem como marcação suplementar uma etiqueta indicando sua classe e alguns traços gramaticais que lhe são associados[i]. A inclusão de um etiquetador de palavras na ferramenta torna eDictor um instrumento particularmente útil e amistoso para quem está interessado no estudo morfossintático dos textos. O exemplo 1 ilustra a aplicação dessas etiquetas à versão padronizada de uma sentença do texto "Atas dos Brasileiros"[ii].

(1) Do texto original à versão etiquetada

a. Original:

a prezentou huma Planta offerecida pello Il*ustríssi*mo Se*nhor* Jozé Corréa Machado arquiteto da Provincia, munto digno Socio Protetor da nossa Sociedade e por este mesno S*enhor* foi nos derigido o Conpetente Orcamento da Obra

b. Versão etiquetada

apresentou/VB-D uma/D-UM-F planta/N oferecida/VB-AN-F pelo/P+D Ilustríssimo/ADJ-S Senhor/NPR José/NPR Corréa/NPR Machado/NPR arquiteto/N da/P+D-F província/N ,/, muito/Q digno/ADJ sócio/N protetor/ADJ da/P+D-F nossa/PRO$-F Sociedade/NPR e/CONJ por/P este/D mesmo/ADJ senhor/NPR foi/SR-D nos/CL dirigido/VB-AN o/D competente/ADJ-G orçamento/N da/P+D-F obra/N

De a. para b. percebemos que foram aplicados processos de edição, como a junção de palavra em *apresentou* e a padronização da ortografia. As etiquetas são atribuídas a essas palavras normalizadas. Percebe-se que essas etiquetas têm uma base, VB para verbo, N para nome, ADJ para adjetivo, etc... à qual se acrescenta em vários casos uma ou várias sub-etiquetas separadas por hífen, como em VB-AN-F onde AN significa "particípio passado" e F significa "feminino". O símbolo +, encontrado em da/P+D-F, expressa a contração entre duas palavras, no caso a preposição *de* e o artigo feminino *a*. Como mencionado logo abaixo a respeito da anotação sintática, a categorização das palavras é em parte tradicional e em parte inspirada da teoria gerativa, como no caso das palavras relativas e interrogativas cujas etiquetas começam com W, por serem palavras que em inglês começam com W (o Q das línguas românicas).

A etiquetagem é feita por um etiquetador automático probabilístico, treinado com dados do português. Sua taxa de acerto é de cerca de 95%, o que significa que uma correção manual é necessária. De novo, eDictor propicia um ambiente confortável para essa tarefa, ao permitir correr de etiqueta para etiqueta, modificando-as quando necessário. Na Figura 4, observa-se que é possível ir corrigindo as etiquetas erradas, com a ajuda de funções como "substituir similares".

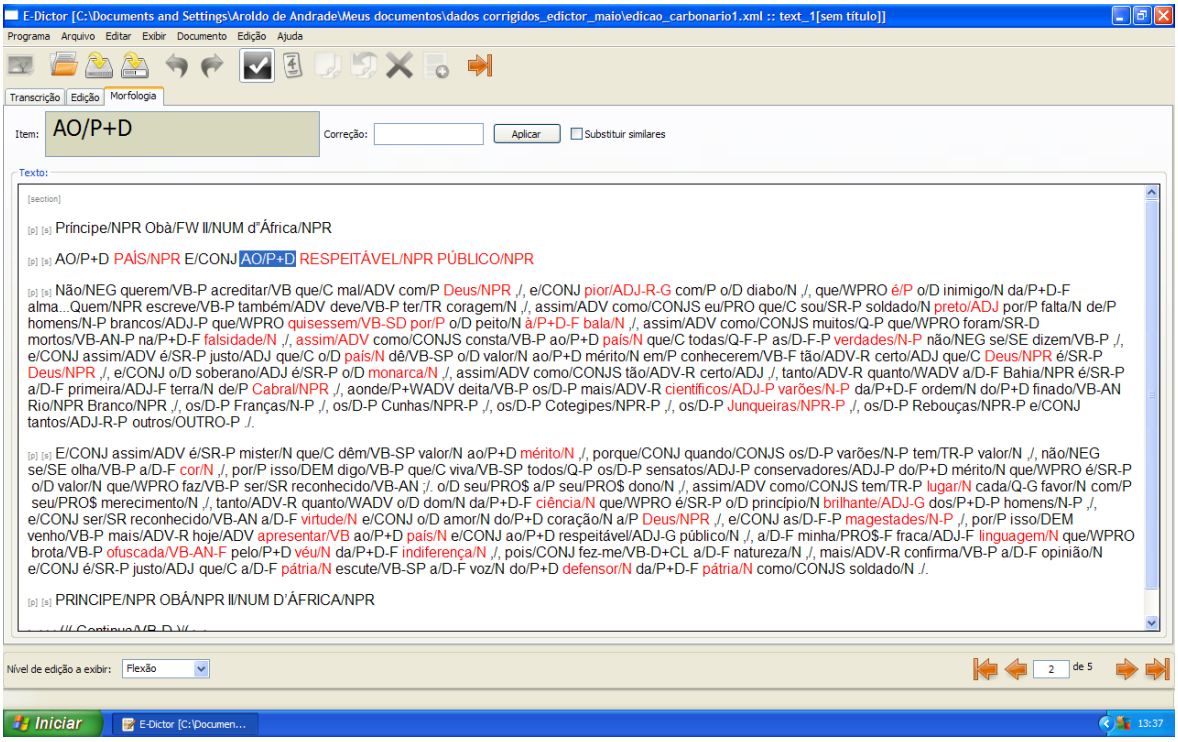

Figura 4- Correção de etiquetas de palavras com eDictor
Fonte: Elaboração própria

A versão etiquetada dos textos pode ser usada para certos tipos de buscas (cf. nota xiv), mas é sobretudo a base para a aplicação de analisadores sintáticos, ou "parsers", que associam uma estrutura sintática a cada frase. Por exemplo, (2) representa a estrutura associada ao exemplo (1):

(2) Representação da estrutura sintática

((IP-MAT (VB-D Apresentou)
(NP-ACC (D-UM-F uma)
(N planta)
(ADJP (VB-AN-F oferecida)
(PP (P pel@)
(NP (D @o)
(ADJ-S Ilustríssimo)
(NPR Senhor)
(NP-PRN (NPR José) (NPR Corréa) (NPR Machado))
(NP-PRN (N arquiteto)
(PP (P da@)
(NP (D-F @a) (N província))))
(, ,)
(NP-PRN (ADJP (Q muito) (ADJ digno))
(N sócio)
(NP-PRN (N protetor))
(PP (P da@)

- (NP (D-F @a) (PRO$-F nossa) (NPR Sociedade)))))))))
- (ID VA_002_SPL,19.113))

Além das etiquetas já presentes em (1), a representação em (2) contém categorias sintagmáticas como IP-MAT, NP-ACC, NP-PRN, ADJP, PP. A primeira é o nó máximo de uma oração afirmativa não dependente. As outras são projetadas, respectivamente, a partir dos núcleos N, ADJ e P. O sistema articula teoria gramatical de inspiração gerativista e gramática tradicional. Da teoria gramatical vem a própria representação sintagmática, com os rótulos categoriais correspondendo à projeção dos núcleos[iii], e para as orações, a nomenclatura IP e CP (cf. CP-QUE para a oração interrogativa no exemplo 3). A gramática gerativa também inspira a representação de categorias vazias para marcar, por exemplo, o movimento de palavras interrogativas ou relativas (cf. ex. 3), ou de pronomes clíticos em construções ditas de alçamento (ex.4) ou ainda o sujeito nulo (ex.5)[iv]. Essas categorias vazias são respectivamente marcadas por *T*, ***pro* e *** [v].

(3) ((CP-QUE **(WPP-1 (P A)**
(NP (WPRO quem)))
(IP-IND **(PP *T*-1)**
(VB-D deu)
(NP-SBJ (D este) (N criminoso))
(NP-ACC (D-F essa)
(N soma)
(PP (P de)
(NP (N dinheiro)))))
(. ?))
(POST SCRIPTUM; ID CARDS0094,.7))

(4) ((IP-MAT (NP-SBJ (NPR Deus))
(NP-1 (CL os))
(VB-D queria)
(IP-INF **(NP-ACC *-1)**
(VB juntar))
(. !))
(TYCHO BRAHE; ID A_003_PSD,45.654))

(5)((IP-MAT**(NP-SBJ *pro*)**
(VB-D-1P Fazíamos)
(NP-ACC (D-F-P as)
(N-P maranhas)

(PP (P @de)
(NP (D-P @os) (N-P cobertores))))
(. .))
(CORDIAL-SIN; ID MST18,23))

Da gramática tradicional vem a marcação das funções como sujeito e objeto (NP-SBJ, NP-ACC), uma vez que, sendo a estrutura bastante simplificada em relação à teoria sintática[vi], é necessário associar às categorias as funções que elas desempenham na oração, como sujeito (SBJ), objeto direto (ACC), aposto (PRN), etc... A anotação utilizada é amplamente baseada no sistema desenvolvido nos *Penn Parsed Corpora of Historical English*, coordenados por Anthony Kroch na Universidade da Pensilvânia[vii]. Adaptações foram feitas para o português nos diversos corpora anotados. Uma apresentação exaustiva do sistema encontra-se no *Portuguese Syntactic Annotation Manual* [viii], com exemplificação extraída dos quatro corpora do português sintaticamente anotados atualmente disponíveis[ix].

A versão de eDictor apresentada aqui não inclui analisador sintático. Deve-se, portanto, exportar uma versão txt da versão etiquetada do texto para rodar o analisador que gera a versão sintaticamente anotada. Até recentemente, usamos para isso o "parser" de Dan Bikel, treinado com os dados do português. Mesmo com o crescimento do corpus de treinamento, o desempenho da ferramenta estagnou em 80% de acertos[x], o que significa um importante trabalho de correção manual a ser realizada. Durante muitos anos, esse trabalho foi feito com *Corpus Draw*, uma interface gráfica incluída no pacote *Corpus Search*[xi]. A Figura 5 mostra a frase representada em (1) e (2) em formato de árvore numa tela de Corpus Draw.

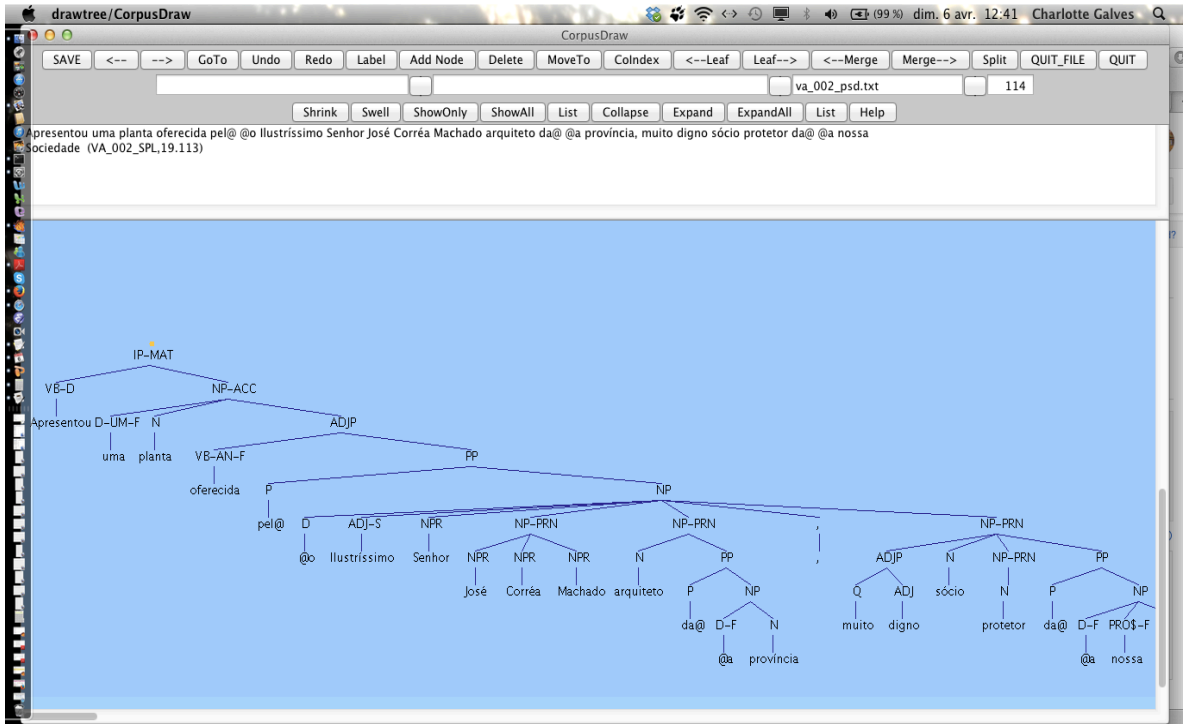

Figura 5- Tela da ferramenta de revisão Corpus Draw
Fonte: Elaboração própria

Os botões da barra superior são usados para realizar as revisões necessárias das estruturas geradas pelo analisador: "Label" permite modificar as etiquetas de palavras e de sintagmas. "Add Node" cria, e "Delete" apaga, nós na estrutura. "Move to" desloca unidades para outras posições. "Coindex" atribui índices idênticos a dois ou vários elementos da estrutura. "Leaf" à esquerda e "Leaf" à direita criam posições vazias. Finalmente, "Merge" junta frases, e "Split" separa frases em duas unidades. O grande mérito de *Corpus Draw* é a excelente visualização que proporciona da estrutura das frases. Mas as categorias inseridas na estrutura têm que ser digitadas na barra branca abaixo dos botões de comando, o que leva tempo e é sujeito a erros. Por isso, *Corpus Draw* vem sendo substituído por uma outra ferramenta de correção, intitulada *Annotald*[xii]. onde as funções não são mais ativadas por botões e comandos manuais, mas por atalhos no teclado e listas de categorias registradas no sistema. O resultado é um ganho enorme de tempo.

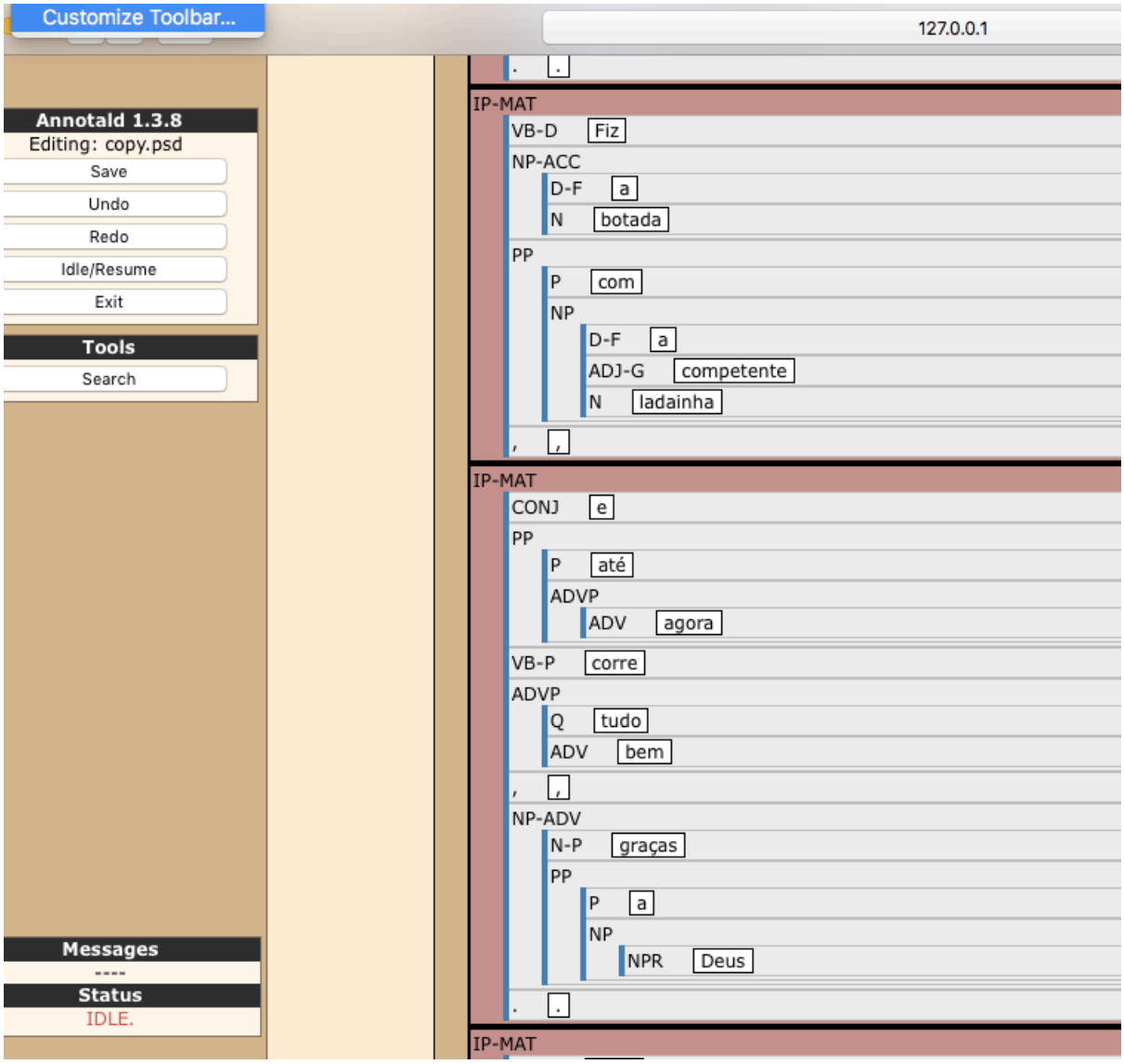

Figura 6- Tela da ferramenta de revisão Annotald
Fonte: Elaboração própria

Além de optar por uma interface de revisão mais ágil, desistimos de recorrer a uma ferramenta probabilística para gerar a anotação sintática. Mesmo que esse tipo de programa seja mais interessante a longo prazo, ele requer uma quantidade de dados demasiado grande para se tornar realmente eficiente num lapso de tempo compatível com os prazos da pesquisa. Foi por isso que foi recentemente adotada uma nova metodologia de trabalho, lançando mão de um analisador baseado na função de revisão da ferramenta *CorpusSearch*, implementado para o português por Catarina Magro[xiii]. O sistema consiste em 11 blocos ordenados de regras que têm a forma de buscas-revisão: dado um certo contexto, aplica-se uma determinada função de revisão de *CorpusSearch*: criar ou apagar nó, gerar uma categoria vazia, co-

indexar, etc... A vantagem de tal sistema é que as regras são feitas para produzir exatamente o que está previsto no sistema de anotação, inclusive categorias vazias e coindexações, que eram introduzidas manualmente na metodologia anterior. Obviamente, a revisão se torna assim menos custosa e menos sujeita a erros. Além disso, o sistema de blocos prevê uma revisão progressiva (no final de cada bloco), que assegura a melhor eficiência do processo nos passos seguintes. Seguindo essa metodologia, a revisão final se torna muito leve, incidindo somente em alguns pontos que escaparam nos passos intermediários.

3- Fazer buscas no Corpus Tycho Brahe

Corpus Search é uma ferramenta desenhada para fazer buscas em corpora sintaticamente anotados[xiv]. Ela permite extrair dos textos construções sintáticas complexas que são relevantes para a análise de determinados fenômenos e, em última instância, para o estudo da dinâmica da língua ao longo do tempo. A guisa de ilustração, apresento uma busca que permite a recuperação de fenômenos de ordem envolvendo a posição relativa de sujeitos pós-verbais, e de advérbios de maneira. Esse tipo de dados é particularmente relevante quando se estuda a estrutura da oração, pois os advérbios de maneira são considerados como ocupando uma posição logo à esquerda do sintagma verbal. Em consequência, se o sujeito pós-verbal precede esses advérbios, isso significa que o sujeito se moveu para fora do sintagma verbal (doravante VP). É importante saber que em línguas românicas como o italiano, os sujeitos pós-verbais sempre seguem os advérbios de maneira, o que sugere fortemente que eles ocupam uma posição interna ao VP, onde recebem uma interpretação marcada de foco informacional. Ao contrário, em línguas escandinavas como o islandês, o sujeito posposto sempre precede o advérbio, o que tende a mostrar que, mesmo estando à direita do verbo o sujeito saiu do VP[xv]. Isso é condizente com a propriedade "V2" das línguas escandinavas que implica uma posição muito alta do verbo, nalgum ponto da periferia esquerda da oração[xvi]. Partindo da hipótese de que o português clássico é uma língua de tipo V2[xvii], é importante verificar a posição relativa dos advérbios de maneira e dos sujeitos pospostos. É claro que a busca manual desse tipo de fenômenos é uma tarefa extremamente custosa se for aplicada a uma quantidade grande de frases, por se tratar de fenômeno não muito frequente, nem muito fácil de localizar. Pela mesma razão, se o corpus for pequeno, o risco é grande de não se achar nenhum dado.

Usando a linguagem de Corpus Search, a busca terá a seguinte forma:

(6) Busca de sentenças com sujeito pós-verbal seguido de advérbio de maneira

```
define: port.def
print_indices: t
node: IP*

query: (tns_vb2 HasSister ADVP*)
AND (ADVP* iDomsOnly ADV)
AND (ADV iDominates *mente)
AND (tns_vb2 HasSister NP-SBJ*)
AND (tns_vb2 precedes NP-SBJ*)
AND (NP-SBJ* precedes ADVP*)
```

Deixando provisoriamente de lado o cabeçalho, observamos nas três primeiras linhas da busca ("query") a referência a ADVP, o sintagma adverbial, e a ADV o advérbio. A primeira condição é de que o sintagma adverbial seja "irmão" ("HasSister") do verbo, referido aqui pela expressão "tns_vb2", que remete a um reagrupamento de etiquetas verbais presente no arquivo de definição chamado "port.def", mencionado no cabeçalho. A irmandade entre nós implica que eles sejam imediatamente dominados pela mesma categoria, no caso IP* (cf. a árvore em 7). Esse mesmo ADVP não pode conter mais (iDomsOnly) que um advérbio. Isso exclui sintagmas adverbiais compostos por mais de um elemento. Finalmente ADV tem que dominar uma palavra terminada em "*mente" (Note que na linguagem de CorpusSearch, * significa "qualquer coisa"). Nas duas linhas seguintes diz-se que o verbo tem como irmão um sujeito, e o verbo precede o sujeito, o que equivale a dizer que o sujeito é pós-verbal. Enfim, a última condição é que o sujeito preceda o sintagma adverbial. Um exemplo de frase achada nos textos por essa busca está em (7), com sua estrutura.

(7) Referiu o depois em Carta sua o mesmo Padre VIEIRA **formalmente** assim:

```
(B_001_PSD,184.1458)
*~/
/*
1 IP-MAT:  2 VB-D, 27 ADVP, 28 ADV, 29 formalmente, 18 NP-SBJ
*/
( (1 IP-MAT (2 VB-D Referiu)
```

(4 NP-ACC (5 CL o))
(7 ADVP (8 ADV depois))
(10 PP (11 P em)
(13 NP (14 N Carta) (16 PRO$ sua)))
(18 NP-SBJ (19 D o) (21 ADJ mesmo) (23 NPR Padre) (25 NPR VIEIRA))
(27 ADVP (28 ADV formalmente))
(30 ADVP (31 ADV assim))
(33 . :))
(35 ID B_001_PSD,184.1458))

(7) mostra o formato em que *Corpus Search* apresenta o resultado das buscas. Primeiro, temos a frase, com o código de seu autor e localização no texto. Aqui trata-se da sentença 1458 da p. 184 do autor André de Barros (B_001). Nessa frase o advérbio de maneira *formalmente* segue o sujeito posposto *o mesmo Padre Vieira.* Na linha seguinte, *Corpus Search* indica onde estão as categorias e palavras requeridas na busca. O domínio é IP-MAT, ou seja, uma oração matriz. Isso faz referência ao cabeçalho em (6) onde o nó sintático pertinente é definido como IP*, ou seja, qualquer tipo de IP. Cada nó da estrutura recebe um índice numérico que permite sua identificação mais rápida na árvore (veja no cabeçalho da busca a indicação de que tal índice tem que estar presente: "print_indices: t")[xviii]. O nó IP-MAT, que domina todos os outros, tem o índice 1. O verbo (VB-D) tem o índice 2, o sintagma adverbial (ADVP) tem o índice 27, o advérbio (ADV) 28, a palavra *formalmente* 29, e finalmente o sujeito (NP-SBJ) tem o índice 18.

Note-se enfim que para achar as frases em que o advérbio não mais segue, mas precede, o sujeito posposto, só precisa modificar a última linha da busca, trocando *AND (NP-SBJ* precedes ADVP*)* por *AND (ADVP* precedes NP-SBJ*).*

Essas buscas podem ser rodadas em grandes quantidades de dados, permitindo achar todos os casos correspondentes às condições expressas, e só esses. É, portanto, uma metodologia rápida e confiável. Além do mais, uma vez que a busca codifica exatamente o que a ferramenta procura no texto, sabemos precisamente o que foi procurado. Pode acontecer, aliás, que a primeira rodada de uma busca retorne dados não esperados. Nesse caso, a busca não codifica corretamente o que se procura e deve ser corrigida, mas isso pode ser feito simplesmente, modificando, tirando ou acrescentando uma ou várias condições. Uma busca também pode ser estendida a outras categorias, por exemplo, ficando menos restritiva

em relação ao tipo de advérbio. Nesse caso, deve-se só tirar as duas linhas que especificam o ADVP[xix]. Enfim, a mesma busca pode ser usada para replicar a pesquisa num outro corpus, o que assegura uma máxima comparabilidade dos resultados, e uma maior confiabilidade na pesquisa.

Na próxima seção, apresentarei resultados sobre a história da sintaxe portuguesa que foram obtidos graças ao CTB e outros corpora construídos nos mesmos moldes, bem como à metodologia descrita aqui.

4- Resultados

O CTB foi criado para elucidar algumas questões em aberto na história sintática do português europeu (doravante PE), entre o século 16 e o século 19, período pouco estudado pelos trabalhos anteriores, que se concentravam na fase mais antiga da língua, ou seja, até o século 16, considerado como fim do período arcaico. Uma questão em particular dizia respeito à localização no tempo da emergência da sintaxe da língua moderna. A periodização tradicional hesitava entre o século 16 e o século 18 (cf. Mattos e Silva 1994). Ana Maria Martins na sua tese de doutorado (cf. Martins 1994), observa a frequência muito alta de colocação enclítica nos sermões do Pe Antonio Vieira (1607-1696) e sugere que esse autor já é representativo da língua moderna. Desse ponto de vista, a sintaxe moderna já estaria aparecendo nos textos do século 17. O CTB permitiu testar empiricamente essa hipótese, com base num conjunto de textos e autores nunca considerado até então.

A figura 7, baseada em 16 textos sintaticamente anotados de autores portugueses nascidos entre 1502 e 1836[xx], mostra a evolução da colocação enclítica em contextos de variação ênclise/próclise, em frases de ordem Sujeito-Verbo, como exemplificado em (8), onde a. e b. são casos de ênclise e c. e d. casos de próclise.

(8) a. Elles **conheciam-se**, como homens, (Vieira, n. 1608)

b. Christo **conhecia-os**, como Deus. (Vieira, n. 1608)

c. Ele **me disse** que pasmava como lhe abastava o que tinha (Sousa, n.1554)

d. Ruy Lopes de Villa-Lobos **o recebeo** com muita honra (Couto, n.1542)

Na figura 7, cada losango preto representa a frequência da colocação enclítica num texto situado, no eixo do tempo, na data do seu nascimento do seu autor. Vieira, nascido em 1608, aparece em dois pontos, correspondendo a dois textos distintos: os sermões, que têm 52% de ênclise, e as cartas, que não têm nenhuma ocorrência de ênclise. Note-se que o valor dos sermões (0.52) destoa em relação a todos os outros textos da mesma época, nos quais a ênclise é marginal, não passando de 18%, inclusive com dois textos sem nenhuma ocorrência, como nas cartas do próprio Vieira. Vê-se também que é nos autores nascidos depois de 1700 que se vê o início da curva de mudança no sentido de cada vez mais ênclise, chegando a 97% no último autor, Ramalho Ortigão, nascido em 1836, beirando os 100% da língua moderna.

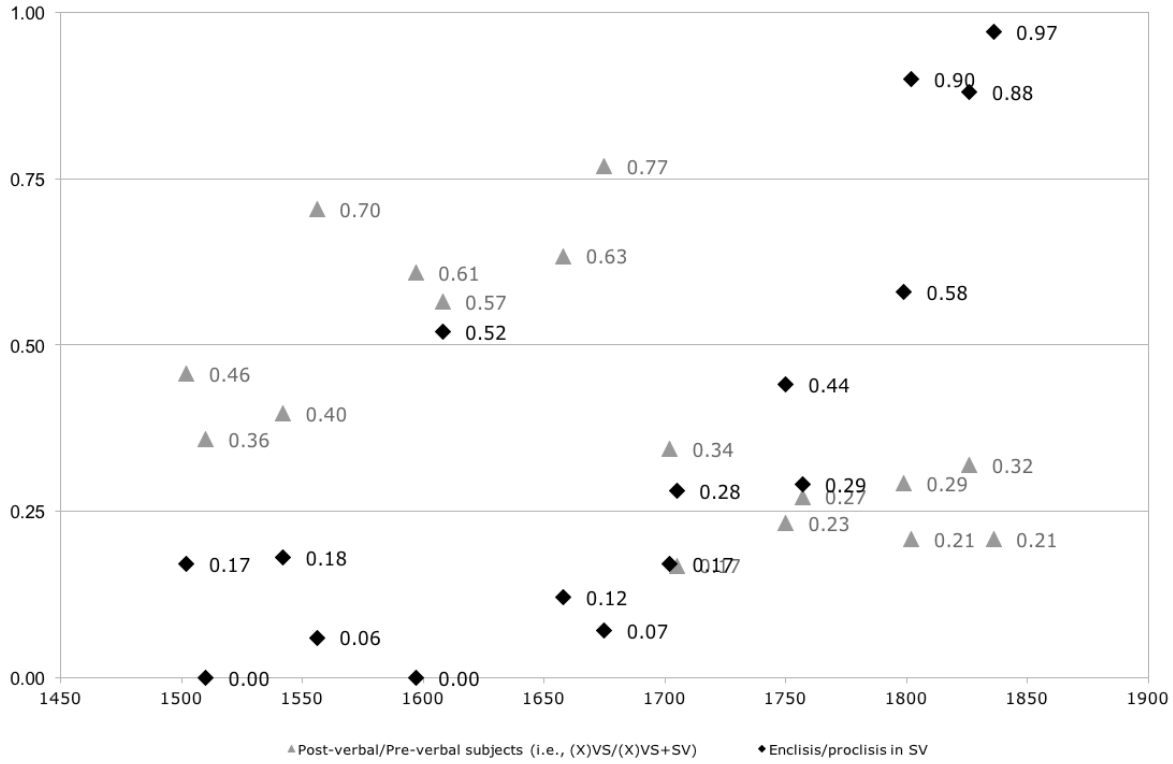

Figura7- a evolução da colocação de clíticos e da posposição do sujeito no português europeu (século 16-19).
Fonte: Galves e Paixão de Sousa (2017)

O que foi então possível mostrar graças a um corpus muito mais extenso do que aquele que tinha sido usado por Martins (1994) para esse período, é que os sermões do Pe Vieira são uma exceção, um ponto fora da curva[xxi]. A conclusão de que a mudança se dá mais tarde, já no século 18, é corroborada pela evolução da sintaxe do sujeito, que mostra uma queda brusca da

frequência da ordem Verbo-Sujeito (VS), em relação a SV, justamente quando a frequência da ênclise aumenta. Isso pode ser observado na mesma Figura 7, onde os triângulos cinzentos correspondem à percentagem de ordem VS. Vê-se claramente que a queda na frequência dessa ordem se dá nos autores nos quais a percentagem de ênclise aumenta.

O mesmo fato pode ser observado na distribuição dos sujeitos nulos, sujeitos pospostos (VS) e sujeitos antepostos (SV), por século, mostrada na Figura 8,onde se vê que a frequência média de sujeitos pospostos (VS) cai de 0.35 no século 17 para 0.12 no século 18, enquanto ao mesmo tempo a frequência de SV passa de 0.17 no século 17 para 0.41 no século 18. É interessante notar que, ao mesmo tempo, a frequência de sujeitos nulos se mantém constante.

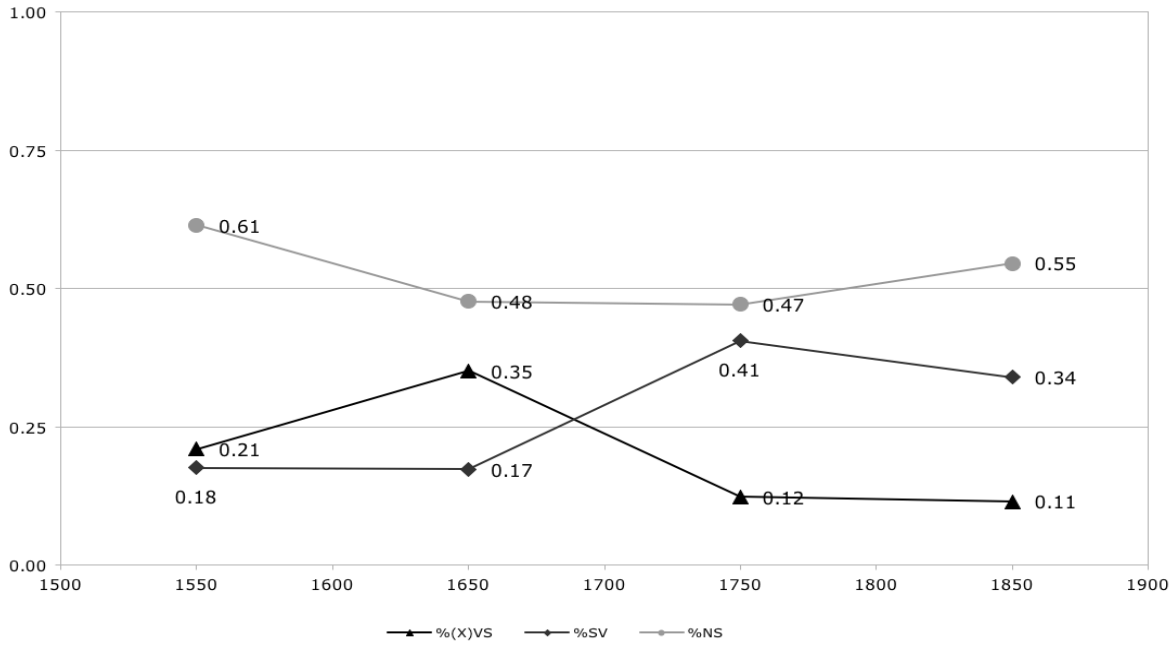

Figura 8- Realização do sujeito (SV/VS, sujeito nulo).
Fonte: Galves e Paixão de Sousa (2017)

A frequência alta de VS no português dos séculos 16 e 17 está associada a uma outra propriedade, da qual já se falou na seção anterior, a propriedade V2, ou seja a recorrência de construções em que o verbo está em segunda posição, mas o que o precede não é o sujeito, como ilustrado nas frases em (9)[xxii]:

(9) a. *Quanto o demonio trabalhou em dous annos* **desfes** Deos em hum instante;

(Maria do Céu,n. 1658)

b. E nos gasalhados e abraços **mostraram** os cardeais legados bem este contentamento; (Sousa, n. 1556)

A figura 9, baseada em 11 textos sintaticamente anotados do CTB, mostra como no século 18 também, a frequência de construções como 9a. e b., chamadas XV (em verde) por oposição a SV (em laranja), cai drasticamente[xxiii].

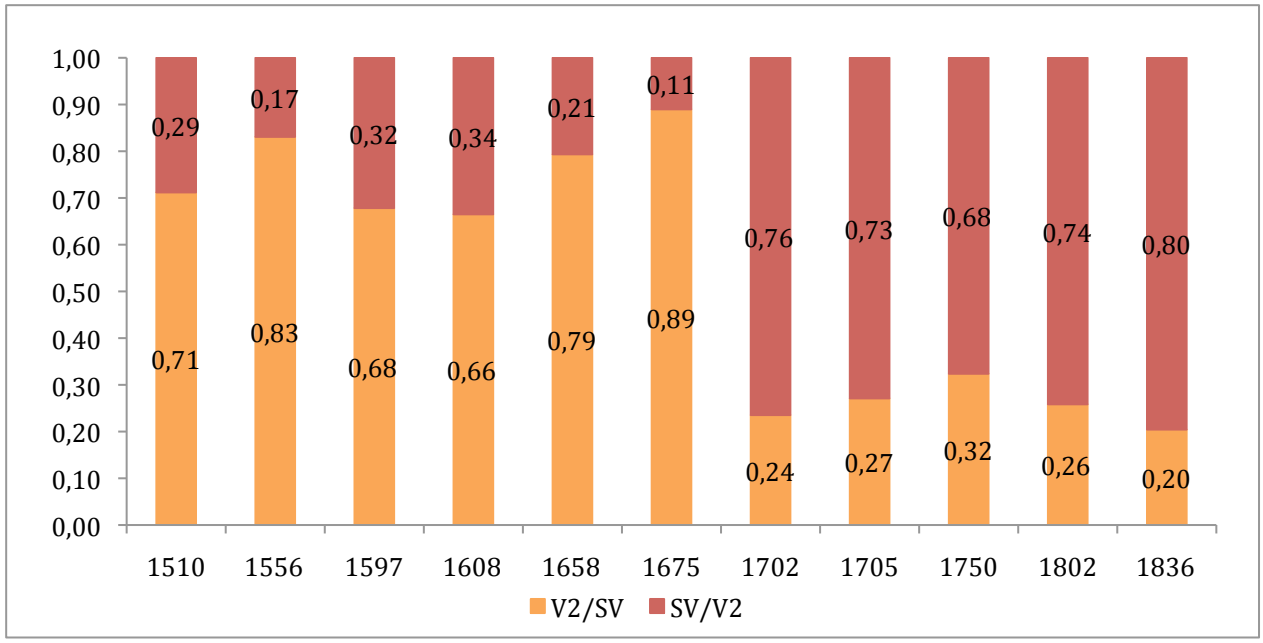

Figura 9- Frequência relativa de SV e XV, por autor
Fonte: Cavalcante; Galves; Paixão de Sousa (2010)

Além de trazer evidência empírica muito forte de que a sintaxe do português europeu moderno aparece nos textos dos autores nascidos depois de 1700, respondendo à pergunta levantada no início da seção, o CTB também permite descrever o estado de língua que chamamos de português clássico, ou seja a língua escrita, e por hipótese falada, pelos autores nascidos nos séculos 16 e 17. Ou seja, além de permitir um trabalho diacrônico sobre a língua, ele também nos dá elementos para escrever gramáticas sincrônicas de fases passadas. Com o crescimento do corpus anotado, foi possível realizar um trabalho de descrição detalhado da distribuição dos padrões sentenciais com base em 11 textos de autores nascidos nos séculos 16 e 17[xxiv], trabalho nunca antes empreendido. A tabela a seguir mostra os resultados dessa

pesquisa no que diz respeito à posição do verbo e a realização do sujeito nas orações principais.

	VS	SV	Sujeitos nulos	TOTAL	
V1	1477	/	936	2413	28%
V2	1417	1388	2356	5161	59%
V3	361	524	256	1141	13%
TOTAL	3255	1912	3548	8715	100%

Tabela 1: Posição do verbo e realização do sujeito no português clássico
Fonte: Galves (a sair)

A distribuição das orações com verbo em primeira, secunda e terceira posição tem sido a fonte de grande debate sobre a sintaxe das línguas românicas antigas, e em particular sua possível natureza V2 (cf. nota xvi). O trabalho sobre o português clássico é particularmente relevante já que se trata da língua românica que instanciou o fenômeno até um período mais tardio, já numa época em que os textos são numerosos e muito diversificados, o que não é sempre o caso com as línguas mais antigas, em que só alguns géneros são disponíveis. Observe-se que foram analisadas 8715 orações, distribuídas em 11 textos representativos dos géneros historiografia, biografia, obras filosóficas e religiosas, cartas, gazeta. O CTB permitiu assim uma janela ímpar sobre o fenômeno, trazendo com muito fundamento empírico o português para dentro da discussão sobre a história das línguas românicas.

Além disso, como foi mencionado na Seção I, outros corpora do português receberam a mesma anotação. Torna-se então possível fazer as mesmas perguntas aos textos do corpus *Post-Scriptum*, ou aos documentos medievais do corpus *Wochwel* (cf. nota ix). O primeiro reúne cartas pessoais escritas do século 16 ao século 19, por hipótese mais próximas da língua falada do que os textos disponíveis no CTB, de cunho marcadamente literários. Isso permite verificar até que ponto o eixo [+ literário/- literário] tem efeito sobre a sintaxe da época, ou seja, até que ponto os resultados apresentados na Tabela 1 são reproduzidos nas cartas do *Post-Scriptum*. Estudos preliminares mostram que o perfil sintático destas, sem ser

drasticamente diferente, apresenta aspectos distintos, em particular uma frequência mais alta da ordem SV, o que pode prenunciar o enfraquecimento do fenômeno V2.

Os textos sintaticamente anotados do português antigo, por sua vez, foram cotejados com os textos do CTB por Medeiros (2018), que observou uma diferença importante na realização do fenômeno V2 no período arcaico e no período clássico. O paralelismo entre as orações principais e as orações subordinadas, existente no primeiro, deixa de existir no segundo. Em outros termos, o português arcaico é uma língua V2 simétrica e o português clássico uma língua V2 assimétrica. Essa conclusão, importante para a compreensão da dinâmica diacrônica da língua portuguesa, não poderia ser obtida sem o recurso aos corpora anotados, que possibilitam uma comparação de fenômenos com base em dados mais numerosos, mais processáveis e mais confiáveis.

5- Considerações finais

Neste breve artigo, mostrei a importância dos corpora sintaticamente anotados para a pesquisa em sintaxe histórica. Graças ao trabalho coletivo de vários grupos, dispomos agora de uma constelação de corpora do português que nos permite entender muito melhor a dinâmica da língua ao longo de sua história. Os textos, bem como as ferramentas de construção e de exploração destes corpora estão disponíveis para todos os pesquisadores envolvidos com a história da língua. O caminho descrito nas seções 1 e 2 está sendo agora trilhado por vários estudiosos do português do Brasil. Graças à metodologia unificada de anotação e busca que garante uma máxima comparabilidade entre os diversos corpora, podemos vislumbrar uma história comparada do português europeu e do português brasileiro, o que certamente nos permitirá entender melhor a dinâmica de sua separação, determinando em particular uma cronologia dos fenômenos de mudança que afetaram o português no Brasil, ou seja uma periodização baseada em fatos linguísticos e não mais somente em fatos sócio-históricos[xxv]. O cruzamento dessas duas periodizações será de grande valia para entender os caminhos da língua portuguesa em território brasileiro. O sucesso dessa empresa depende da integração nos corpora de documentos oriundos de diversas regiões e de diversas origens sociais, chegando aos mais populares. A convergência entre as ferramentas computacionais e conceituais apresentadas aqui e o minucioso trabalho de levantamento, edição e análise sócio-histórica de documentos não literários efetuado no âmbito de projetos de pesquisa como o

CE-DOHS[xxvi] e de outros grupos associados ao *Projeto para a História do Português Brasileiro*, será essencial para o êxito dessa jornada no túnel do tempo.

Para terminar, vale mencionar que está atualmente em desenvolvimento uma nova versão do CTB, que permitirá acessar a todas as suas funcionalidades on-line, com interfaces facilitadoras das tarefas de anotação e busca. Esse projeto envolve a construção de uma plataforma aberta à criação de novos corpora, com o auxílio das mesmas ferramentas, parametrizáveis para as diferentes línguas. Já está em estágio de desenvolvimento um corpus da língua indígena kadiweu, nos mesmos moldes, com as devidas adaptações, tanto na anotação linguística, quanto no tipo de documentos disponibilizados[xxvii].

6- Referências

ANTONELLI, A. **Sintaxe de posição do verbo e mudança gramatical na história do português europeu**. 2011. Tese (Doutorado em Linguística), Instituto de Estudos da Linguagem, Universidade de Campinas, Campinas

BELLETTI, A. Aspects of the low IP area. In: RIZZI, L. (org.) **The Structure of CP and IP**, Oxford: Oxford University Press, 2004. p. 16-51.

CAVALCANTE, S.; GALVES, C.; PAIXÃO DE SOUSA, M.C. Topics, subjects and grammatical change: from Classical to Modern European Portuguese, In: **Subjects in diachrony: Grammatical change and the expression of subjects Conference**, Universidade de Regensburg, Alemanhã, 4-5/11/2010. http://rhssl1.uni-regensburg.de/SlavKo/conferences/gces/abstracts/Galves_Paix2206o%20de%20Sousa_de%20Oliveira%20Cavalcante.pdf

FARIA, P.; GALVES, C. Criando "bancos de árvores": o sistema de anotação e o processamento automático. **Cadernos de Estudos Linguísticos**, Campinas, v. 58, n.2, p. 299-315, 2016.

GALVES, C. Syntax and Style: clitic-placement in Padre Antonio Vieira. **Santa Barbara Portuguese Studies,** Santa Barbara**,** vol.6, p. 387-403, 2002.

GALVES, C. Relaxed Verb Second in Portuguese. In: WOLFE, S.; WOODS, R. (orgs.) **Rethinking Verb Second**, Oxford: Oxford University Press, a sair.

GALVES, C.; PAIXÃO DE SOUSA, M.C. The position of the verb in the history of Portuguese: Subject position, Clitic placement and Prosody, **Language**, vol, 93, n.3, p.152-180, 2017.

LUCCHESI, D. A periodização da história sociolinguística do Brasil. **DELTA-Documentação em Linguística Teórica e Aplicada**, v. 33, n.2**,** 2017.

MAGRO, C. ParsPort. Revision queries for parsing Portuguese. Centro de Linguística da Universidade de Lisboa, 2017. URL: http://parsport.sourceforge.net

MARTINS, A.M. **A história dos clíticos em português**. 1994. Tese (Doutorado em Linguística), Faculdade de Letras, Universidade de Lisboa, Lisboa.

MATTOS E SILVA, R.V. Para uma caracterização do período arcaico do português, **D.E.L.T.A-Documentação em Linguística Teórica e Aplicada**, São Paulo, vol.10 n. especial, p.247-276, 1994.

MEDEIROS, C. **A sintaxe da ordem no português medieval**. 2018. Tese (Doutorado em Linguística), Instituto de Estudos da Linguagem, Universidade de Campinas, Campinas.

OLIVEIRA, K. **Negros e escrita no Brasil do século XIX: sócio-história, edição filológica de documentos e estudo linguístico**. 2006. Tese (Doutorado), Universidade Federal da Bahia, Salvador.

PAIXÃO DE SOUSA, M.C. O Corpus Tycho Brahe: contribuições para as humanidades digitais no Brasil, **Filologia e Linguistica Portuguesa**, v. 16 n. esp, p. 53-93, 2014.

PAIXÃO DE SOUSA, M.C.; KEPLER, F.; FARIA, P. E-Dictor: Novas perspectivas na codificação e edição de corpora de textos históricos. In: Shepherd, T.; Berber Sardinha, T.; Veirano Pinto, M. (Orgs.). **Caminhos da linguística de corpus**, Campinas: Mercado de Letras, 2010.

Sobre a autora

Charlotte Galves. Doutora em Língua Portuguesa pelo Université Paris-Sorbonne, França (1980), Professor titular da Universidade Estadual de Campinas, Bolsista de Produtividade em Pesquisa CNPq – Nível 1, tem experiência na área de Linguística, com ênfase em Descrição e Análise do Português, atuando principalmente nos seguintes temas: descrição comparativa do português europeu, português brasileiro e português clássico no quadro teórico da gramática gerativa; história gramatical da língua portuguesa nas suas diversas vertentes; interface fonologia-sintaxe e seu papel na mudança linguística; elaboração e uso de grandes corpora eletrônicos anotados de língua; além de modelagem probabilística em linguística. Charlotte foi pioneira em investir na formação de grandes corpora anotados em Língua Portuguesa, precursora do primeiro corpus sintaticamente anotado do Português: O Corpus Histórico do Português Tycho Brahe.

Notas

[i] Ver o manual de anotação morfológica em
http://www.tycho.iel.unicamp.br/~tycho/corpus/manual/pos2016.html

[ii] Esse é o único texto brasileiro presente, nesta data, no CTB, com anotação sintática. Trata-se de atas da *Sociedade Protetora dos Desvalidos*, escritas na Bahia no século 19, por negros alforriados. Cf. Oliveira (2006)

[iii] Note-se, porém, que isso não é aplicado de maneira sistemática. Por exemplo, não há projeção de VP. Remeto o leitor interessado para a introdução do Manual de Anotação Sintática para uma discussão (cf. https://sites.google.com/site/portuguesesyntacticannotation/1/1-1-introductory-remarks

[iv] Os exemplos são retirados do manual de anotação sintática, cf. nota viii.

[v] Note-se que os vestígios são sistematicamente inseridos no início do IP.

[vi] Por exemplo, não há marcação do sintagma verbal, o que tem como consequência o sujeito e os complementos estarem ambos no mesmo nível do verbo (cf. nota iii).

[vii] Cf. https://www.ling.upenn.edu/~kroch/
Note-se que esse mesmo sistema está sendo usado para várias outras línguas (cf. https://www.ling.upenn.edu/hist-corpora/)

[viii] Cf. https://sites.google.com/site/portuguesesyntacticannotation/

[ix] Além do CTB, são o corpus *Wochwel* http://alfclul.clul.ul.pt/wochwel/, o corpus *Post Scriptum*: http://ps.clul.ul.pt/pt/index.php e o corpus *Cordial-Sin*: http://www.clul.ulisboa.pt/en/10-research/314-cordial-sin-corpus .

[x] Cf. Faria & Galves (2016).

[xi] Cf. corpussearch.sourceforge.net

[xii] Cf. http:// annotald.github.com/user.html

[xiii] Cf. Magro (2017). A primeira ferramenta desse tipo foi elaborada para o francês por Beatrice Santorini na Universidade da Pensilvânia.

[xiv] Está disponível no site do CTB uma interface gráfica que permite fazer buscas em textos etiquetados (cf. http://www.tycho.iel.unicamp.br/~tycho/corpus/texts/csquery/csquery.html). Apesar da sua facilidade de uso, essa ferramenta tem limitações que são problemáticas quando se quer trabalhar com a sintaxe histórica. Com efeito, uma vez que os textos não têm anotação sintática, as informações contidas na busca só podem dizer respeito à ordem e às categorias e/ou às palavras presentes nos textos. Noções como "sujeito", "objeto", orações "matriz" e "subordinada", bem como objetos complexos como sintagmas ou categorias vazias não podem ser referidas nas buscas. Para isso, necessita-se de uma anotação completa. Além do mais, uma só propriedade afetando um termo ou uma sequência de termos pode ser definida em cada busca. Essa ferramenta é, contudo, útil para pesquisas exploratórias, para certos tipos de fenômenos de ordem, e além disso, como um primeiro contato com a lógica de *Corpus Search*.

[xv] Cf. Belletti (2004).

[xvi] O termo "língua V2" foi atribuído inicialmente às línguas nas quais o verbo aparece sempre em segunda posição nas orações independentes, como o alemão e várias outras línguas germânicas. Nessas línguas, o sujeito é pós-verbal sempre que um outro elemento da oração se encontra na posição pré-verbal. Para alguns autores, a caracterização de uma língua como V2 deriva mais de propriedades estruturais do que da ordem linear. Para eles, a propriedade essencial das línguas V2 é o movimento do verbo para uma posição da periferia esquerda da oração. Desse ponto de vista, línguas em que o verbo não está sempre linearmente na segunda posição podem ser consideradas línguas V2 se for mostrado que elas compartilham essa propriedade estrutural.

[xvii] Cf. por exemplo Galves & Paixão de Sousa (2017)

[xviii] t significa "true" ("verdadeiro").

[xix] Para uma apresentação completa da sintaxe de *Corpus Search*, veja-se o manual do usuário em http://corpussearch.sourceforge.net/CS-manual/Contents.html

[xx] Lista dos textos compondo essa versão:
- século 16: Pero Magalhães Gândavo (n.1502, g_008) "Historia da Província de Santa Cruz vulgarmente chamada Brasil"; Fernão Mendes Pinto (n.1510, p_001), "Peregrinação"; Diogo do Couto (n.1542, c_007) "Décadas"; Frei Luís de Sousa (n.1556, s_002), "A vida de Frei Bertolameu dos Mártires";
- século 17: Manuel Galhegos (n.1598, g_001), "Gazeta"; Pe Antonio Vieira (n.1608), "Sermões" (v_004); Maria do Céu (n.1658, c_002), "Vida e morte de Madre Elena da Cruz"; André de Barros (n.1675, b_001), "Vida do apostólico Pe Antonio Vieira";

- século 18: Cavaleiro de Oliveira (n. 1702, c_001), “Cartas”; Matias Aires (n. 1705, a_001), “Reflexões sobre a vaidade dos homens”; Marquesa de Alorna (n. 1750, a_004), “Cartas”; J. D. Rodrigues da Costa (n. 1757, c_005), “Entremezes de cordel”;
- século 19: Almeida Garrett (n. 1799, g_004), “Teatro”; Marquês de Fronteira e Alorna (n. 1802, a_003), “Memórias”; Camilo Castelo Branco (n. 1825, b_005), “Maria Moisés”; Ramalho Ortigão (n. 1836, 0_001), “Cartas a Emília”.

[xxi] Para uma explicação desse fenômeno, remeto o leitor a Galves (2002).

[xxii] Note-se a presença do advérbio de modo “bem” depois do sujeito pós-verbal em 9b. Para uma discussão da ordenação dos advérbios e da sintaxe V2 do português clássico, ver também Antonelli (2011).

[xxiii] Outros estudos baseados no CTB mostraram o mesmo resultado com outros fenômenos. Por falta de espaço, não me deterei neles aqui.

[xxiv] Àqueles mencionados na nota xx para os séculos 16 e 17, foram adicionados os seguintes textos: Francisco Rodrigues Lobo (n.1579, l_001) "Côrte na Aldeia e Noites de Inverno "; Pe Antonio Vieira (n.1608, v_002) “Cartas”; José da Cunha Brochado (n.1651, b_008) “Cartas".

[xxv] Para uma revisão das periodizações propostas e uma nova visão da questão, ver Lucchesi (2017).

[xxvi] Cf. http://www5.uefs.br/cedohs/

[xxvii] O corpus kadiweu inclui, por exemplo, as gravações sonoras das sentenças Cf. http://www.tycho.iel.unicamp.br:8180/fmk/index