Filol. Linguíst. Port., São Paulo, v. 16, n. spe, p. 53-93, dez. 2014
http://dx.doi.org/10.11606/issn.2176-9419.v16ispep53-93

O *Corpus Tycho Brahe*: contribuições para as humanidades digitais no Brasil

The Tycho Brahe Corpus: contributions for the digital humanities in Brazil

Maria Clara Paixão de Sousa *
Universidade de São Paulo, São Paulo, São Paulo, Brasil

Resumo: O processo de aproximação entre o campo filológico e o campo computacional nos estudos históricos da língua portuguesa, observado desde os anos 1990, configura hoje um horizonte em franca expansão, tornando oportunas as reflexões sobre as transformações produzidas pelo o tratamento computacional na tradição do trabalho filológico e linguístico. Este artigo se propõe a uma reflexão nesse sentido, partindo da exploração detalhada da tecnologia de codificação de textos usada no *Corpus Anotado do Português Histórico Tycho Brahe*, buscando examinar as diferentes combinações de procedimentos filológicos, linguísticos e computacionais envolvidos em sua construção, e discutindo as implicações metodológicas desses procedimentos. Abordamos os corpora eletrônico anotados não como coleções 'de' textos, mas sim como bancos de dados 'sobre' textos, que englobam diferentes camadas de representação sobre sua linguagem e sobre sua materialidade. Essa abordagem nos permite vislumbrar as especificidades do trabalho em ambiente digital no campo da filologia e da linguística histórica, e sugerir alguns caminhos para o debate sobre os desafios e perspectivas que se abrem para esse campo a partir dos projetos pioneiros que descrevemos.

* Professora do Departamento de Letras Clássicas e Vernáculas da Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo – USP, São Paulo, São Paulo, Brasil; mariaclara@usp.br.

ISSN 1517-4530

Palavras-chave: Humanidades digitais. Filologia. Linguística histórica. Corpus eletrônico anotado. Anotação sintática.

Abstract: The confluence between philology and computation in Portuguese historical linguistics, in process since the 1990s, is now an expanding horizon, making this a good moment for the reflection about the transformations produced by computational technologies on traditional philological and linguistic work . This paper proposes to reflect on this, starting with a detailed exploration of the text technologies used in the *Tycho Brahe Corpus of Historical Portuguese*, and examining the combinations of philological, linguistic and computational procedures involved in its construction. We approach electronic corpora as databases about texts, rather than collections of texts; and this allows us to envisage the particularities involved in work with digital media in philology and linguistics. We try to start debate on the challenges and perspectives opened to these fields in Brazil, after the considerations on the pioneer projects described.

Keywords: Digital humanities. Philology. Historical linguistics. Annotated corpora. Syntactic annotation.

1 INTRODUÇÃO[1]

A pesquisa histórica sobre a língua no Brasil alcançou importância central no final do século XX, com a intensificação do interesse pela perspectiva diacrônica e a renovação da relevância dos estudos da mudança linguística em diferentes quadros teóricos (entre outros, com Mattos e Silva 1988, Kato & Roberts 1996, Castilho 1998, Galves 1998) – processo que trouxe, como consequência, o adensamento do trabalho com textos antigos (conforme observaram Megale & Cambraia, 1999), marcando um período de crescimento do interesse pelas áreas da filologia e da crítica textual no país. Também na transição entre o final da década

1 Este trabalho é resultado de uma experiência de quinze anos de colaboração com a Profª. Drª. Charlotte Galves e com a equipe que construiu o *Corpus Tycho Brahe*. Desde minha iniciação científica (1996-1998), passando por meu trabalho de tese (1999-2004), até um estágio de pós doutoramento (2004-2007) – todos sob orientação da Profª. Charlotte e financiados pela Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) – pude acompanhar o processo de desenvolvimento do *Corpus* e as pesquisas feitas com base nele por colegas de iniciação, pós-graduação e pós-doutorado. O artigo tenta mostrar um pouco desse processo coletivo de trabalho, seus resultados e perspectivas.

de 1990 e o início dos anos 2000, começam a ampliar-se as pesquisas nas áreas de linguística de corpus e da linguística computacional, com a construção dos primeiros grandes corpora anotados da língua portuguesa, até hoje ativos (como mostram Shepherd et al., 2010). Na confluência entre a retomada da perspectiva diacrônica e a ampliação do trabalho computacional, surgem, na mesma época, os primeiros corpora anotados de textos históricos portugueses – no Brasil, o *Corpus Anotado do Português Histórico Tycho Brahe* (a partir de Galves, 1998); em Portugal, o *Corpus Electrónico Português do Período Clássico* (a partir de Verdelho, 2001) e o *Corpus Informatizado do Português Medieval* (a partir de Xavier, 2002), para ficarmos nos exemplos emblemáticos. Mais de uma década depois, ao pioneirismo dessas primeiras iniciativas com o trabalho computacional sobre o texto antigo em português se juntaram diversos novos projetos, cenário discutido detalhadamente em Banza e Gonçalves (2013). Entretanto, essa união entre os estudos históricos da língua, o trabalho filológico e o tratamento computacional do texto traz desafios inéditos, e costuma ser acomapanhada de intensos debates – de fato, para alguns estudiosos contemporâneos (entre outros, Baumann & Crane 2010), o trabalho com o texto antigo no ambiente digital faz surgir uma nova filologia, a *e-philology* (Crane et al., 2008); para outros, as rupturas são de tal forma profundas que chegam a determinar o nascimento de um campo inteiramente novo de investigação intelectual: as *Humanidades Digitais* (Schreibman et al., 2004, entre outros). Como quer que nos posicionemos frente a esse debate, um ponto se destaca: os impactos das tecnologias digitais de difusão de textos sobre os campos da filologia e dos estudos históricos da língua tendem a aumentar nas próximas décadas. Como bem observa Crane (2010), as iniciativas de digitalização de acervos históricos no mundo inteiro ainda estão em crescimento[2], e isso pode tornar a relevância do trabalho filológico no meio digital cada vez mais clara para a comunidade acadêmica nos próximos anos. No ambiente da língua portuguesa, o crescimento das iniciativas de digitalização foi também sensível na última década[3], e acreditamos, com Crane (2010), que o adensamento do volume de documentos históricos disponíveis em formato digital tende a levar a um crescimento da percepção da necessidade de transformar o material digitalizado

2 Notemos por exemplo que, mesmo com os inúmeros projetos de digitalização realizados nos últimos anos, calcula-se que hoje menos de 10% dos documentos em arquivos europeus já tenham sido digitalizados (cf. Pekel, 2012).

3 Para ficar apenas entre as iniciativas já consolidadas, lembramos a digitalização dos acervos do *Arquivo Nacional da Torre do Tombo* (http://antt.dglab.gov.pt/) e da *Biblioteca Nacional* (http://bnd.bn.pt/) em Portugal, e do *Arquivo do Estado de São Paulo* (http://www.arquivoestado.sp.gov.br/acervo_digitalizado.php) e da *Biblioteca Nacional* (http://bndigital.bn.br/) no Brasil.

em textos *efetivamente trabalháveis* – legíveis pelas pessoas, e processáveis pelos computadores. Assim, aquele processo de aproximação entre o campo filológico e o campo computacional nos estudos históricos da língua portuguesa, cujos primórdios observamos desde o final do último século, aparece, no cenário atual, como horizonte em franca expansão. O momento, portanto, parece mais que maduro para a reflexão sobre as rupturas (e as continuidades) produzidas pelo o tratamento computacional de documentos históricos na tradição do trabalho filológico e linguístico.

Este artigo se propõe a essa reflexão, partindo de uma exploração detalhada da tecnologia de codificação de textos usada no *Corpus Anotado do Português Histórico Tycho Brahe* (Galves e Faria, 2010), buscando examinar as diferentes combinações de procedimentos filológicos, linguísticos e computacionais envolvidos em sua construção, e discutindo as implicações metodológicas desses procedimentos. Essa proposta envolve o desafio de abordar com alguma propriedade técnicas originárias de campos do conhecimento até pouco tempo muito distantes entre si; para empreendê-la, optaremos por abordar os procedimentos computacionais de modo panorâmico, usando uma linguagem corrente, dirigindo a discussão mais ao leitor afeito à filolologia e à linguística histórica que ao público da computação. Buscamos, assim, contribuir um pouco para o diálogo e para a aproximação entre essas esferas, que devem se intensificar nos próximos anos.

2 A ANOTAÇÃO FILOLÓGICA E LINGUÍSTICA NO *CORPUS TYCHO BRAHE*

2.1 Panorama geral

O *Corpus Anotado do Português Histórico Tycho Brahe*, construído a partir do projeto de Galves (1998), foi uma iniciativa pioneira no âmbito da língua portuguesa, e permanece hoje como o maior corpus eletrônico anotado de textos históricos em português. O *Corpus* compreende atualmente cerca de sessenta obras, disponíveis para pesquisa livre em diferentes formatos: versões para leitura, com transcrição conservadora e modernizada, e versões com anotação linguística – morfossintática e sintática. A seleção dos documentos começou seguindo os objetivos de Galves (1998), reunindo um conjunto de textos portugueses escritos por autores nascidos entre os séculos XVI e XIX como subsídio para estudar mudanças gramaticais sofridas pelo Português Europeu naquele período; mas, com o desenvolvimento de projetos subsequentes, o horizonte do corpus se expandiu no tempo e no espaço, e a coleção passou a incluir obras de autores brasileiros e africanos, assim como obras representativas de fases mais recuadas da

língua. Hoje, a coleção inclui textos escritos por autores portugueses, brasileiros e africanos, nascidos entre 1380 e 1845. Nesta seção, apresentamos o processo que se desenrola entre os textos-fonte e o produto final das anotações no *Corpus*, começando por um panorama amplo, e passando em seguida à descrição de cada etapa de anotação. O conjunto de textos anotados que forma o *Corpus Tycho Brahe* (doravante CTB) tem como objetivo principal possibilitar de forma ampla a *recuperação de informações* filológicas e linguísticas dos textos. Para apresentar o processo de anotação realizado com essa finalidade, tomaremos como exemplo um trecho da obra "*História da Província de Santa Cruz a que vulgarmente chamamos Brasil*", de Pero Magalhães de Gandavo (1502[?]-1579), impressa em 1576, e editada eletronicamente no CTB em 2004 a partir do fac-simile digital da *Biblioteca Nacional de Portugal* (Gandavo, 1576), descrevendos as três etapas básicas de sua anotação: a *anotação de edição*, a *anotação morfossintática* e a *anotação sintática*, resumidas nas figuras abaixo. A Figura 1 mostra, na página 22 do fac-simile, a sentença "*A frui-/ta della ſe chama banánas: parecenſe na feiçam com pe/pinos, & criamſe em cachos: algũs delles há tam gran-/des que tem de cento & cincoenta banánas pera cima*" (linhas 5 a 8).

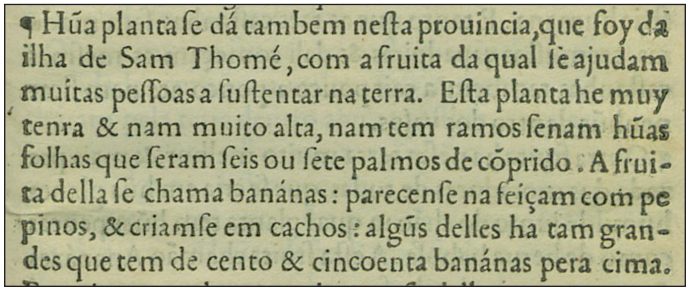
¶ Hũa planta ſe dá tambem neſta prouincia, que foy da
ilha de Sam Thomé, com a fruita da qual ſe ajudam
muitas peſſoas a ſuſtentar na terra. Eſta planta he muy
tenra & nam muito alta, nam tem ramos ſenam hũas
folhas que ſeram ſeis ou ſete palmos de cõprido. A frui-
ta della ſe chama banánas: parecenſe na feiçam com pe
pinos, & criamſe em cachos: algũs delles ha tam gran-
des que tem de cento & cincoenta banánas pera cima.

Figura 1. Detalhe do fac-simile original
(fonte: Gandavo, 1576, p. 22, <http://purl.pt/121/3/#/22>)

As Figuras 2 a 4 mostram as anotações de edição (2), morfossintática (3) e sintática (4) dessa sentença. Destacaremos, no que segue, dois pontos do processo aqui ilustrado: as etapas de anotação são incrementais, cada uma se fundando nos resultados da anterior; e cada etapa envolve procedimentos automatizados e procedimentos de intervenção humana em diferentes medidas.

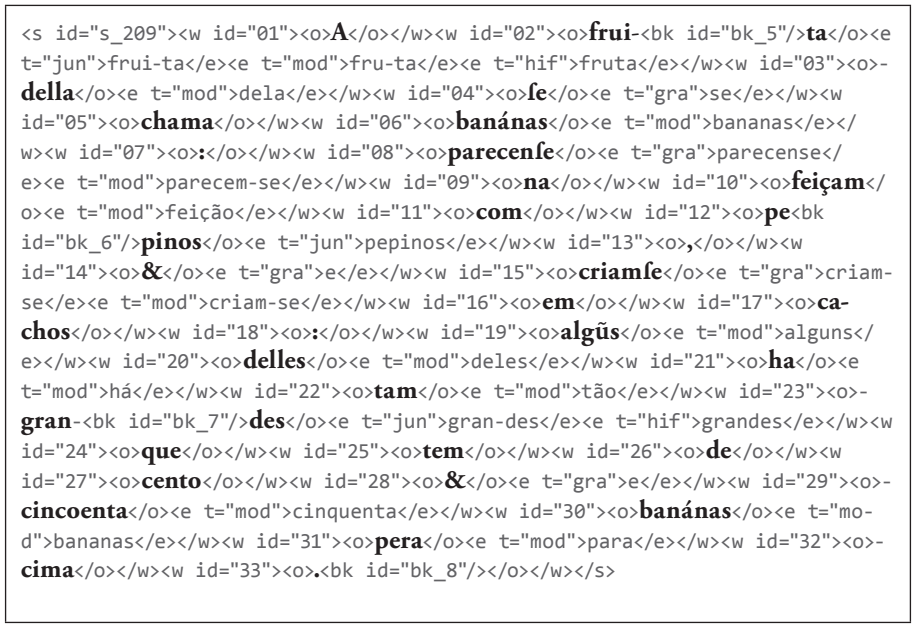
<s id="s_209"><w id="01"><o>**A**</o></w><w id="02"><o>**frui**-<bk id="bk_5"/>**ta**</o><e t="jun">frui-ta</e><e t="mod">fru-ta</e><e t="hif">fruta</e></w><w id="03"><o>-**della**</o><e t="mod">dela</e></w><w id="04"><o>**ſe**</o><e t="gra">se</e></w><w id="05"><o>**chama**</o></w><w id="06"><o>**banánas**</o><e t="mod">bananas</e></w><w id="07"><o>**:**</o></w><w id="08"><o>**parecenſe**</o><e t="gra">parecense</e><e t="mod">parecem-se</e></w><w id="09"><o>**na**</o></w><w id="10"><o>**feiçam**</o><e t="mod">feição</e></w><w id="11"><o>**com**</o></w><w id="12"><o>**pe**<bk id="bk_6"/>**pinos**</o><e t="jun">pepinos</e></w><w id="13"><o>**,**</o></w><w id="14"><o>**&**</o><e t="gra">e</e></w><w id="15"><o>**criamſe**</o><e t="gra">criam-se</e><e t="mod">criam-se</e></w><w id="16"><o>**em**</o></w><w id="17"><o>**ca-chos**</o></w><w id="18"><o>**:**</o></w><w id="19"><o>**algũs**</o><e t="mod">alguns</e></w><w id="20"><o>**delles**</o><e t="mod">deles</e></w><w id="21"><o>**ha**</o><e t="mod">há</e></w><w id="22"><o>**tam**</o><e t="mod">tão</e></w><w id="23"><o>-**gran**-<bk id="bk_7"/>**des**</o><e t="jun">gran-des</e><e t="hif">grandes</e></w><w id="24"><o>**que**</o></w><w id="25"><o>**tem**</o></w><w id="26"><o>**de**</o></w><w id="27"><o>**cento**</o></w><w id="28"><o>**&**</o><e t="gra">e</e></w><w id="29"><o>-**cincoenta**</o><e t="mod">cinquenta</e></w><w id="30"><o>**banánas**</o><e t="mod">bananas</e></w><w id="31"><o>**pera**</o><e t="mod">para</e></w><w id="32"><o>-**cima**</o></w><w id="33"><o>**.**<bk id="bk_8"/></o></w></s>

Figura 2. Anotação de edição da senteça 209

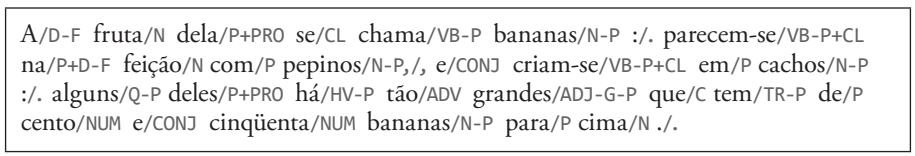
A/D-F fruta/N dela/P+PRO se/CL chama/VB-P bananas/N-P :/. parecem-se/VB-P+CL na/P+D-F feição/N com/P pepinos/N-P,/, e/CONJ criam-se/VB-P+CL em/P cachos/N-P :/. alguns/Q-P deles/P+PRO há/HV-P tão/ADV grandes/ADJ-G-P que/C tem/TR-P de/P cento/NUM e/CONJ cinqüenta/NUM bananas/N-P para/P cima/N ./.

Figura 3. Anotação morfossintática

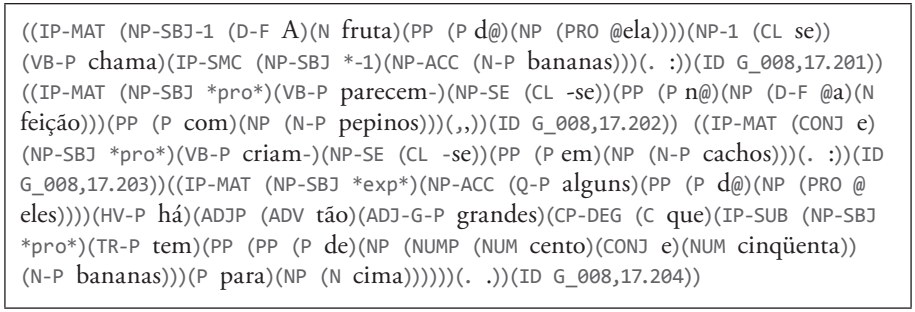
((IP-MAT (NP-SBJ-1 (D-F A)(N fruta)(PP (P d@)(NP (PRO @ela))))(NP-1 (CL se))(VB-P chama)(IP-SMC (NP-SBJ *-1)(NP-ACC (N-P bananas)))(. :))(ID G_008,17.201))((IP-MAT (NP-SBJ *pro*)(VB-P parecem-)(NP-SE (CL -se))(PP (P n@)(NP (D-F @a)(N feição)))(PP (P com)(NP (N-P pepinos)))(,,))(ID G_008,17.202))((IP-MAT (CONJ e)(NP-SBJ *pro*)(VB-P criam-)(NP-SE (CL -se))(PP (P em)(NP (N-P cachos)))(. :))(ID G_008,17.203))((IP-MAT (NP-SBJ *exp*)(NP-ACC (Q-P alguns)(PP (P d@)(NP (PRO @eles))))(HV-P há)(ADJP (ADV tão)(ADJ-G-P grandes)(CP-DEG (C que)(IP-SUB (NP-SBJ *pro*)(TR-P tem)(PP (PP (P de)(NP (NUMP (NUM cento)(CONJ e)(NUM cinqüenta))(N-P bananas)))(P para)(NP (N cima))))))(. .))(ID G_008,17.204))

Figura 4. Anotação sintática

2.2 A anotação de edição

A primeira etapa de anotação do Corpus, a *anotação de edição*, engloba a codificação de informações de duas ordens sobre o texto original: primeiro, as informações relativas às decisões editoriais (conjecturas de leitura, etc.) e à estrutura de texto (quebras de linha, parágrafos seções, etc.); segundo, intervenções interpretativas diversas (atualização grafemática, expansão de abreviaturas, atualização ortográfica, etc.)[4]. Essa etapa é realizada em um processador especialmente voltado para a edição filológica e a codificação linguística eletrônicas, o *eDictor* (Paixão de Sousa, Kepler & Faria, 2013). Concebido originalmente para o CTB, o programa é usado atualmente por outros cinco grupos de pesquisa no Brasil e em Portugal (Carneiro, 2014; Lopes, 2014; Marquilhas, 2014; Namiuti, 2014; Paixão de Sousa, 2014), e pretende oferecer uma interface amigável aliada a um alto nível de controle e flexibilidade na codificação de textos eletrônicos com finalidade de pesquisa linguística. Na versão atual (1.0 Beta 10), o *e-Dictor* combina um editor XML a um etiquetador morfossintático, e inclui funcionalidades para gerar versões dos textos correspondentes aos diferentes planos de anotação.

O módulo Transcrição do *e-Dictor* (Figura 5) oferece uma tela dupla, na qual é possível visualizar o fac-simile de um texto original e transcrevê-lo, aplicando-se informações estruturais básicas (quebras de linha, etc.) e comentários. É possível ainda corrigir a transcrição e anotação estrutural básica de textos transcritos em outros ambientes, manual ou automaticamente1. O módulo Edição (Figura 6) oferece uma interface para se trabalhar o arquivo gerado pelo módulo anterior, editando-o na profundidade desejada. No CTB, essa edição vai até o plano da modernização da grafia. No exemplo em destaque, o termo *parecenſe*, do original, é normalizado grafematicamente para *parecense*, e em seguida tem a grafia atualizada para *parecem-se.*

4 Importa notar que o CTB compreende documentos manuscritos e impressos, alguns deles trabalhados desde a etapa da edição no âmbito do corpus, outros incluídos graças à parceria com projetos e pesquisadores dedicados à edição de manuscritos, como Carneiro, 2014. A anotação eletrônica de edição é aplicada uniformemente seja aos textos editados no âmbito do CTB (como é o caso do texto que usamos como exemplo), seja aos textos cedidos por outros projetos; não é nosso objetivo discutir a metodologia de edição em cada caso, apenas resumir a *anotação* das edições.

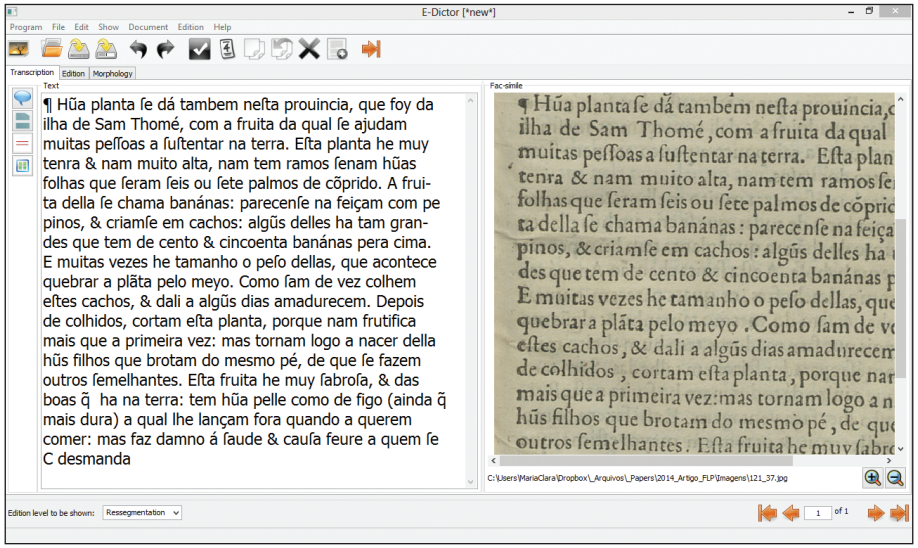

Figura 5. *e-Dictor*: Interface da anotação de transcrição

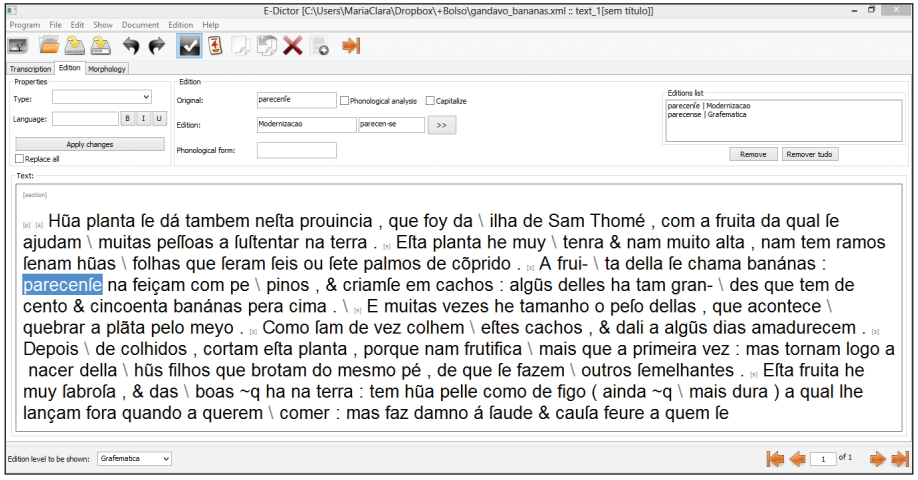

Figura 6. *e-Dictor*: Interface da anotação de edição

A opção pela atualização das grafias nos textos do CTB está ligada à interação entre os requerimentos filológicos fundamentais de fidedignidade em relação aos documentos originais e o atendimento aos requerimentos computacionais das etapas

seguintes de processamento: as ferramentas de análise linguística automática não são capazes, atualmente, de manipular satisfatoriamente textos que apresentem padrões muito amplos de variação nas grafias. Entretanto, este módulo do *e-Dictor* é bastante flexível, pois o inventário de intervenções possíveis é inteiramente aberto, e pode ser adaptado aos objetivos de cada editor ou grupo de editores; além da normalização das grafias, outros grupos de usuários tem aplicado intervenções relativas à correção de reconhecimento automático (Paixão de Sousa, 2011), a análises grafemáticas mais sofisticadas (Silva e Lopes, 2012), entre outros. De todo modo, seja qual for o grau de intervenção editorial aplicado aos textos, é fundamental observar que no sistema de anotação realizado no *e-Dictor* todas as etapas do processamento ficam registradas, de modo que a informação sobre a forma original dos termos nos textos não se perde. De fato, notemos que, se a interface do *e-Dictor*, acima ilustrada, simula um editor de textos normal, a ferramenta é, em essência, um anotador, que aplica uma linguagem de marcação sobre os textos, o XML (eXtended Markup Language, W3 2014). Dessa forma, à edição da sentença "*A frui- ta della ſe chama banánas: parecenſe na feiçam com pe pinos, & criamſe em cachos: algũs delles há tam gran- des que tem de cento & cincoenta banánas pera cima*", do trecho de exemplo, corresponde o código XML ilustrado na Figura 2 mais acima, e abaixo reproduzido em detalhe quanto ao trecho "*parecenſe na feiçam com pe pinos*". Na sintaxe básica da anotação XML, cada "elemento" de anotação fica identificado por rótulos apresentados entre parêntesis angulares, e esses rótulos podem circundar, ou conter, elementos textuais, convencionando-se um rótulo com a forma <nomedoelemento> como abertura, e um rótulo com a forma <*/nomedoelemento*> como fechamento. Dentro do rótulo indicador de abertura de cada elemento, podem-se codificar ainda atributos, com a sintaxe <*nomedoelemento atributo="x"*>. Em (1), por exemplo, cada linha mostra uma palavra, codificada como elemento pelo código <w>...</w>; o atributo *id*="..." mostra a numeração automática de cada palavra; dentro do elemento <w>, o código <o>..</o> indica forma original; <bk/> indica quebras de linha. Essa parte do código é gerada automaticamente pelo *e-Dictor* a partir da transcrição feita no primeiro módulo (Figura 5):

(1)
```
<w id="01"><o>A</o></w>
<w id="02"><o>frui-<bk id="bk_5"/>ta</o></w>
<w id="03"><o>della</o></w>
<w id="04"><o>ſe</o></w>
<w id="05"><o>chama</o></w>
<w id="06"><o>banánas</o></w>
<w id="07"><o>:</o></w>
<w id="08"><o>parecenſe</o></w>
<w id="09"><o>na</o></w>
```

```
<w id="10"><o>feiçam</o></w>
<w id="11"><o>com</o></w>
<w id="12"><o>pe<bk id="bk_6"/></o></w>
<w id="12"><o>pinos</o></w>
```

Por cima da transcrição assim codificada, uma segunda camada de anotação pode ser aplicada no interior de cada elemento <w>...</w>, correspondendo às intervenções aplicadas pelos editores, no segundo módulo (Figura 6). Cada intervenção é codificada entre <e>...<e/>, e recebe um atributo no formato *t*="...", indicando com a natureza da intervenção (sendo esse o passo que pode ser adaptado a diferentes objetivos de edição). No nosso exemplo do CTB, as intervenções incluiram a atualização grafemática (sigla t="*gra*"), a junção de termos (t="*jun*"), a modernização ortográfica (t="*mod*"), e a retirada de hífens depois de junções (t="*hif*"):

(2)
```
<w id="01"><o>A</o></w>
<w id="02"><o>frui-<bk id="bk_5"/>ta</o>
          <e t="jun">frui-ta</e><e t="mod">fru-ta</e><e t="hif">fruta</e></w>
<w id="03"><o>della</o><e t="mod">dela</e></w>
<w id="04"><o>ſe</o><e t="gra">se</e></w>
<w id="05"><o>chama</o></w>
<w id="06"><o>banánas</o><e t="mod">bananas</e></w>
<w id="07"><o>:</o></w>
<w id="08"><o>parecenſe</o><et="gra">parecense</e><e t="mod">parecem-se </e></w>
<w id="09"><o>na</o></w>
<w id="10"><o>feiçam</o><e t="mod">feição</e></w>
<w id="11"><o>com</o></w>
<w id="12"><o>pe<bk id="bk_6"/>pinos</o><e t="jun">pepinos</e></w>
```

Observemos em detalhe o código completo em (2) quanto a *parecenſe* (palavra número 08), por exemplo: a anotação final representa as diferentes informações relativas ao termo, tais sejam, a transcrição conservadora, *parecenſe*, e as atualizações grafemática, *parecense*, e ortográfica, *parecem-se* – e pode ser lida assim: "*o termo (elemento <w>...</w>) contém três camadas de informação: (1) o original (elemento <o>... </o>) é parecenſe; (2) a primeira intervenção editorial (primeiro elemento <e>...</e>) é a atualização grafemática (atributo t="gra"), parecense; (3) a segunda intervenção editorial (segundo elemento <e>...</e>) é a modernização ortográfica (atributo t="mod"), parecem-se*". Esse procedimento permite armazenar informações relativas à transcrição conservadora e à edição modernizada em

um mesmo arquivo, como diferentes camadas de informação. Com base nesse arquivo, diferentes versões para leitura, análogas às tradicionais edição diplomática, semi-diplomática, interpretativa, etc., podem ser produzidas, como visualizações das diferentes camadas da anotação, na forma de novos arquivos de texto, cada um contendo apenas as informações codificadas como "*original*" ou "*editado*", (além disso, é possível gerar também arquivos com o glossário completo das edições realizadas – cf. Paixão de Sousa, 2013[b] para um detahamento). A Figura 7 ilustra uma versão conservadora do texto, gerada a partir do código XML parcialmente ilustrado em (2), e a Figura 8, uma versão modernizada, gerada a partir do mesmo código – essa última, a versão que segue para a etapa de anotação morfossintática, que se descreve a seguir:

A frui- |
ta della ſe chama banánas: parecenſe na feiçam com pe |
pinos, & criamſe em cachos: algũs delles há tam gran- |
des que tem de cento & cincoenta banánas pera cima.

Figura 7. Trecho de versão conservadora gerada da anotação de edição

A fruta dela se chama bananas: parecem-se na feição com pepinos, e criam-se em cachos: alguns deles há tão grandes que tem de cento e cinqüenta bananas para cima.

Figura 8. Trecho de versão modernizada gerada da anotação de edição

2.3 Anotação morfossintática

A etapa da **anotação morfossintática** consiste na identificação e codificação das classes de palavras em um texto. Nessa etapa, o texto modernizado ilustrado na Figura 8, gerado pelo código XML parcialmente mostrado em (2), por exemplo, receberia a seguinte anotação:

(3) `A/D-F fruta/N dela/P+PRO se/CL chama/VB-P bananas/N-P :/. parecem-se/VB-P+CL na/P+D-F feição/N com/P pepinos/N-P,/, e/CONJ criam-se/VB-P+CL em/P cachos/N--P :/. alguns/Q-P deles/P+PRO há/HV-P tão/ADV-R grandes/ADJ-G-P que/C tem/TR-P de/P cento/NUM e/CONJ cinqüenta/NUM bananas/N-P para/P cima/N ./.`

A anotação morfossintática é hoje a mais automatizada de todas as etapas de anotação no *Corpus*, sendo realizada pelo programa desenvolvido por Kepler

(2005, 2010), um analisador morfossintático automático com taxa de acerto de 95%, acoplado ao *e-Dictor*.

Além de gerar automaticamente o arquivo anotado, o último módulo do *e-Dictor* (Figura 9) permite corrigir as imprecisões restantes. O analisador automático é também chamado de "*etiquetador*", já que as siglas convencionadas para as classes morfossintáticas são afixadas a cada palavra como se fossem "*etiquetas*", com a sintaxe *palavra*/ETIQUETA, como se nota em (3) e na Figura 9.

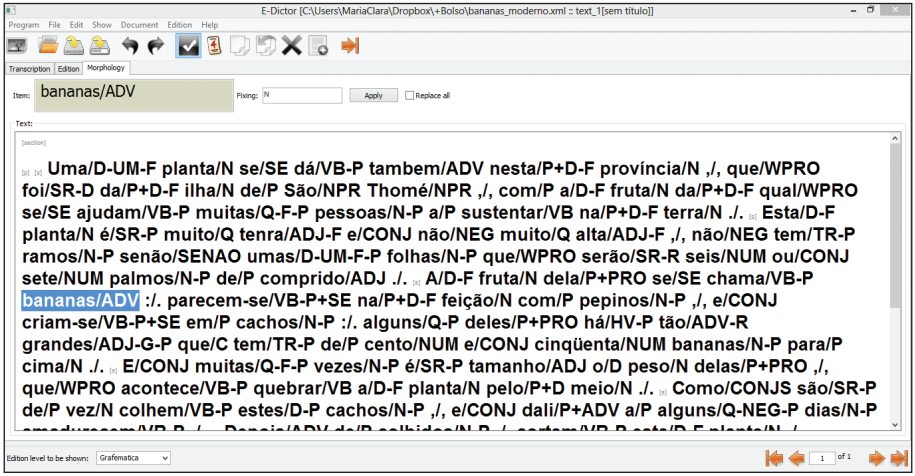

Figura 9. *e-Dictor*: Interface de correção de etiquetação
(destaque: correção de [bananas/ADV] para [bananas/N-P])

No CTB, são usadas 381 "*etiquetas*", que remetem a classes básicas (*Nome*, *Verbo*, *Preposição*, etc.), flexões (tempo, número) e aglutinações (*Preposição* + *Determinante*, etc.)[5]. O processo automático de identificação dessas categorias envolve procedimentos computacionais que não poderemos explorar com propriedade (remetemos para isso a Kepler, 2010, Kepler e Finger, 2006, Kepler, 2005, Finger, 2000 e Finger, 1998); vamos aqui apenas resumir a ideia geral do funcionamento. As programações de etiquetação podem seguir basicamente dois métodos: o método baseado em regras pré-estabelecidas, e o método probabilístico. Os etiquetadores usados no CTB, de Finger (1998) a Kepler (2010), são do tipo probabilístico, ou seja, envolvem algoritmos que calculam a probabilidade

5 Veja-se a lista completa de etiquetas em http://www.tycho.iel.unicamp.br/~tycho/corpus/manual/tags.html

da etiqueta de cada palavra, considerando a própria palavra e seu contexto. O cálculo probabilístico é aperfeiçoado por meio de sucessivas rodadas de treinamento, por tentativa e erro, no processo denominado *aprendizado de máquinas* – no caso, aprendizado de máquinas *supervisionado* ou *parcialmente supervisionado*. O aprendizado se inicia fornecendo-se, a uma versão preliminar do etiquetador, textos com as classes de palavras *manualmente* anotadas. Em cima dessa anotação manual, o etiquetador tenta inferir regras fundadas em padrões recorrentes; sucessivas rodadas de treinamento são então realizadas, até que se considere que a capacidade de inferência do etiquetador atingiu seu limite. Vamos então tentar aqui um breve resumo de como um etiquetador probabilístico poderia atribuir a cada "*palavra*" sua etiqueta nesse processo, tomando o trecho em (3) como exemplo, e destacando três casos emblemáticos. Comecemos com o caso mais simples: as palavras com frequência robusta e categoria morfossintática inequívoca – no trecho acima, seriam bons exemplos as preposições *com, em, de, para*, e a conjunção *e*, itens muito frequentes nos textos em geral, e sempre correspondentes às mesmas etiquetas (respectivamente, P e CONJ):

(4)

parecem-se na feição **com/P** pepinos	P	= preposição
e criam-se **em/P** cachos	P	= preposição
tão grandes que tem **de/P** cem a...	P	= preposição
cento e cinquenta bananas **para/P** cima	P	= preposição
e/CONJ criam-se em cachos	CONJ	= conjunção coordenativa
cento **e/CONJ** cinquenta	CONJ	= conjunção coordenativa

Nesses casos, o etiquetador, depois de analisar o corpus manualmente anotado e se deparar com um número robusto de *com*/P, *em*/P, *de*/P, *para*/P, e *e*/CONJ, sem nenhuma outra opção de etiquetação, calcula que, em um texto desconhecido qualquer, *com, em, de, para* terão 100% de probabilidade de corresponder a P, e *e* a CONJ; e aplica essa previsão. O segundo caso emblemático, e já mais complexo, é o das palavras menos frequentes nos textos em geral – no nosso trecho, por exemplo, os nomes *bananas*, *pepinos*, ou mesmo *fruta* e *cachos*:

(5)

A **fruta/N-P** dela se chama bananas	N	= nome
A fruta dela se chama **bananas/N-P**	N-P	= nome, -plural
parecem-se na **feição/N** com pepinos	N	= nome
parecem-se na feição com **pepinos/N-P**	N-P	= nome, -plural
e criam-se em **cachos/N-P**	N-P	= nome, -plural

É bastante provável que o etiquetador nunca tenha encontrado *bananas, pepinos, fruta* ou *cachos* no corpus de exemplo; e, portanto, o fato de a categoria morfossintática básica dessas palavras ser inequívoca (i.e., sempre N) é irrelevante para o programa, pois ele nunca registrou essa correspondência. Nesses casos, entram em cena cálculos probabilísticos baseados em *condições de contexto* – ou seja: nas etiquetas que seguem ou antecedem a sequência-alvo. Nos casos acima, o cálculo poderia por exemplo considerar as etiquetas antecedentes (P, em *com*/P *pepinos*, *em*/P *cachos*; VB-P, em *chama*/VB-P *bananas*; D-F, em *a*/D-F *fruta*; P+D-F, em *na*/P+D-F *feição*) e calcular, em cada caso, qual a probabilidade das palavras seguintes corresponderem a cada uma das 381 etiquetas do inventário – compensando, assim, seu desconhecimento das sequencias-alvo por seu conhecimento dos padrões possíveis em cada um desses ambientes. Nessa "aposta", o cálculo poderia acertar (como acertou a etiqueta /N, para *pepinos*, *fruta* e *cachos*), ou errar, (como errou em bananas, originalmente etiquetado como advérbio, /ADV – cf. Figura 9). O terceiro caso emblemático é o das sequências que apresentam ambiguidade morfossintática, i.e., que podem corresponder a mais de uma etiqueta – como, nesse trecho, *A*, *se*, *chama*, e *que*:

(6)

A/D-F fruta dela se chama bananas	D-F	= determinante, -feminino
A fruta dela **se/CL** chama bananas	CL	= pronome clítico
A fruta delas se **chama/VB-P** bananas	VB-P	= verbo, -presente
alguns deles há tão grandes **que/C** tem	C	= complementizador

No caso de *a*, a ambiguidade se estabelece entre a etiqueta D-F, como no exemplo, e a etiqueta P, para a preposição a (no exemplo hipotético "*disse a*/P *ela que se chama bananas*"); no caso de se, a ambiguidade é entre a etiqueta CL, como no exemplo, e as etiquetas CONJS, para conjunção subordinativa (ex., "*Seria bom se*/CONJS *pudéssemos comer bananas*") ou WQ, para interrogativas indiretas (ex., "*Perguntei se*/WQ *a fruta se chama bananas*"); no caso de *chama*, entre VB-P, verbo no presente, como no exemplo, ou N, nome (ex. "*a chama*/N *da vida*"); por fim, para *que*, a ambiguidade se estabelece entre C, complementizador, como no exemplo, e WPRO, para pronomes relativos ("*Essa fruta que*/WPRO *se chama bananas*"). Em casos como esses, a etiquetação se dá também pelo cálculo de probabilidades fundado nas condições de contexto, a exemplo do que dissemos para as palavras desconhecidas; pois, mesmo que o etiquetador conheça essas palavras ambíguas, ele registrou mais de uma correspondência *palavra*/ETIQUETA para cada uma, e portanto precisará decidir, a cada nova ocorrência, qual das possíveis correspondências está em jogo. Assim, no caso das palavras ambíguas, a fequência é menos importante que a consistência do entorno – e sequências ambíguas e muito

frequentes, como *a* e *que*, podem ser mais desafiantes que sequências ambíguas e menos frequentes, como chama, a depender da qualidade do contexto em que cada uma apareça. Por exemplo, o contexto imediato de *chama* no exemplo acima (depois de CL, e antes de N: *se*/CL *chama*/? *bananas*/N-P) é um ambiente mais provável para verbos que para nomes, e pode tornar esse caso mais fácil para a desambiguação que, por exemplo, o caso de *a*, cujo contexto imediato (antes de N, *A*/? *fruta*/N) é ambiente tão provável para preposições como para determinantes. Com isso, podemos compreender uma característica importante dos analisadores automáticos probabilísticos: sua taxa de erros não é uniforme, mas sim concentra-se em itens desconhecidos e, principalmente, em itens ambíguos. No caso do etiquetador usado atualmente no Tycho Brahe, por exemplo, segundo Kepler (2010), 25% de todos os erros remetem a duas palavras apenas: justamente, as palavras *a* e *que* (que, reiteramos, são itens muito frequentes nos textos).

O aspecto relevante para nossa discussão, neste ponto, é que isso nos mostra os fatores que determinam a precisão da etiquetação automática: primeiro, na ponta da programação, podemos ver que um fator central nessa precisão será a capacidade de *desambiguação* do algoritmo probabilístico, ou (em termos menos leigos) seu poder discriminatório para decisões de desambiguação. Esse poder discriminatório depende de diferentes aspectos, um exemplo dos quais seria a *amplitude* do ambiente de "contexto" considerado no cálculo: alguns algoritmos consideram uma etiqueta antecedente e uma subsequente, outros duas; no caso do CTB, o etiquetador (desde Kepler, 2005) é capaz de decidir caso a caso quantas etiquetas precedentes e subsequentes precisa considerar (em termos técnicos, considera cadeias de alcance variável): assim, para as decisões mais difíceis, como a etiquetação de *a* ou *que*, o analisador leva em conta um contexto maior do que considera nas decisões relativamente mais fáceis. Na outra ponta desse processo, entretanto, está o fator linguístico: além da boa programação, a precisão da etiquetação depende da qualidade do inventário de etiquetas. Primeiro, e notando o óbvio, para que seja possível extrair um padrão lógico de correspondências entre as palavras e suas etiquetas, é necessário que essa correspondência seja consistente. Segundo, e o que é menos óbvio, o inventário de etiquetas precisa ser construído levando-se em conta o limite do aprendizado da máquina como fator de equilíbrio do grau de detalhamento da análise linguística pretendida. Em resumo, quisemos ressaltar aqui dois pontos: primeiro, a precisão da análise morfossintática automática depende da aliança entre a boa programação e a boa análise linguística; e segundo, a análise linguística que se materializa numa anotação que envolva processamentos automáticos não é, nem pretende ser, a melhor análise possível do ponto de vista linguístico – mas, sim, a melhor análise linguística possível dentro dos limites do processamento por máquinas. Voltaremos a isso

depois de falarmos na última etapa da anotação linguística, a anotação sintática, na qual esses dois pontos estão também em jogo.

2.4 Anotação sintática

A terceira e última etapa de preparação dos textos no CTB é a **anotação sintática**, ou seja, a identificação e codificação de sua estrutura sintagmática. A exemplo da anotação morfossintática, esta é também uma etapa com componentes automatizados; entretanto, no CTB, a anotação sintática está atualmente menos automatizada que a etiquetação, como veremos. Do ponto de vista computacional, a tarefa de produzir uma anotação sintática automática é mais complexa que a de produzir uma etiquetação morfossintática: trata-se de construir algoritmos que que não apenas reconheçam identidades em sequências linearmente dispostas, mas ainda, que reconheçam padrões de agrupamento hierárquico – um "*parser*", ou *parseador*[6]. Aqui, novamente, não temos condições de expor com propriedade os aspectos computacionais envolvidos no funcionamento dos programas de análise sintática automática, e apenas destacaremos a ideia básica do processo.

O parseador atualmente em treinamento para o português no CTB foi desenvolvido por Bikel (2004), no âmbito do sistema *Penn-Treebank* (2014). Trata-se de um analisador probabilístico com aprendizado parcialmente supervisionado – ou seja, a exemplo do etiquetador morfossintático, o parseador começa identificando padrões consistentes e recorrentes em um corpus de treinamento, e em seguida procura calcular a probabilidade da recorrência desses padrões em instâncias inéditas (i.e., em novos textos). No início desse processo de aprendizado, o corpus consiste em um texto etiquetado morfossintaticamente nos moldes descritos mais acima, e anotado manualmente quanto à estrutura sintática. O resultado da análise automática das primeiras rodadas é então corrigido por um pesquisador, e reenviado ao parseador – e assim sucessivamente, até que o reconhecimento da máquina se aprimore; o objetivo final é obter uma boa anotação automática simplesmente a partir de um texto etiquetado. O sistema de anotação sintática aplicado aos textos do CTB é uma adaptação, para o português, do sistema *Penn-Treebank* – mais especificamente, do *Penn-Helsinki Parsed Corpus of Middle English* (Kroch & Taylor, 2000), metodologia seguida também por outros

6 Segundo Harper (2014) e Merriam-Webster (2014), o uso do verbo "*to parse*" em inglês data do século XVI, já com o sentido de "*realizar a análise sintática*"; provavelmente, segundo Harper, como reflexo da formulação "*Quae pars orationis?*", presente nas gramáticas latinas. Assim, "*to parse*" seria indiretamente derivado do nominativo latino *pars*. Nesse artigo usarei as formas aportuguesadas "parsear", "parseador", etc., em referência à análise automática.

corpora em português, como o *Cordial-Sin* (Martins, 2014) e o *Post-scriptum* (Marquilhas, 2014). Exemplos da anotação usada no CTB estão nas Figuras 10 a 13 abaixo, que repetem, por partes, a anotação do trecho já mostrado superficialmente na Figura 3. Tentaremos,no que segue, explicitar essa anotação em degraus de profundidade – das maiores às menores unidades, para facilitar a exposição; e focalizaremos essa exposição na distinção entre os aspectos automáticos e não-automáticos da análise.

Os códigos básicos da anotação são os parêntesis, (...), indicando o limite de cada estrutura e seu encaixamento hierárquico, e as siglas em maiúsculas, (X ...), indicando as categorias de cada estrutura. A disposição dos níveis em recuos tabulares pretende apenas facilitar a visualização – em termos computacionais, esse é simplesmente um esquema (X (Y (Z)(W))).

Assim, estão anotadas, em cada exemplo, uma estrutura principal, marcada como IP-MAT (i.e., "*Inflectional Phrase, Matrix*", correspondente à "oração principal"); e, no interior de cada uma, estruturas menores em diferentes níveis – do nível mais alto, diretamente encaixado a cada (IP-MAT ...), subdividindo-se em agrupamentos subsequentes, até quantos forem necessários em cada caso. A configuração de cada estrutura em categorias segue duas formas básicas: pode ser formada imediatamente a partir da classe morfossintática do elemento contido – é o caso dos verbos e conjunções, no nível superior, ex. (VB-P *chama*), (CONJ *e*), e das estruturas nos níveis mais baixos de encaixamento, ex. (N *pepinos*)) – ou pode ser "projetada" indiretamente a partir da classe morfossintática de um dos elementos contidos, ex. (NP (N *pepinos*)). Algumas estruturas são categorizadas adicionalmente quanto à função sintática, em sub-anotações como -SBJ (*Sujeito*), -ACC (*Acusativo*).

```
( (IP-MAT (NP-SBJ-1 (D-F A)
                    (N fruta)
                    (PP (P d@)
                    (NP (PRO @ela))))
          (NP-1 (CL se))
          (VB-P chama)
          (IP-SMC (NP-SBJ *-1)
                  (NP-ACC (N-P bananas)))
          (. :))  (ID G_008,17.201))
```

Figura 10. Anotação sintática da oração 17.201 (7)

7 Veja-se www.tycho.iel.unicamp.br/~tycho/corpus/texts/psd/txt/g_008_psd.txt para a anotação completa deste texto.

```
( (IP-MAT (NP-SBJ-1 *pro*)
          (VB-P parecem-)
          (NP-SE (CL -se))
          (PP (P n@)
              (NP (D-F @a) (N feição)))
          (PP (P com)
              (NP (N-P pepinos)))
          (, ,))  (ID G_008,17.202))
```

Figura 11**.** Anotação sintática da oração 17.202

```
( (IP-MAT (CONJ e)
          (NP-SBJ *pro*)
          (VB-P criam-)
          (NP-SE (CL -se))
          (PP (P em)
              (NP (N-P cachos)))
          (. :))  (ID G_008,17.203))
```

Figura 12**.** Anotação sintática da oração 17.203

```
( (IP-MAT (NP-SBJ *exp*)
          (NP-ACC (Q-P alguns)
                  (PP (P d@)
                      (NP (PRO @eles))))
          (HV-P há)
          (ADJP (ADV tão)
                (ADJ-G-P grandes)
                (CP-DEG (C que)
                        (IP-SUB (NP-SBJ *pro*)
                                (TR-P tem)
                                (PP (PP (P de)
                                        (NP (NUMP (NUM cento)
                                                  (CONJ e)
                                                  (NUM cinqüenta))
                                            (N-P bananas)))
                                    (P para)
                                    (NP (N cima))))))
          (. .))  (ID G_008,17.204))
```

Figura 13. Anotação sintática da oração 17.204

Começamos, então, com a identificação das estruturas maiores, (IP -MAT). O domínio máximo da análise do parseador é a "sentença" identificada automaticamente com base nos sinais de pontuação – notemos, nas Figuras 10 a 13, os sinais (:), (,) e (.) no final de cada IP-MAT, cada um correspondendo a uma de duas etiquetas possíveis, [/.] ou [/,], sendo [/,] aplicada apenas à própria vírgula, indicando pontuação não-final, e [/.], aplicanda a [.], [:], [;], [?] e [!], indicando pontuação final. Um texto sem pontuação final seria simplesmente tomado pelo parseador como uma imensa sentença, um único domínio de análise sobre o qual ele aplicaria seus cálculos, até encontrar uma etiqueta [/.]. Dentro do domínio assim determinado, as maiores unidades anotadas são as orações principais (IP -MAT), e em seu interior, as subordinadas (como [*que tem de cento e cinquenta bananas para cima*], no exemplo):

```
(7) ( (IP-MAT A fruta dela se chama bananas                )       (:) )
    ( (IP-MAT parecem-se na feição com pepinos             )       (,) )
    ( (IP-MAT e criam-se em cachos                         )       (:) )
    ( (IP-MAT alguns deles há tão grandes)
              (CP que tem de cento e cinquenta bananas para cima))(.) )
```

O próximo "*degrau*" seriam os sintagmas maiores, como ilustramos abaixo para as três primeiras orações principais do exemplo (sem nos ocuparmos no momento da categoria de cada sintagma, e dispensando os recuos tabulares):

```
(8) ( (IP-MAT (A fruta dela) (se)    (chama)     (bananas)     )     (:) )
    ( (IP-MAT (parecem)      (-se)   (na feição) (com pepinos) )     (,) )
    ( (IP-MAT (e)            (criam) (-se)       (em cachos)   )     (:) )
```

Observemos um ponto interessante, sobre as unidades formadas no exemplo (8): elas parecem não equivaler perfeitamente às unidades herdadas do plano morfossintático, como mostrariam os casos de *parecem-se* e *criam-se*, que no módulo morfossintático compunham unidades coesas – *parecem-se*/VB-D+CL e *criam-se*/VB-D+CL – e agora aparecem como termos independentes entre si – (*parecem*) e (*-se*) de um lado, e (*criam*) e (*-se*) de outro. Evidentemente, isso corresponde a um requerimento da análise sintática: (*parecem*) e (*-se*) e (*criam*) e (*-se*) são unidades diferentes dessa perspectiva, correspondendo a verbos de um lado e seus argumentos de outro. Outro caso em que a unidade do plano morfológico não corresponde à do sintático são as preposições aglutinadas a determinantes e pronomes. Podemos ver isso seguindo o parseamento em mais um degrau, fazendo a análise interna dos sintagmas maiores mostrados em (8) acima – (*A fruta*

dela), (*na feição*), (*com pepinos*), e (*em cachos*). Notem-se aí mais dois exemplos de sequências coesas no plano morfossintático que agora aparecem ressegmentadas: (*dela*), analisado como (*d*)(*ela*), e (*na*), analisado como (*n*)(*a*):

```
(9) ((IP-MAT ((A)(fruta)((d)(ela))) (se)     (chama)          (bananas)       ) (:))
    ((IP-MAT (parecem)              (-se)    (n((a)(feição))) ((com)(pepinos)) ) (,))
    ((IP-MAT (e)                    (criam) (-se)             ((em)(cachos))  ) (:))
```

Novamente, essa ressegmentação é necessária para que os termos possam receber uma análise sintática apropriada (que se descreve a seguir). Aqui importa notar apenas que, entre os planos morfossintático e sintático, algumas formações passam por um ajuste de segmentação – um procedimento inteiramente automático, visto que se aplica apenas sobre processos muito regulares na língua[8]. Podemos agora passar para o problema de como o parseador pode reconhecer agrupamentos como os ilustrados em (8) e subagrupamentos como os de (9). O programa trabalha considerando a combinação entre os termos etiquetados e suas etiquetas: nas primeiras rodadas sobre o corpus manualmente anotado, ele procura padrões regulares de agrupamento desses conjuntos, e nas etapas seguintes calcula a probabilidade de agrupamentos em sequências semelhantes em novos textos. Se completarmos a análise de dois dos sintagmas analisados apenas parcialmente em (9), ((*com*)(*pepinos*)) e ((*em*)(*cachos*)), acrescentando suas etiquetas e a categoria maior gerada – PP, *sintagma preposicional*, imediatamente percebemos a regularidade potencial:

```
(10) com/P pepinos/N-P:     (PP (P com)  (NP (N-P pepinos))  )
     em/P cachos/N-P:       (PP (P em )  (NP (N-P cachos ))  )
```

A regularidade da estrutura dos PPs se nota facilmente no caso de (*com pepinos*) e (*em cachos*), onde além de tudo as unidades sintagmáticas mínimas correspondem perfeitamente às unidades da anotação morfológica. Para observarmos o mesmo em (*dela*) e (*na feição*), precisamos lembrar o ajuste da segmentação de (*dela*) para (*d*)(*ela*) e de (*na*) para (*n*)(*a*) – de modo a separar em cada caso a preposição (i.e., o núcleo do PP) do restante da "*palavra*":

---

8 A segmentação é marcada com um sinal @ nos pontos de ruptura – lendo-se portanto, na anotação sintática completa (cf. Figuras 10 a 13), `(PP (P d@) (NP (PRO @ela))))`, etc. Na versão do corpus após 2013, os textos morfossintaticamente anotados já são colocados à disposição com essa segmentação marcada (cf. o de Gandavo, em http://www.tycho.iel.unicamp.br/~tycho/corpus/texts/pos/txt/g_008_pos.txt).

```
(11) dela/P+PRO          > d/P ela/PRO:   (PP (P d ) (NP (N-P ela)          ) )
     na/P+D-F feição/N   > n/P a/D-F:     (PP (P n ) (NP (D-F a)(N-P feição)) )
```

Feita a ressegmentação, a estrutura desses PPs será análoga à dos outros dois, como fica evidente se abstrairmos a estrutura interna dos NPs contidos em cada um (de que já trataremos) – ou seja, temos sempre, axiomaticamente, (PP (P)(NP…)) – o que faz dos PPs um bom exemplo de estrutura regular, cuja análise pode ser feita automaticamente com alta precisão.

```
(12) (PP (P)(NP ... ) )     exs.:   (PP (P d )  (NP ela     ) )
                                    (PP (P n )  (NP a feição) )
                                    (PP (P com) (NP pepinos ) )
                                    (PP (P em ) (NP cachos  ) )
```

Passando à estrutura interna dos NPs contidos nos PPs acima, veremos que ela é já menos regular: dois são formados por um nome, um por um nome precedido de determinante, um por um pronome – e, ainda, há o NP (*A fruta dela*), formado por um determinante, um nome e um PP (*dela*, que por sua vez contém o NP *ela*):

```
(13) (NP (N)             )  ex.:   (NP (N-P pepinos)                    )
     (NP (PRO)           )  ex.:   (NP (PRO ela)                        )
     (NP (D) (N)         )  ex.:   (NP (D-F a)    (N-P feição)          )
     (NP (D) (N) (PP ...) )  ex.:   (NP (D-F A)    (N fruta  ) (PP dela) )
```

Embora não seja tão simples como a dos sintagmas preposicionais, a estrutura dos sintagmas nominais segue ainda padrões bastante regulares; além disso, são estruturas de elevada frequência – assim, apresentam também boas chances de serem analisadas com precisão pelo parseador a partir do corpus de aprendizado. Para outras estruturas, menos regulares que os sintagmas preposicionais ou nominais, o parseador apresentará, naturalmente, menor precisão: assim como no caso da etiquetação morfossintática, o programa precisará de diversas rodadas de treinamento para poder calcular com precisão as probabilidades de agrupamento nessas formações mais complexas. O último aspecto da anotação a comentarmos é ainda mais complexo: a anotação das categorias sintáticas, que, à diferença das anotações que vimos até este ponto, não derivam simplesmente das informações do plano morfológico, mas sim dependem da identificação da função de determinados sintagmas *em relação* aos demais na estrutura. Observemos

por exemplo a anotação dos "sujeitos" nas três primeiras orações principais do nosso exemplo (e abstraíndo-se a marcação dos demais elementos):

(14) (IP-MAT (**NP-SBJ *A fruta dela***) *se chama bananas*)
(IP-MAT (**NP-SBJ *pro***) *parecem-se na feição com pepinos*)
(IP-MAT *e* (**NP-SBJ *pro***) *criam-se em cachos*)

Na primeira oração, "-SBJ" identifica o NP (*A fruta dela*) como sujeito, "NP-SBJ"; nas outras duas orações, (*parecem-se na feição com pepinos*) e (*criam-se em cachos*), a marca NP-SBJ não aparece associada a qualquer dos sintagmas anteriormente identificados – mas a uma nova entidade, **pro**, em (NP-SBJ **pro**), que não estava presente no texto original, nem no texto etiquetado. Essa identificação não é feita automaticamente pelo parseador, mas sim por um pesquisador trabalhando sobre o resultado do processamento automático. O mesmo vale para as demais categorias vazias anotadas no sistema[9] – no âmbito dos sujeitos, além de **pro**, aplicada aos nulos referenciais, há ainda as categorias **exp**, para os não-referenciais. Notemos um exemplo de sujeito nulo não-referencial (ou seja, "*expletivo*", **exp**) na última oração do nosso trecho de exemplo, (*Alguns deles há tão grandes, que tem de cento e cinquenta bananas para cima*); observe-se abaixo, em destaque, apenas a anotação do sujeito da matriz, (NP-SBJ **exp**), e de (*alguns deles*) como acusativo, NP-ACC:

(15) (IP-MAT **(NP-SBJ *exp*)(NP-ACC *alguns deles*)** *há tão grandes (que tem ...)*)

Assim, no caso de (15), a anotação de sujeito remete a um elemento que não apenas não "está", materialmente, no texto (como também é o caso da anotação dos sujeitos **pro**, em 14), mas que nem mesmo podem receber uma interpretação referencial – ou seja essa anotação remete a uma categoria puramente gramatical. Essa é, portanto, uma opção de anotação com um grau de abstração maior do que as que vimos antes, e que leva a diversas perguntas interessantes, algumas das quais abordaremos mais à frente. Antes disso, precisamos discutir as principais finalidades da anotação filológica e linguística do CTB: pois o processo de anotção do corpus – e a anotação sintática, em particular – não pode ser bem compreendido se não considerarmos sua principal finalidade: compor um registro plenamente recuperável de informações linguísticas sobre os textos.

---

9 Aqui não pretendemos detalhar cada uma – remetemos ao manual de anotação do Corpus, seção 7, para uma lista completa: http://www.tycho.iel.unicamp.br/~tycho/corpus/manual/syn-frm.html

2.5 Objetivos finais da anotação no Corpus Tycho Brahe

O "*produto final*" das três etapas de anotação do *Corpus Tycho Brahe*, cujo breve panorama acabamos de apresentar, é o conjunto dos documentos anotados, colocados à disposição para pesquisas em três formatos: o documento com anotação de edição (e como derivados, a transcrição conservadora, a modernizada e o glossário de edições), o documento com anotação morfossintática, e o documento com anotação sintática. Nesse conjunto, torna-se possível a ampla recuperação de informações – da perspectiva linguística, em particular, a recuperação das estruturas morfossintáticas e sintáticas anotadas. Este é, de fato, o objetivo fundamental do *Corpus* do ponto de vista linguístico. Desde o início da construção do CTB, diversos trabalhos fizeram uso das possibilidades de pesquisa oferecidas pela anotação morfossintática – dentre os quais citaríamos, apenas como exemplos, algumas das teses e dissertações defendidas entre 2004 e 2010 (Andrade, 2010; Gravina, 2008; Namiuti, 2008; Cavalcante, 2006; Carneiro, 2005; Paixão de Sousa, 2004) e alguns dos artigos e capítulos de livros publicados no mesmo período (Galves et al. 2005(a), Galves et al. 2005(b), Galves et al 2006). Esses trabalhos se desenvolveram, sobretudo, a partir da iniciativa da elaboração, por Namiuti (2004), de um sistema de buscas automáticas que opera sobre as etiquetas morfossintáticas; atualmente, ainda a partir da técnica assim estabelecida, é possível construir buscas na própria interface web do corpus, de forma bastante intuitiva, pelo sistema de montagem gráfica de buscas. O objetivo final e principal da anotação no CTB, entretanto, é possibilitar a pesquisa fundada na anotação sintática, e os primeiros trabalhos a fazer uso desse potencial vem aparecendo desde Galves et al 2010, Cavalcante et al. 2010 e Cavalcante et al. 2011. A busca nos textos sintaticamente anotados demanda mecanismos mais complexos que a busca pelas etiquetas morfossintáticas; no caso do CTB, as buscas nesses textos são realizadas por meio do programa "*Corpus Search*" (Randal et al., 2009), que envolve uma gramática compatível com a anotação do corpus – ou seja, que reconhece seus elementos nodais (como IP-MAT, NP-SBJ, NP-ACC, PP, etc.) e lexicais (N, P, D, etc.), e pode identificar padrões de combinação entre esses nós, por meio de operadores indicativos da posição relativa entre os elementos (*precedência*, *precedência imediata*, etc.). Imaginemos, como exemplo ilustrativo, uma busca fundada na anotação de "*sujeitos*": como vimos mais acima, o sujeito, (NP-SBJ (...)), pode ser indicado como nulo referencial, (NP-SBJ (**pro**)); nulo não-referencial (NP-SBJ (**exp**)); ou lexical, (NP-SBJ (*x*)) – a rigor, "*não-nulo*". Abaixo estão dois exemplos de buscas simples fundadas apenas nessas características, usando as siglas de anotação que já vimos e os termos de busca "*node*" (indicando qual o elemento nodal a ser considerado), "*query*" (indicando a busca), "AND" ("*concomitância*"), "*iDominates*" ("*precede imediatamente*"), e "!" ("*não*"):

(16) **(a)** *Busca por sujeitos nulos em orações matrizes:*

```
node: IP-MAT
query: (IP-MAT* iDominates NP-SBJ-*)
AND   (NP-SBJ-* iDominates *pro*)
```

(b) *Busca por sujeitos lexicais em orações matrizes*

```
node: IP-MAT
query: (IP-MAT* iDominates NP-SBJ-*)
AND   (NP-SBJ-* iDominates !*pro*|*exp*)
```

A busca em (a) pode ser lida como "*considerando as orações matrizes, procure as que contenham um sujeito, e esse sujeito seja nulo e referencial*", e a busca em (b), "*considerando as orações matrizes, procure as que contenham um sujeito, e esse sujeito não seja nulo*". Combinando-se indicações sobre a anotação do sujeito com indicações sobre o ambiente de entorno (por exemplo, sujeitos nulos referenciais em orações matrizes ou encaixadas; sujeitos lexicais *que precedam ou antecedam os verbos*; etc.), e/ou com o interior da estrutura NP-SBJ (sujeitos lexicais formados por um *nome*, por um *pronome*, etc.), seria possível construírem-se buscas mais complexas, permitindo-se portanto recuperar do corpus uma vasta gama de padrões estruturais relevantes para estudo. Nesta altura, podemos retomar questões que deixamos em aberto acima sobre a anotação dos sujeitos na nossa sentença de exemplo – vamos relembrá-la, quanto orações principais:

(17)
```
((   (NP-SBJ A fruta dela) se chama bananas (:) ))
((   (NP-SBJ *pro*)        parecem-se na feição com pepinos (,) ))
(( e (NP-SBJ *pro*)        criam-se em cachos (:)))
((   (NP-SBJ *exp*)        alguns deles há tão grandes que
     (NP-SBJ *pro*)        tem de cento e cinquenta bananas para cima (.)))
```

Como vimos, o rótulo de função "*sujeito*" pode ser atribuído a sintagmas lexicais (NP-SBJ *A fruta dela*), ou a "*categorias vazias*". Nesses casos, a anotação de sujeito é uma abstração com relação ao texto materialmente considerado, e remete à interpretação gramatical da sentença por parte de um pesquisador. A categoria (NP-SBJ **pro**) se aplica a sujeitos referenciais, ou seja, que tem uma contrapartida semântica; a categoria (NP-SBJ **exp**) se aplica aos não referenciais, ou seja, àqueles sem contrapartida semântica – trata-se aqui de representar um item estrutural, não um dos argumentos de um verbo. Além de construções com o verbo *haver*, a anotação (NP-SBJ **exp**) é aplicada a construções com

verbos como acontecer, amanhecer, etc.[10], e em certas construções específicas com os demais verbos – notadamente, construções que envolvam o pronome clítico *se* com verbos intransitivos ou complementos sentenciais (em contraste com as construções em (16) mais acima, com *se* e verbos transitivos, onde os sujeitos nulos foram anotados **pro**)[11]. No caso de *haver*, evidentemente, a anotação (NP-SBJ **exp**) remete à análise de que esse verbo, nessa construção, não apresenta argumento externo (ou seja, é "*impessoal*", nos termos clássicos). Em todos os casos, a anotação de um expeletivo remete a opção de exigir-se, para toda sentença, um sujeito gramatical. Imediatamente podem surgir, nesse ponto, diversas questões sobre a propriedade dessa análise: *terá mesmo toda sentença um sujeito gramatical, sem correspondência semântica na estrutura argumental do verbo?* E se isso for verdadeiro, *aplica-se efetivamente a (haver), numa construção como "alguns deles há tão grandes..."? Será certo identificar um sujeito nulo não referencial e argumento acusativo nessa construção?* Notemos, agora, que essas são perguntas (muito!) interessantes do ponto de vista da teoria gramatical, e é nesse âmbito que precisam ser discutidas: entretanto, elas não cabem no âmbito da discussão sobre **anotação sintática**. A opção por anotar um sujeito gramatical em cada sentença não é uma afirmação teórica da existência de um sujeito gramatical em cada sentença: é uma opção metodológica com a finalidade de permitir a recuperação de diferentes construções tornadas *comparáveis* pela anotação. Por exemplo, e ainda no caso de *haver*: um dos efeitos da decisão de

---

10 Alguns exemplos seriam as seguintes sentenças de Gandavo, 1576 (apenas os sujeitos em destaque):

(i) *E quando* (NP-SBJ *exp*)(VB-P acontece) *alagar-se alguma, os mesmos Índios, se lançam ao mar* (ID G_008,36.691)
(ii) *E quando* (NP-SBJ *exp*)(VB-P amanhece) *as mais das vezes está o ceu todo coberto de nuvens,* (ID G_008,8.31)
(iii) *e tanto que* (NP-SBJ *exp*)(VB-P anoiteceu)*, o mesmo Principal se apartou da companhia com dez ou doze flecheiros escolhidos de que ele mais se confiava* (ID G_008,39.715)

11 Ou seja, construções que possam receber uma interpretação de "*indeterminação do sujeito*". Nesses casos, além disso, o pronome "se" é indexado ao sujeito expletivo. Alguns exemplos seriam os seguintes, (a) com verbo intransitivo, (b) e (c) com complementos sentenciais (IP-INF); a co-indexação se mostra pelos índices numéricos, -1:

(i) (NP-SBJ-1 *exp*) *Até aqui* (NP-SE-1 (CL se))(VB-P navega) *por ele*
(ii) *e* (NP-SBJ-1 *exp*)(VB-P pode-)(NP-SE-1 (CL -se)(IP-INF *navegar por ele até sessenta léguas como já se navegou*)
(iii) (NP-SBJ-1 *exp*) *Pois daqui* (NP-SE-1 (CL se))(VB-P pode)(IP-INF *inferir quanto mais serão acrescentadas as fazendas daqueles que tiverem duzentos, trezentos escravos)*

anotar algumas construções com esse verbo como envolvendo um sujeito nulo e não-referencial e um argumento não-sujeito (acusativo, como em (NP-ACC *alguns deles*)) é que se pode, muito facilmente, comparar, na série temporal formada pelo *Corpus*, construções desse tipo com aquelas construções muito presentes no português mas antigo, com o verbo *haver* "pleno", em seu sentido original de "*ter, pertencer*". Vejamos abaixo um exemplo de oração com *haver* "pleno" no texto de Gandavo (1576), "*Alguns deles houveram já os portugueses às mãos*" – na qual o NP *os portugueses* é anotado como sujeito – comparada lado a lado com a da sentença anterior:

(18) (IP-MAT (**NP-ACC Alguns deles**) *houveram já* (**NP-SBJ os portugueses**) às mãos)
(IP-MAT (**NP-SBJ *exp***)(**NP-ACC alguns deles**) *há tão grandes (que tem ...)*)

Do pouco que já vimos sobre o sistema de buscas no corpus, podemos ver como se poderia montar uma busca que diferenciasse cada tipo de construção: "*busque sentenças com o verbo haver com sujeitos marcados como* (NP-SBJ **exp**)", de um lado, e : "*busque sentenças com o verbo haver com sujeitos* ***não*** *marcados como* (NP-SBJ **exp**)", de outro. Com isso, seria possível estudar os usos de *haver* numa e noutra construção ao longo da série temporal do *Corpus* – e, também, avaliar a **consistência** da aplicação de cada anotação nos textos. Chegamos então, de fato, à régua pela qual se deve medir a qualidade da anotação: a boa anotação é a anotação *consistente*, que esteja associada aos mesmos ambientes sintáticos ao longo de um corpus, formando padrões plenamente recuperáveis. A concepção das categorias de anotação linguística é portanto moldada fundamentalmente pela necessidade de conciliar os requerimentos de uma boa descrição gramatical aos requerimentos computacionais de processamento – tanto os requerimentos ligados ao treinamento do parseador, como *os do sistema de buscas*. Assim, o grande exercício, no desenvolvimento da anotação sintática, não é encontrar a melhor análise, a análise mais correta, mas sim a melhor anotação, a que possa ser aplicada de forma mais consistente possível sem gerar inconsistências nos dados. No caso de *haver*, a regra de aplicação da anotação de sujeitos é bastante simples: se houver concordância explícita no verbo, haverá um sujeito NP; se não houver concordância explícita, um sujeito expletivo (note-se que isso não significa que se afirme que nenhuma construção com *haver* sem concordância explícita possa ser uma construção com *haver* pleno – significa, apenas, que nos casos ambíguos, se escolhe uma das anotações por defeito). Voltando agora à pergunta sobre o erro ou acerto da análise de um sujeito expletivo em certas construções com *haver*, notemos primeiro que – no limite – se quisermos, podemos interpretar (NP-SBJ **exp**) como "*não ter sujeito*". E, além disso, se algum dia se decidir que esta ainda assim não é uma boa análise,

todos os casos assim anotados podem ser encontrados e reanotados segundo uma nova análise – desde que estejam anotados consistentemente. Com isso, podemos agora passar a refletir com um pouco mais de profundidade sobre as vantagens e desafios do trabalho linguístico com corpora anotados.

3 O "CORPUS ANOTADO" COMO INSTÂNCIA DE REPRESENTAÇÃO DO CONHECIMENTO

Um corpus eletrônico anotado, mais que uma coleção "*de*" textos, é um banco de dados "*sobre*" textos. A "*anotação*" é a codificação usada para representar informações nesse banco de dados, registrando-as de modo controlado, eplícito e recuperável. Um corpus eletrônico histórico, por apresentar a particularidade de ser formado a partir de textos originalmente compostos em tecnologia não digital – i.e., textos em papel, manuscritos ou impressos – inclui, entre as camadas de representação de informações, também a representação material do texto original. É essa contingência geral do corpus eletrônico (somada a essa contingência particular dos corpora eletrônicos históricos) que orienta e condiciona os procedimentos metodológicos de sua construção, e determina os efeitos de seu uso final como fonte para estudos linguísticos, como veremos.

Entretanto, para compreendermos esses impactos em profundidade, será importante antes de tudo destacarmos uma dimensão por vezes esquecida do trabalho de construção de corpora eletrônicos: a dimensão do tempo e da complexidade do trabalho de análise humana encerrados em seus produtos finais. Notamos o esquecimento dessa dimensão, em particular, quando observamos que, como principal "*vantagem*" do uso de textos anotados, muitas vezes se aponta a "*rapidez*" da pesquisa que eles permitem. Aqui será importante ponderarmos esse fator da "*rapidez*", lembrando dois apectos: o primeiro deles se depreende já dos breves exemplos de buscas sintáticas com fórmulas bastante simples que mostramos mais acima. Como se pode notar ali, a assim-chamada busca "*automática*" não é uma tarefa banal: para que dê bons resultados, é preciso não só que se conheça bem a linguagem de busca, mas, sobretudo, que se conheça muito bem a anotação aplicada aos textos. A gama de conhecimentos necessários para se realizarem as buscas nos textos com anotação sintática é portanto o primeiro fator a indicar a necessidade de ponderar o aspecto da "*rapidez*" propiciada pelos copora anotados; mas se passamos, além disso, a considerar o tempo dispendido na elaboração do corpus, veremos que esse fator se relativiza ainda mais. Pensemos por exemplo no processo atualmente mais automatizado na anotação do CTB – como vimos, a etiquetação morfossintática: de fato, hoje, aperta-se um botão, e, em segundos, surge o texto com todas as suas palavras identificadas

como *Nomes*, *Verbos*, *Preposições*, etc... Entretanto, essa produção '*instantânea*' de uma análise linguística, que aparece hoje quase como um passe de mágica, esconde quinze anos de trabalho dos cientistas da computação e linguistas que, desde 1995, trabalharam na anotação manual dos primeiros textos, experimentaram os primeiros resultados da análise automática, corrigiram-nos, debateram as decisões sobre as categorias de anotação, etc. Processo semelhante se dá ainda hoje no treinamento do parseador – todo ele fundado, vale destacar, nas primeiras rodadas realizadas sobre um texto inteiramente anotado por um pesquisador (com o trabalho de Britto, 1998); desde então, o parseador segue sendo treinado, sempre por meio da aplicação de sucessivas rodadas de correção humana e reprocessamento pelo programa. Espera-se que, com o aprofundamento do treinamento, o parseador chegue a um ponto mais automatizado, a exemplo do etiquetador. Entretanto, mesmo nesse momento futuro do amadurecimento do sistema, será importante levar-se em conta os anos tomados na sua construção desde a iniciativa de Galves (1998), antes de depositar na *rapidez* a principal vantagem da pesquisa fundada nos textos anotados. Fundamentalmente, de fato, o tempo dispendido na anotação vale a pena apenas na medida em que o sistema se torna *permanente* e *extensível* a novas pesquisas. Quando compreendemos essa dimensão do tempo e da complexidade do trabalho de análise humana encerrados nos corpora anotados, começamos também a compreender que os impactos efetivamente relevantes desse tipo de corpus remetem, mais que à rapidez, à *qualidade* do trabalho linguístico que eles permitem. Essa qualidade está ligada à característica mais importante da anotação eletrônica: ela configura uma **explicitação** da interpretação e da análise do texto, do plano filológico aos planos linguísticos. A anotação se pretende sistemática e explícita, e se oferece abertamente à crítica, ao teste e à reformulação – e o percurso que leva a essa explicitação é repleto de decisões importantes, formando uma relação muito delicada entre interpretação, análise e anotação, que discutiremos aqui apenas muito brevemente. Do ponto de vista do trabalho linguístico, por exemplo, o aspecto mais desafiador do trabalho de construção de um sistema de anotação automática – seja da anotação sintática, seja da morfossintática – é a concepção das categorias de análise que alimentarão as ferramentas computacionais. As discussões apresentadas em Britto et al. (1995), Galves e Britto (1999) e, em particular, Britto, Finger e Galves (2002), sobre os aspectos principais na concepção das categorias de análise morfossintática usadas no CTB, mostram um pouco do imenso trabalho de análise encerrado na determinação do inventário atual de etiquetas. Nesse processo, as categorias morfossintáticas que viriam a ser usadas na análise foram intensamente interrogadas e experimentadas, e em alguns casos, reformuladas, atendendo ao equilíbrio entre os requerimentos linguísticos e computacionais, até se chegar às

381 categorias hoje em uso. Um primeiro exemplo desse processo seriam as categorias atribuídas à palavra *que*: no sistema atual, duas etiquetas são possíveis nesse caso (como vimos, C e WPRO); no início, havia ainda duas outras possibilidades, CONJ, conjução de coordenação (para o *que* explicativo, como em "*eu gosto de bananas, que*/CONJ *eu não sou bobo*"), e WD, para o *que* interrogativo ("*que*/WD *fruta você gosta*?"); essas possibilidades foram abandonadas, pois o desafio de desambiguação automática era imenso, e causava uma queda considerável no desempenho geral do etiquetador. Nesse caso, portanto, a decisão de anotação precisou privilegiar um critério computacional em detrimento da análise linguística mais detalhada inicialmente desejada (e, ainda assim, podemos considerar o caso de *que* como o mais difícil desse sistema; experimentos conduzidos por Kepler, 2010, chegaram a um limite de 84% de acertos para a desambiguação entre essas duas etiquetas restantes – os mesmos experimentos, entretanto, conseguiram elevar bastante a taxa de acerto de outros itens ambíguos, como *a*, chegando a até 95% de precisão). Isso não quer dizer que não se considere que os diferentes usos de que acima não sejam reconhecidos ou valorizados – mas, apenas, que sua implementação computacional não foi factível. Um exemplo oposto, no qual os critérios linguísticos sobrepujaram o critério da eficiência computacional, seria o caso das etiquetas complexas, como P+PRO, preposição aglutinada a pronome (*dela*/P+PRO), ou P+D-F, preposição aglutinada a determinante (*na*/P+D-F), e as sub-etiquetas referentes à flexão, como *a*/D-F, determinante feminino. Essas etiquetas complexas e sub-etiquetas representaram um desafio considerável da perspectiva computacional, como mostram Finger (1998) e Britto, Finger e Galves (1999): por conta delas, o etiquetador precisou ser treinado em duas etapas, uma etapa inicial dedicada ao aprendizado das etiquetas básicas e unitárias, e uma etapa de refinamento, incluindo as sub-categorias e aglutinações. Entretanto, por se julgar imprescindível, para uma boa descrição morfossintática da língua portuguesa, que a análise automática contemplasse sua riqueza morfológica, optou-se por um sistema que pudesse identificar as flexões e os sintagmas morfologicamente aglutinados, mesmo em deterimento da rapidez do cálculo. Notemos agora que essa decisão tem também consequências importantes para a próxima etapa de anotação, a anotação sintática: como vimos, para a análise sintagmática, as aglutinações entre preposições e outras categorias precisam ser desmembradas, possibilitando assim análise dos sintagmas preposicionais, como (PP *dela*), analisado como (PP (P *d*) (NP (PRO *ela*))), e outros tratados acima. Obsevemos também um segundo ponto interessante: essa manipulação e remanipulação das unidades vai tornando a anotação gradativamente mais abstrata, conforme passamos da anotação morfossintática para a anotação sintática: a análise tomará como unidades elementos que não eram unidades no texto

– onde não havia, por exemplo, (*d*)(*ela*), mas, sim, (*dela*). Essas unidades mínimas de análise no plano sintagmático formam unidades lógicas, mais do que materiais – e isso inclui não só a análise independente de itens materialmente coesos, como também a análise de categorias *lógicas* que não estavam presentes no plano material do texto, e são acrescentadas nesse ponto da anotação (como os sujeitos nulos, que também discutimos).

A anotação sintática, portanto, é um exemplo bastante claro de que a anotação não é exatamente uma descrição do texto, mas, de fato, uma **representação sobre o texto**. Entretanto, essa qualidade representacional não é exclusividade do plano da anotação sintática: todas as etapas de construção de um corpus eletrônico são representações sobre um texto – no caso de um corpus de textos históricos, incluindo a própria transcrição. De fato: se voltarmos um pouco, e lembrarmos o texto original do nosso exemplo, observaremos que, se ali não há (*d*)(*ela*), tampouco há, por exemplo, (*pepinos*) – mas sim, (*pe*)/(*pinos*), unidades separadas por uma quebra de linha; que essas sequências formam uma só "*palavra*" é uma interpretação do editor.

Repetimos, na Figura 14, o fac-símile do texto, para lembrar sua segmentação original. No trecho, há três casos de segmentação intravocabular: (*fui-*)/(*ta*), entre as linhas 5 e 6; (*pe*)/(*pinos*), entre 6 e 7; e (*gran-*)/(*des*), entre 7 e 8 – e ressalte-se, no caso de (*pe*)/(*pinos*), a falta da indicação de separação por hífen no original.

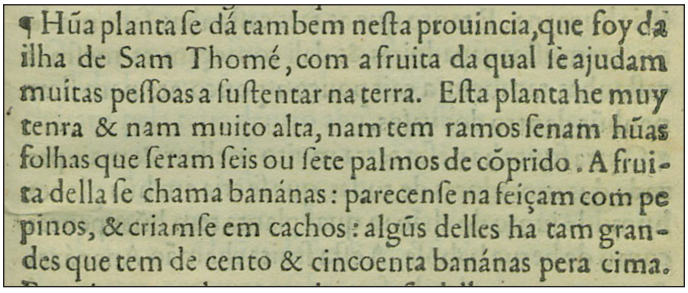
¶ Hũa planta ſe dá tambem neſta prouincia, que foy da
ilha de Sam Thomé, com a fruita da qual ſe ajudam
muitas peſſoas a ſuſtentar na terra. Eſta planta he muy
tenra & nam muito alta, nam tem ramos ſenam hũas
folhas que ſeram ſeis ou ſete palmos de cõprido. A frui-
ta della ſe chama banánas: parecenſe na feiçam com pe
pinos, & criamſe em cachos: algũs delles ha tam gran-
des que tem de cento & cincoenta banánas pera cima.

Figura 14. Detalhe do fac-simile original de Gandavo 1576

Naturalmente, a segmentação de "*palavras*", em qualquer texto escrito, é uma análise gramatical[12]. A edição filológica, ao trabalhar textos de outras épocas para a finalidade de estudos, procura ser o mais fiel possível à segmentação

12 Para um dos importantes estudos sobre o surgimento da segmentação intervocabular nas tradições de escrita da Europa Ocidental no século XIV, e sobre a interação entre essa tecnologia e os estudos gramaticais coetâneo, cf. Saenger (2004).

original, lançando mão, para isso, do conhecimento sobre a língua da época do texto – mas, também, do conhecimento sobre os constrangimentos técnicos do suporte original (no caso dos impressos, sabemos, por exemplo, que a separação do texto em unidades linguísticas entra em conflito em alguns pontos com o requerimento técnico do encaixamento na mancha gráfica da página – é esse o caso de (*fui-*)/(*ta*), (*pe*)/(*pinos*), (*gran-*)/(*des*)). Assim, é porque sabemos que (*pepinos*) é uma palavra, e porque interpretamos que a separação (*pe*)/(*pino*) não corresponde a algum fato gramatical importante, mas apenas remete a um constrangimento técnico, que mesmo em uma transcrição "conservadora" podemos transcrever (*pepino*) como unidade, e de alguma forma indicamos sua separação original. Em contraste, procuramos conservar com exatidão as segmentações que interpretamos como correspondentes a unidades de sentido – por exemplo, interpretamos que a unidade gráfica (*parecenſe*) provavelmente remete a uma unidade no plano morfológico, e talvez fonético, e por isso mantemos essa unidade na transcrição. Até aqui, não tocamos em nenhuma novidade; trata-se apenas da metodologia tradicional da edição conservadora. A diferença, na edição digital, será a codificação usada para representar a leitura. Em um corpus anotado, cada unidade de texto será indicada explicitamente; no caso do CTB, por exemplo, anotada como <o>, ou "*item original*" – como mostra (19), retomando do exemplo (1) da Seção 1 apenas a anotação dos itens originais da Figura 14 acima, mais exatamente nos pontos das quebras de linha:

(19)
```
<w><o>frui-<bk/>ta</o></w>
<w><o>pe<bk/></w><w></o><o>pinos</o></w>
<w><o>gran-<bk/>des</o></w>
```

Entretanto, surge a pergunta – qual a relevância dessa indicação por meio de tantos códigos? Por que não transcrever o texto "*exatamente como ele era*", com espaços onde havia espaços, e quebrando as linhas onde as linhas quebravam, formando assim simplesmente um texto limpo e legível? Porque, num arquivo eletrônico, não há "*texto sem código*". A opção que se coloca é apenas entre um código que controlamos, e um código externamente imposto. Ao transcrever um texto originalmente escrito em papel para um arquivo eletrônico, estamos de fato operando uma mudança de códigos: do código gráfico-visual (no qual a sequência de sinais gráficos justapostos forma o código visualmente interpretável como texto) para uma codificação matemática dos caracteres e de sua organização, que funciona em dois passos: um estágio de manipulação artificial da informação produz, em um segundo estágio, o código visualmente interpretável como texto. O primeiro estágios de codificação, entretanto, é quase sempre obliterado pela

forte ligação entre a representação digital e as características visuais do texto impresso, já no momento do processamento do texto.

A Figura 15 representa a visualização em uma tela de computador da versão do texto em 12 quanto digitada em um processador comum. O texto "*limpo*" da Figura 15 é composto por tantos ou mais códigos que o exemplo em (19) acima.

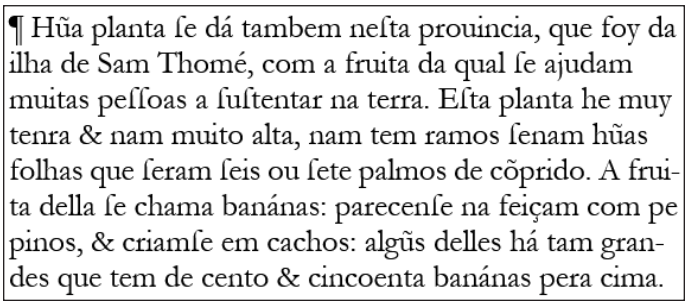
¶ Hũa planta ſe dá tambem neſta prouincia, que foy da ilha de Sam Thomé, com a fruita da qual ſe ajudam muitas peſſoas a ſuſtentar na terra. Eſta planta he muy tenra & nam muito alta, nam tem ramos ſenam hũas folhas que ſeram ſeis ou ſete palmos de cõprido. A fruita della ſe chama banánas: parecenſe na feiçam com pepinos, & criamſe em cachos: algũs delles há tam grandes que tem de cento & cincoenta banánas pera cima.

Figura 15. Tela de um processador Word 2008

De fato, e paradoxalmente, quanto mais um texto em uma tela de computador se parecer a um texto em uma página de papel impressa, mais complexa pode ser a codificação que o sustenta.O que visualizamos na tela de um computador é a última camada de um código digital *simulado* em código espacial. Nessa camada simulada, não há, por exemplo, "espaços", nem "quebras de linha"; mas sim códigos numéricos e instruções que levam à visualização desses códigos como"espaços" em "quebras de linha". Os comandos que ativamos no teclado ao digitar um texto, basicamente, inscrevem códigos numéricos no arquivo – códigos que nunca "*vemos*", já que os processadores de texto imediatamente operam sua representação em texto legível (os códigos para cada caracter alfabético remetem a imagens de letras, o código para espaço remete à visualização de um espaço, etc.), operando para isso manipulações simbólicas em diferentes graus de complexidade. O menor grau corresponde ao processamento da *codificação de caracteres*, ou seja, a codificação dos caracteres alfabéticos, do espaçamento e possivelmente das quebras de linha – da perspectiva computacional, o fato de um arquivo conter esse tipo específico de informação é o que diferencia de fato um arquivo "*de texto*" de outros arquivos. Um exemplo de formato de arquivo comum que inclui apenas a codificação de caracteres é o "*.txt*"; mas mesmo nesses arquivos, a codificação não é aparente no processamento. Por cima dessa codificação básica, arquivos mais complexos podem conter informações mais sofisticadas, relativas por exemplo à formatação e à organização espacial do texto.

Essas informações já não fazem parte do sistema de codificação de caracteres, e são adicionadas como instruções em uma camada superior, a *codificação de texto* ("*text encoding*"). Arquivos desse tipo são usados nos processadores comerciais, cujo código na maior parte dos casos é fechado – ou seja: nos quais o sistema de correspondência entre a codificação matemática de base e o texto humanamente legível não só não é visível durante o processamento, mas também é inacessível para programações fora do ambiente do processador – um exemplo é o formato *.doc* ou *.docx*, dos processadores *Word* da empresa *Microsoft*.

A Figura 16 mostra um arquivo *.txt* contendo o parágrafo ilustrado na Figura 15, conforme lido no editor de códigos HxE. Cada grupo do código hexadecimal (à direita) remete a um caracter alfabético, um espaço ou uma quebra de linha (à esquerda). A Figura 17 mostra um arquivo em formato *.docx* com o mesmo trecho da Figura 16, aberto no editor HxE. Nesse caso, na coluna onde deveria aparecer o texto correspondente à codificação, vemos a misteriosa sequência "*ËnÂ0.E÷•ú.'·Ubè¢ª**...*".

```
00000212  2E 00 20 00 41 00 20 00 66 00  .. .A. .f.
0000021C  72 00 75 00 69 00 2D 00 20 00  r.u.i.-. .
00000226  0D 00 0A 00 74 00 61 00 20 00  ....t.a. .
00000230  64 00 65 00 6C 00 6C 00 61 00  d.e.l.l.a.
0000023A  20 00 7F 01 65 00 20 00 63 00   ...e. .c.
00000244  68 00 61 00 6D 00 61 00 20 00  h.a.m.a. .
0000024E  62 00 61 00 6E 00 E1 00 6E 00  b.a.n.á.n.
00000258  61 00 73 00 3A 00 20 00 70 00  a.s.:. .p.
```

Figura 16. Visualização de codificação hexadecimal (.txt)

```
0000023A  94 CB 6E C2 30 10 45 F7 95 FA  "ËnÂ0.E÷•ú
00000244  0F 91 B7 55 62 E8 A2 AA 2A 02  .'·Ubè¢ª*.
0000024E  8B 3E 96 2D 52 E9 07 18 7B 02  ‹>–-Ré..{.
00000258  56 FD 92 C7 BC FE BE 13 02 51  Vý'Ç¼þ¾..Q
00000262  55 01 91 0A 6C 22 25 33 F7 DE  U.'.l"%3÷Þ
0000026C  33 56 C6 83 D1 DA 9A 6C 09 11  3VÆƒÑÚšl..
00000276  B5 77 25 EB 17 3D 96 81 93 5E  µw%ë.=–.“^
00000280  69 37 2B D9 D7 E4 2D 7F 64 19  i7+Ù×ä-.d.
```

Figura 17. Visualização de codificação hexadecimal (.docx)

Para a nossa discussão, são dois os pontos de relevância dessa pequena digressão sobre as entranhas do "*texto digital*" (mais detalhada em Paixão de Sousa, 2013[b]). Primeiro, notemos que, se no uso cotidiano dos arquivos eletrônicos de texto (por exemplo, para compor documentos cuja finalidade principal é serem impressos) a falta de acesso à sua codificação pode não ser um problema, no

uso com finalidade de pesquisa – por exemplo, na composição de documentos eletrônicos com objetivo de subsidiar estudos filológicos e linguísticos – a possibilidade de acessar, controlar e manipular as codificações ganha relevância. Em uma edição filológica armazenada num arquivo eletrônico de texto construído em um processador comercial comum, informações essenciais como a transliteração exata dos caracteres, o espaçamento entre as palavras, as quebras de linha, a formatação (que classicamente é usada, inclusive, para indicar convenções sobre interferências editoriais), etc., estão confiadas a um código secreto sobre o qual os editores não tem nenhum controle. Já aí se nota um primeiro impacto importante: a falta de controle sobre a codificação implica, de saída, na possibilidade da perda das informações codificadas, seja porque os códigos podem se tornar obsoletos (e portanto inteira ou parcialmente ilegíveis), seja porque a codificação se altera na transposição para outros ambientes de processamento, etc. A codificação de textos controlada, padronizada e aberta surge principalmente como resposta a esse problema: ela possibilita, à pequisa filológica e linguística, trabalhar as informações codificadas nos arquivos eletrônicos de modo controlado (cf. TEI, 2014 para uma discussão detalhada). Basicamente, isso funciona adicionando-se, a arquivos eletrônicos em formatos abertos com os caracteres codificados, codificações de texto em linguagens padronizadas – como, por exemplo, o XML, já mencionado aqui em diferentes pontos. Portanto: a "*anotação*", nesse sentido, é simplesmente a aplicação controlada de codificação de textos.

Isso significa dizer que a diferença entre um corpus eletrônico "*anotado*" e um corpus eletrônico composto por documentos construídos em processadores comerciais não é que o corpus anotado contém codificações sobre o texto – isso, ambos contém. A diferença é que, no corpus anotado, as codificações são *conhecidas* e *controladas*. Assim, é preciso ter consciência de que, quando compomos um texto em ambiente eletrônico, estamos sempre construindo um banco de dados – o arquivo eletrônico conhecido como "*texto digital*" é sempre, efetivamente, um banco de dados *sobre o texto*. Assim, nos corpora eletrônicos de textos históricos, a "*anotação de transcrição*" equivale à codificação controlada de informações relativas à representação material e linguística de um texto original nesse banco de dados, e isso envolve codificações não menos complexas que as codificações da anotação sintática – essa com objetivos diferentes, rementendo a uma representação da estrutura lógica do texto.

4 COMENTÁRIOS FINAIS

Vimos como, nos corpora eletrônicos anotados, as diferentes camadas de anotação, da transliteração e anotação de edição à anotação linguística, equivalem à

explicitação dos diferentes graus de interpretação do texto, graças à nossa compreensão e à nossa análise da língua e dos códigos gráficos de representação da linguagem (e, no caso de textos antigos, de como eles funcionam em diferentes períodos históricos). A anotação, em última instância, é a representação artificial de uma interpretação humana. Restaria, agora, refletirmos um pouco sobre como a construção dessas representações reflui, inversamente, sobre nossas análises e nossas interpretações dos textos. As considerações mais correntes sobre as consequências do trabalho computacional sobre os estudos históricos da língua costumam estar ligadas à consideração de uma característica de "rapidez" no trabalho automatizado com corpora, que já procuramos ponderar – e que, notamos agora, pode ser lembrada como ponto positivo e como ponto negativo. Há os que enxerguem, na aplicação de procedimentos automáticos sobre o trabalho filológico e linguístico, uma certa banalização do ofício tradicional das análises humanistas, como se a automatização fosse correlata a um olhar superficial sobre o texto. Nesse ponto, insisto na valorização da dimensão do trabalho de pesquisa e análise encerrado nos processos aparentemente automáticos, para mostrar que, na construção desses processos, tudo é muitas vezes pensado, muitas vezes ponderado e reconsiderado, fazendo desse trabalho uma atividade intensa de reflexão que fica muito longe de qualquer possibilidade de superficialidade. Mas isso não significa dizer que a incursão de disciplinas humanísticas tradicionais como a filologia e a linguística ao reino das técnicas computacionais não traga consequências importantes para essas disciplinas – ao contrário, as consequências são imensas, mas de outra natureza.

De fato, acredito que a expedição ao reino das tecnologias computacionais demanda das humanidades, hoje, uma profunda autoanálise crítica. O trabalho com processamentos artificial sobre o texto nos obrigam a fiar com novas fibras aquele fio que forma o tecido do trabalho das humanidades: o fio dos nossos olhares de leitura e dos nossos mecanismos interpretativos. O esforço da explicitação encerrado nas nossas metodologias de anotação, transforma profundamente nosso olhar sobre o texto: Unsworth (2004) trata dessas modificações do olhar com profundidade, ao falar nas mudanças provocadas pelo trabalho computacional sobre as nossas "*formas de atenção*" a objetos e instrumentos de conhecimento tradicionais nas humanidades, como o catálogo, o glossário, e o corpus. Por sobre a mudança na nossa "*atenção*", depositam-se as mudanças na nossa hermenêutica interpretativa: o objetivo de explicitar uma análise muda profundamente a natureza da análise. Num corpus automatizado, isso acontece, principalmente, já que o objeto dessa explicitação é uma máquina, uma calculadora superdesenvolvida, uma programação artificial que processará simbolicamente a anotação por meio de algoritmos – sem espaço para inferências fora do

que está instruído, sem intuição linguística, sem interpretação. A consequência disso é um esforço imenso e constante de decomposição instrucional – e nesse ponto, encontramos a "*imensa recompensa do exercício da consistência estúpida*" discutido por Unsworth (2001). De fato, o encontro com as limitações heurísticas da máquina nos obriga a refletir a todo momento sobre passos interpretativos antes inconscientes, e nos faz portanto, (re)descobri-los. Esses efeitos da limitação da máquina sobre nossa atitude analítica talvez componham o aspecto mais interessante, embora pouco discutido, da confluência entre os trabalhos linguístico, filológico e computacional.

Aqui voltamos ao que havíamos salientado no início dessa conversa: para alguns autores contemporâneos, esses e outros efeitos da aliança entre o humanístico e o computacional são tão significativos, que fizeram nascer um campo inteiramente novo de investigação, chamado amplamente de "*Humanidades Digitais*". Aqui não se pretende discutir as inúmeras definições em debate sobre esse campo ainda um tanto difuso (cf. Paixão de Sousa, 2011, para uma tentativa nesse sentido), e apenas salientamos que ele surge a partir da constatação das mudanças metodológicas, e talvez epistemológicas, provocadas pelo impacto das ferramentas digiais sobre o trabalho nas humanidades. Importa ressaltar, sobretudo, que esses impactos são colocados pela transformação cultural constituída pelo advento da difusão digital do texto – e as primeiras pesquisas que buscaram adaptar as metodologias tradicionais a essa nova forma de difusão são talvez, simplesmente, as que sentiram esses impactos mais precocemente. Assim, a não ser que recusemos a tecnologia digital de difusão dos textos como um todo, parece inevitável que as disciplinas do texto em geral venham a se deparar, mais cedo ou mais tarde, com seus efeitos, e precisarão refletir sobre as rupturas e os avanços colocados por esse encontro. Trouxemos, aqui, apenas algumas pinceladas desse debate, sem explorá-lo com a profundidade merecida, com duas intenções principais.

Primeiro, será importante termos em mente a discussão sobre os efeitos das tecnologias digitais no ambiente dos estudos históricos e filológicos nos próximos anos, quando, como já sugerimos, a disponibilidade de fontes documentais digitalizadas tende a aumentar. Caberá a nós decidir se esse crescimento significará de fato um adensamento de bons materiais de trabalho, permitindo-nos aprofundar o conhecimento sobre a língua e sua história, ou se irá se efetivar apenas como um aumento de volume de itens dispersos e mal codificados. Nesse sentido, a tomada de consciência sobre a importância do controle da codificação de textos, por exemplo, pode vir a se revestir de crescente importância em nossa área – como observa Crane (2010), agora, mais que nunca, "*precisamos de editores!*", ou seja, precisamos de linguistas e filólogos bem-formados, que estejam dispostos a enfrentar o desafio de sistematizar com qualidade o dilúvio de textos

digitalizados que ainda nos espera no horizonte. O trabalho com o texto antigo no âmbito da linguística histórica, nos novos moldes permitidos pela aliança entre a tradição do conhecimento filológico e as inovações do tratamento computacional, tem portanto um compromisso ético importante frente a essa nova realidade material do texto. Procurei, com esse artigo, mostrar o exemplo de um projeto pioneiro de anotação de textos históricos em língua portuguesa, no qual a aliança entre a linguística histórica e a computação foi construída de modo consistente ao longo de muitos anos de trabalho – envolvendo um grande esforço coletivo de análises e estudos que formou uma geração de "*linguistas-computeiros*", e que nos mostrou como o trabalho em linguística histórica pode ser fascinante e desafiador. Desejei ainda, com essa reflexão, homenagear os pesquisadores pioneiros das *Humanidades Digitais* no Brasil. Nesse momento em que a fortuna textual da língua portuguesa está se transmutando em arquivos eletrônicos, fica aqui, para os pesquisadores em formação, o exemplo desses primeiros projetos, que abriram o caminho para que a nossa herança cultural rematerializada em dígitos continue significativa por gerações.

REFERÊNCIAS

Andrade A. A subida de clíticos em português: um estudo sobre a variedade europeia dos séculos XVI a XX [tese]. Campinas: Universidade Estadual de Campinas; 2010.

Bikel D. On The parameter space of generative lexicalized statistical parsing models. [tese]. Philadelphia: University of Pennsylvania; 2004.

Britto HS, Finger M, Galves C. Computational and linguistics aspects of the construction of the Tycho Brahe Historical Corpus of Portuguese. In: Claus D. Pusch & Wolfgang Raible, editores Romance corpus linguistics, corpora and spoken language 2002. Tubingen: Gunter Narr Verlag; 2002. [137-146].

Britto H.S. Posição dos clíticos no português europeu dos séculos XVI a XX: análise de dados em mudança para formalização da mudança gramatical [projeto de pesquisa]. Campinas: Universidade Estadual de Campinas/Fundação de Amparo à Pesquisa do Estado de São Paulo; 1998.

Carneiro ZON. Corpus eletrônico de documentos históricos do sertão [base de dados da Internet]. Feira de Santana: Universidade Federal de Feira de Santana. Acessada em: [28/10/2014]. Disponível em: <http://www2.uefs.br/cedohs>.

Carneiro ZON. Cartas Brasileiras (1809-1904): Um estudo linguístico-filológico [tese]. Campinas: Universidade Estadual de Campinas; 2005.

Cavalcante SRO, Paixão de Sousa MC. Construções de "SE-passivo" na história do português e a posição de sujeitos e complementos. In: Costa A et al, organizador. Textos Seleccionados do XXVI Encontro Nacional da Associação Portuguesa de Linguística. Lisboa: Associação Portuguesa de Linguística/FCT; 2011. p. 153-167.

Cavalcante SRO. O uso de 'se' com infinitivo na história do português: do português clássico ao português europeu e português brasileiro modernos [tese]. Campinas: Universidade Estadual de Campinas; 2006.

Cavalcante SRO, Galves C, Paixão de Sousa MC. Topics, subjects and grammatical change: from classical to modern european portuguese. Conference Subjects in Diachrony: Grammatical Change and the Expression of Subjects; Regensburg; 2010.

Crane G (et al.). ePhilology: when the books talk to their readers. In: Siemens R, Schreibman S, editores. Blackwell companion to digital literary studies. Oxford: Blackwell; 2008. Acessado em: [28/10/2014]. Disponível em: <http://www.digitalhumanities.org/companionDLS/>.

Crane G. Give us editors! Re-inventing the edition and re-thinking the humanities. In: Online umanities Scholarship: the shape of things to come. University of Virginia: Mellon Foundation; 2010. Acessado em: [28/10/2014]. Disponível em: <http://cnx.org/content/m34316/latest/>.

Finger M. Técnicas de otimização da precisão empregadas no etiquetador Tycho Brahe. In: Nunes MGV, editor. V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR 2000); Atibaia, SP; 19-22 de Novembro de 2000. São Paulo: ICMC/USP; [2000]. pp. 141-154.

Floripi SA. Estudo da variação do determinante em sintagmas nominais possessivos do português médio ao português europeu moderno [tese]. Campinas: Universidade Estadual de Campinas; 2008.

Galves C. Padrões rítmicos, fixação de parâmetros e mudança linguística [projeto de pesquisa]. Campinas: Universidade Estadual de Campinas/Fundação de Amparo à Pesquisa do Estado de São Paulo; 1998. Acessado em: [28/10/2014]. Disponível em: <http://www.tycho.iel.unicamp.br/~tycho/prfpml/fase1>.

Galves C, Faria PPF. Tycho Brahe parsed corpus of historical portuguese [base de dados na Internet]. Campinas: Universidade Estadual de Campinas. Acessado em: [28/10/2014]. Disponível em: <http://www.tycho.iel.unicamp.br/~tycho/corpus>.

Galves C. Tycho Brahe corpus syntactic annotation system. Campinas: Universidade Estadual de Campinas. Acessado em: [28/10/2014]. Disponível em: <http://www.tycho.iel.unicamp.br/~tycho/corpus/manual/syn-frm.html>.

Galves C, Britto H. A construção do corpus anotado do português histórico Tycho Brahe: o sistema de anotação morfológica. In: Rodrigues I, Quaresma P, editores. Actas do IV Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada (PROPOR'99); Évora; 20-21 de Setembro. 1999. pp. 81-90.

Galves C, Paixão de Sousa, MC. The loss of verb-second in the history of portuguese: subject position, clitic placement and prosody. XII Diachronic Generative Syntax Conference (DiGS); 2010; Cambridge.

Gravina A. A ordem VS e as restrições de sujeito nulo no PB: estudo em um corpus diacrônico [tese]. Campinas: Universidade Estadual de Campinas; 2014.

Harper D. Online etimology dictionary [*homepage* na Internet]. Acessado em: [28/10/2014]. Disponível em: <http://www.etymonline.com/index.php>.

Kepler FN, Finger M. Part-of-speech tagging of portuguese based on variable length Markov Chains. In: Vieira R, Quaresma P, Nunes MGV, Mamede NJ, Oliveira, Dias MC, editors. Computational processing of the portuguese language: 7th International Workshop, PROPOR 2006. Berlin/Heidelberg: Springer Verlag; 2006. pp. 248-25

Kroch A, Taylor A. penn-helsinki parsed corpus of middle english [base de dados na Internet]. Philadelphia: University of Pennsylvania. Atualizado em: [30/01/2010]. Acessado em: [28/10/2014]. Disponível em: <http://www.ling.upenn.edu/hist-corpora/PPCME2-RELEASE-3/index.html/>.

Lopes, CRS. Laboratório de história do português brasileiro [base de dados na Internet]. Rio de Janeiro: Universidade Federal do Rio de Janeiro. Acessado em: [28/10/2014]. Disponível em: <http://www.letras.ufrj.br/laborhistorico. 2014>.

Marquilhas R. Projeto arquivo digital de escrita quotidiana em portugal e espanha na época moderna. Lisboa: Universidade de Lisboa. Acessado em: [28/10/2014]. Disponível em: <http://www.clul.ul.pt/pt/investigacao/462-post-scriptum-home>.

Menezes G. A colocação de clíticos nas orações coordenadas no corpus do português Tycho Brahe. Revista da Abralin. 2010;19(1): [89-105].

Merriam-Webster. The Merriam-Webster online dictionary [*homepage* na Internet]. Acessado em: [28/10/2014]. Disponível em: <http://www.merriam-webster.com/>
Namiuti C. Memória conquistense [banco de dados na Internet]. Vitória da Conquista: Universidade Estadual do Sudoeste da Bahia. Acessado em: [28/10/2014]. Disponível em: <http://www.corpora.uesb.br>.

Namiuti C. Aspectos da história gramatical do português: interpolação, negação e mudança [tese]. Campinas: Universidade Estadual de Campinas; 2008.

Namiuti C. O corpus anotado do português histórico: um avanço para as pesquisas em linguística histórica do português. Revista Virtual de Estudos da Linguagem – ReVEL. 2004;2(3):[1-9].

Paixão de Sousa MC. Grupo de pesquisas humanidades digitais [*homepage* na Internet]. São Paulo: Universidade de São Paulo. Acessado em: [28/10/2014]. Disponível em: <http://humanidadesdigitais.org/projetos/filologia>.

Paixão de Sousa MC. Língua barroca: sintaxe e história do português nos 1600 [tese]. Campinas: Universidade Estadual de Campinas; 2008.

Paixão de Sousa MC. A filologia digital em língua portuguesa: alguns caminhos. In: Banza AP, Gonçalves MF, coordenadores. Património textual e humanidades digitais: da antiga à nova filologia. Évora: Centro Interdisciplinar de História, Culturas e Sociedades da Universidade de Évora (CIDEHUS)/Fundação para a Ciência e a Tecnologia (FCT); 2013.

Paixão de Sousa MC. Texto digital: uma perspectiva material. Revista ANPOLL (Associação Nacional de Pós–Graduação e Pesquisa em Letras e Linguística). 2013;1(35):[17-60].

Paixão de Sousa MC, Kepler, FN, Faria, PPF. e-Dictor. [Programa de Computador]. Versão [1.0 beta 10]. Data de Publicação [2013]. Acessado em [28/10/2014]. Disponível em: http://edictor.net/download.

Randall B, Taylor A, Kroch A. CorpusSearch 2 [programa de computador. Versão: [2]. Data de publicação [2009]. Atualizado em: [20/11/2009]. Acessado em: [28/10/2014]. Disponível em: <http://corpussearch.sourceforge.net/>.

Saenger P. La lectura en los últimos siglos de la edad media. In: Chartier R, Cavallo G. Historia de la lectura en el mundo occidental. Madrid : Santillana, 2004. p. 211-260.

Schreibman S, Siemens R, Unsworth J, editores. A companion to digital humanities. Oxford: Blackwell; 2004. Acessado em: [28/10/2014]. Disponível em: <http://www.digitalhumanities.org/companion/>.

Sheperd T, Sardinha TB, Pinto MV, organizadores. Caminhos da linguística de corpus. Campinas: Mercado de Letras; 2010.

Unsworth J. Forms of attention: digital humanities beyond representation. The face of text: computer-assisted text analysis in the humanities. III Conference of the Canadian Symposium on Text Analysis (CaSTA); 2004; McMaster University.

Unsworth J. Knowledge representation in humanities computing. Inaugural E-humanities Lecture at the National Endowment for the Humanities; 2001 abril 03. Acessado em: [28/10/2014]. Disponível em: <http://www.iath.virginia.edu/~jmu2m/KR/>.

Xavier MF. Corpora of medieval portuguese. Tagging and parsing [projeto de pesquisa]. Cidade: Universidade Nova de Lisboa/FCT; 2014. Acessado em: [i28/10/2014]. Disponível em: <http://cipm.fcsh.unl-pt>.