Anotação multicamada para corpora históricos do Português

Multilayer annotation for Portuguese historical corpora

Aline Silva Costa[1]

Cristiane Namiuti[2]

Maria Clara Paixão de Sousa[3]

Resumo

No percurso das pesquisas realizadas com corpora embasados na metodologia de anotação do Corpus Histórico do português Tycho Brahe, foram identificados problemas e desafios a serem transpostos no respectivo sistema de anotação. Tendo como parâmetro o princípio de fidedignidade aos documentos originais para corpora históricos que contribuem também com pesquisas filológicas, este trabalho mostra uma proposta inicial de alteração no esquema de anotação morfossintática e de edições do Corpus Tycho Brahe para solucionar por ora o problema encontrado nas representações da quebra de linha/página e das edições. A solução proposta atende ao requisito de fidedignidade ao mesmo tempo em que alcança maiores conformidade e interoperabilidade.

Palavras-chave: *Corpus Histórico do português Tycho Brahe. anotação multicamada.*

Abstract

In the course of the researches conducted with corpora based on the annotation methodology of the Corpus Histórico do português Tycho Brahe, some problems and challenges in the respective annotation system were identified. Taking as a parameter the principle of reliability of original documents for historical corpora that also contribute to philological research, this work shows an initial proposal to change the morphosyntactic and editions annotation scheme Corpus Tycho Brahe's to solve the problem found in line / page break and text editions. The proposed solution meets the requirement for reliability while achieving greater compliance and interoperability.

Keywords: *Historical Corpus of Portuguese Tycho Brahe. multilayer annotation.*

[1] alinesilvacosta10@gmail.com

[2] cristianenamiuti@uesb.edu.br

[3] mariaclara@usp.br

1 Introdução

No universo das Humanidades Digitais, inúmeras possibilidades para pesquisas são abertas nas áreas de História da Língua e Filologia. A partir do diálogo destas áreas com a Linguística de Corpus, colocam-se desafios inerentes às edições digitais e se disponibiliza um leque de possibilidades para a preservação, disponibilização e análise de textos antigos.

A partir da compilação do Corpus histórico anotado do português Tycho Brahe (CTB) (GALVES; ANDRADE; FARIA, 2017), iniciada em 1998, diversas pesquisas sobre a história da língua baseadas em corpora de textos antigos têm se ancorado na metodologia de compilação e anotação desse corpus, como o Corpus de Documentos Oitocentistas de Vitória da Conquista (DOViC) (SANTOS; NAMIUTI, 2016), o Corpus Eletrônico de Documentos Históricos do Sertão (CE-DOHS) (CARNEIRO; LACERDA et al., 2012); o corpus do catálogo do projeto M.A.P. (Mulheres na América Portuguesa) (PAIXÃO DE SOUSA; MONTE, 2017), entre outros, formando assim uma "família" de corpora do português brasileiro nos moldes do CTB.

O Corpus Tycho Brahe é um corpus histórico composto de textos portugueses dos séculos XIV a XIX, os quais foram portados para o suporte digital, compondo um corpus eletrônico de fontes antigas. O Corpus DOViC é também um corpus histórico, composto de textos da língua portuguesa dos séculos XVIII e XIX. No entanto, neste corpus os textos são documentos manuscritos e a transposição para o suporte digital se dá por meio da fotografia cientificamente controlada, adotando o método Lapelinc (SANTOS; NAMIUTI, 2016). Paixão de Sousa (2006) defende que o texto digital deve manter a fidelidade ao texto original na preparação para o tratamento computacional. Ancorando-se neste preceito de fidedignidade, o sistema de anotação de edições para o CTB foi concebido com o intuito de preservar características originais dos textos, fundamentais para as pesquisas filológicas e que ora são perdidas na fase de preparação e intervenções de edição para o processamento computacional. O sistema define etiquetas para cada estrutura variante no texto, permitindo manter e recuperar os elementos originais e suas versões editadas. Tratando-se de documentos manuscritos antigos, Santos e Namiuti (2017) corroboram o requerimento da fidelidade quando

postulam que o trabalho de investigação para construir corpora deve buscar a fidedignidade entre o documento físico e sua versão digital na transposição do suporte físico do livro manuscrito para o livro digital. Dessa forma, o Corpus DOViC adotou a mesma metodologia de anotação do CTB e definiu o *workflow* de compilação do método Lapelinc, incorporando também o preceito de fidedignidade.

Os textos do Corpus Tycho Brahe também recebem anotações linguísticas nos níveis morfológico e sintático. Para o primeiro nível, é utilizada a linguagem XML (*eXtensible Markup Language*) (W3C, 2015), enquanto que a anotação sintática adota o formato Penn TreeBank (PTB) (SANTORINI, 2010). O Corpus DOViC, embasado na metodologia de anotação do CTB, acrescenta uma anotação sintática em XML para seus textos, a fim de obter as vantagens de homogeneidade e conformidade com padrões. O sistema de anotação do CTB aliado a ferramentas computacionais possibilita buscas automáticas nos textos, contribuindo com investigações em várias áreas do conhecimento, como história e filologia, além da própria linguística histórica. As potencialidades de pesquisa engendradas pelas tecnologias digitais com os textos anotados evidenciaram lacunas ainda não preenchidas pelo sistema de anotação atual do CTB, nos permitindo identificar problemas existentes e trazer à luz alguns desafios que devem ainda ser superados. No sistema atual, a marcação de edições, mais precisamente de segmentação e junção, quebra de linhas e de páginas, e informações morfológicas, impõe algumas limitações ou dificuldades para atender ao requisito de fidedignidade entre material/original e digital, e às potencialidades das buscas automáticas nos textos anotados (COSTA, 2019).

Motivados pelo objetivo de obter maior compartilhamento, mesclagem e comparação de recursos de linguagem, o Grupo de Pesquisas Humanidades Digitais da Universidade de São Paulo – USP e o Laboratório de Pesquisas em Linguística de Corpus (LaPeLinC) da Universidade Estadual do Sudoeste da Bahia - UESB (ambos instituídos em 2009) se uniram em torno de um projeto que visa a concepção e o desenvolvimento de recursos digitais focalizados no estudo histórico da língua e na curadoria digital de acervos memoriais (PAIXÃO DE SOUZA; NAMIUTI, 2019): o "Laboratório Virtual de Humanidades Digitais (LaViHD)" (USP e UESB). No âmbito deste novo projeto, diante dos problemas e desafios encontrados na anotação dos corpora embasados no CTB, reúnem-se esforços para propor uma nova anotação para os corpora históricos da língua

portuguesa anotados em XML. Este trabalho mostra, como recorte de pesquisa de tese em andamento de Costa (2019), a proposta inicial de uma nova representação dos dados de textos antigos, sugerindo alterações no sistema de anotação, especificamente das camadas morfológica e de edições, ancorando-se nas diretrizes TEI (*Text Encoding Initiative*) (TEI CONSORTIUM, 2019) como padrão para anotação estrutural e linguística, visando padronização e conformidade. A proposta tem como pilares as características de múltiplas camadas, homogeneidade na representação, e consonância com os requisitos dos corpora históricos, sobretudo a fidedignidade aos originais, não desconsiderando os aspectos computacionais envolvidos.

2 Metodologia

O desenvolvimento da proposta inicial apresentada neste trabalho considerou a manutenção ou obtenção em maior nível de alguns requisitos basilares demandados pelas pesquisas com os corpora em questão ou para anotação de corpus em geral, a saber: adequabilidade, acurácia, interoperabilidade e conformidade.

Para adequabilidade, buscou-se conservar na proposta inicial a consistência filológica e a fidelidade da representação digital em relação ao original, conforme apresentados por Paixão de Sousa (2006) e Namiuti e Santos (2017). Foi feita a comparação de trechos anotados nos esquemas de anotação atual e proposto para verificar se as informações anotadas em um se mantinham no outro. Para acurácia, certificou-se que as regras sintáticas da linguagem XML foram seguidas, não permitindo documentos XML mal formados.

A manutenção de interoperabilidade foi garantida com a conservação do formato XML já usado na anotação. A conformidade é definida pelo nível de consonância com padrões estabelecidos e para este requisito a proposta considerou as diretrizes de anotação TEI. Ao passo em que a conformidade influencia proporcionalmente na interoperabilidade, procuramos alinhar a proposta com algumas recomendações de nomenclaturas das diretrizes do padrão TEI, ao mesmo tempo em que conservamos muitas outras nesse momento inicial, focando primeiro na estrutura da anotação e deixando para momentos posteriores da pesquisa a decisão sobre os nomes a serem adotados em todo o esquema de anotação. O TEI é um sistema flexível e modular, cuja

infraestrutura permite a cada usuário criar um esquema adequado às suas necessidades, sem comprometer a interoperabilidade dos seus dados (TEI CONSORTIUM, 2019).

Para demonstrar os problemas identificados no sistema de anotação atual do CTB e as soluções propostas, foram utilizados pequenos trechos anotados de um texto anotado em XML do Corpus DOViC, intitulado "Carta de Liberdade do Cabrinha Bernardo", que corresponde a um documento manuscrito do gênero Carta de alforria, datado de 1845. O texto foi selecionado por estar com anotações completas e morfossintática e de edições, e por conter trechos com os problemas identificados e para os quais viemos propor soluções neste trabalho: quebra de linha no meio de uma palavra, com edições de junção, edição de segmentação e modernização.

3 Anotação em corpora digitais

O processo de anotação, também chamado de "taggeamento" ou etiquetagem, consiste em inserir marcações explícitas nos textos digitais para representar informações acerca de características implícitas pertencentes aos mesmos (TEI CONSORTIUM, 2019).

As anotações podem ser de vários tipos, podendo representar a estrutura do texto ou informações de caráter linguístico. As marcações linguísticas podem ser de vários níveis, representando informações morfossintáticas, sintáticas, semânticas, discursivas, etc. A anotação que marca as palavras com suas classes gramaticais é conhecida como *Part-Of-Speech tagging* (POS *tagging*) e se configura como uma anotação de nível morfossintático. *Taggers* (etiquetadores) POS são componentes de software que realizam o trabalho de etiquetar cada palavra em uma sentença com uma *tag* apropriada, que indica se determinada palavra é um substantivo, verbo, adjetivo, etc. (MANNING; SCHUTZE, 2000).

A anotação em nível sintático é realizada por meio da marcação da estrutura sintática de constituintes nas sentenças dos textos. Um corpus com essa anotação é considerado como um *treebank*, dado que a forma mais comum de representar esse tipo de informação é por meio de uma estrutura arbórea. Entretanto, o termo *treebank* não é limitado a corpora contendo representações de estrutura sintática, mas aplica-se a todos os tipos de corpora gramaticalmente analisados em forma de árvore. Existe um variado

número de formatos para representação e anotação da estrutura sintática de textos, como TIPSTER, Penn TreeBank, Susanne e NeGra (MENGEL; LEZIUS, 2000). O formato Penn TreeBank é um esquema de anotação sintática de corpora desenvolvido pela Universidade da Pensilvânia, que utiliza uma representação arbórea delimitada por parênteses etiquetados (SANTORINI, 2010).

A aplicação de diferentes tipos de anotação como anotação estrutural do texto, anotação de edições, e anotações linguísticas (sintática, morfossintática, semântica, etc) num corpus pode se configurar como uma anotação multicamada. Segundo Zeldes (2019), o termo "multicamada" para anotação de corpora não deve ser aplicado se apenas houver mais de um tipo de anotação no corpus, mas sim quando houver diferentes formas de informação ou análise independentes. As informações podem ser criadas independentemente para o mesmo texto, em diferentes épocas ou lugares, ou por diferentes pessoas; não podem, portanto, ser derivadas uma da outra. Tomemos como exemplo as anotações de lema e POS em um corpus. Tais anotações não podem ser consideradas multicamada, uma vez que estas informações estão entrelaçadas. No entanto, a anotação de edições filológicas é totalmente independente das anotações POS ou sintática, e assim, um corpus com essas anotações pode ser considerado multicamada.

Uma anotação pode ser caracterizada ainda como *stand-off* quando as marcações ou metadados são mantidos em locais separados dos dados primários. A junção das anotações com os dados primários para posterior recuperação na anotação *stand-off* se dá por meio de endereçamento de um conjunto de bytes, ou correspondência entre elementos, etc. (IDE; ROMARY; CLERGERIE, 2003). Exemplos dessa categoria são anotações baseadas em *tokens* ou frases que são diretamente anexadas à *string* relacionada com separadores ou *tags* estruturais. Já as anotações consideradas *inline* são aquelas mescladas com o os dados primários em um mesmo arquivo, o que de certa forma os altera. Há também formas híbridas de anotação combinando as abordagens *stand-off* e *inline*, como o formato XML TIGER (ECKART, 2020).

As camadas de anotação também podem ser encapsuladas e armazenadas separadas umas das outras ou mescladas num arquivo único. Sendo assim, é possível compilarmos

uma tabela resumindo as quatro possíveis "configurações" de anotação multicamada, combinando essas características. A tabela 1 mostra essa combinação.

Tabela 1- Caracterizações dos tipos de anotação multicamada

	Anotação Inline	**Anotação Stand-off**
	(dados primários juntos com os metadados)	**(dados primários separados dos metadados)**
Camadas em arquivos separados	*Inline* com arquivos separados	*Stand-off* com arquivos separados
Camadas em arquivo compartilhado	*Inline* com arquivo compartilhado	*Stand-off* com arquivo compartilhado

Fonte: Elaboração dos autores

3.1. Padrões de anotação

A conformidade, definida como o alinhamento com padrões estabelecidos, é de grande importância na definição de um esquema de anotação. Usar padrões bem documentados e aceitos facilitam o intercâmbio de dados e sua conversão para uso com novas ferramentas de anotação. A padronização e homogeneidade na anotação se fazem necessárias para obtenção de interoperabilidade e comunicação entre diferentes sistemas, atributos que permitem compartilhamento, mesclagem e comparação de recursos de linguagem. A representação XML usa uma marcação amplamente aceita para anotação de corpus, trazendo já vantagens de interoperabilidade e legibilidade por máquinas. Conquanto, é interessante considerar também padrões de anotação de textos, e mais especificamente, padrões de anotação linguística para o projeto de um sistema de anotação de corpora de língua (ZELDES, 2019; IDE; ROMARY; CLERGERIE, 2003; W3C, 2015).

A conformidade é crucial ainda para a escalabilidade do corpus, uma vez que formatos padrão trabalham mais facilmente com ferramentas automáticas de PLN (Processamento de Linguagem Natural), como *parsers*, etiquetadores e outros. Representações idiossincráticas reduzem o potencial de compatibilidade com tais ferramentas, dificultando ainda mais a anotação tanto para humanos quanto para máquinas. Um projeto de corpus pode criar um modelo próprio de anotação e

desenvolver ferramentas específicas para o formato criado no âmbito do mesmo projeto. No entanto, uma representação altamente idiossincrática torna menos provável que outros grupos de pesquisadores possam usar esse formato ou estender esse corpus (ZELDES, 2019).

Entre vários formatos ou recomendações de anotação existentes, destacaremos as diretrizes TEI (*Text Encoding Initiative)* e o padrão ISO TC37/SC4 (ISO, 2001). O TEI é um consórcio formado por instituições acadêmicas, projetos de pesquisa e pesquisadores individuais de todo o mundo, com o objetivo de alcançar uma padronização de codificação de textos em suporte eletrônico, particularmente com o propósito de intercâmbio de dados, baseado na SGML (*Standard Generalized Markup Language*). A iniciativa TEI teve início nos anos 80 e veio a tornar-se um padrão de anotação que tem sido largamente utilizado desde 1994. O consórcio desenvolve e mantém coletivamente um padrão para a representação de textos em formato digital, tendo como principal produto um conjunto de diretrizes que especificam métodos de codificação para textos legíveis por máquina, principalmente nas ciências humanas, ciências sociais e linguística, as "TEI Guidelines", as quais têm sido aplicadas na anotação linguística, particularmente na etiquetagem gramatical. O Consórcio TEI fornece uma bibliografia de publicações relacionadas ao padrão TEI e software desenvolvido para ou adaptado a ele (TEI CONSORTIUM, 2019).

O Padrão ISO TC37/SC4 é um *framework* para anotação de informação linguística desenvolvido pela Organização Internacional de Padronização (*International Organization for Standardization*). A ISO formou um subcomitê (SC4) no âmbito da Comissão Técnica 37 (TC37, *Terminology and Other Languages Resources*) com o objetivo de estabelecer padrões internacionais e recomendações para a modelagem de dados, anotação, intercâmbio de dados e avaliação de recursos linguísticos. Dentre os diversos grupos de trabalho do TC37/SC4, um grupo foi criado para prover um *framework* para anotação linguística. A intenção não é definir um esquema ou formato único e definitivo de anotação, mas fornecer uma arquitetura que possa servir de referência para diferentes esquemas de anotação, permitindo a fusão ou a comparação entre eles. A estrutura do *framework* tem como finalidade prover o máximo de flexibilidade para codificadores e anotadores, e ao mesmo tempo permitir e estimular o intercâmbio. O TC37/SC4 de trabalho considera o TEI como um padrão de anotação de

textos. As diretrizes TEI têm sido amplamente utilizadas para criação de corpora contemporâneos ou históricos, para pesquisas lexicográficas, estudos literários, descrições linguísticas, entre outras (ALUÍSIO; ALMEIDA, 2006; ISO, 2001).

4 O Corpus Tycho Brahe e o Corpus DOViC

O Corpus do Português Histórico Tycho Brahe é um corpus digital de textos antigos, compilado por pesquisadores da Universidade Estadual de Campinas (Unicamp) com objetivo de contribuir tanto com pesquisas linguísticas no âmbito da história da língua portuguesa, quanto com investigações de natureza filológica. O corpus é composto por textos em português de autores nascidos entre 1380 e 1845 (GALVES; BRITTO, 2008).

O corpus DOViC (Corpus de Documentos Oitocentistas de Vitória da Conquista) é também um corpus digital de textos antigos, compilado por pesquisadores do Laboratório de Pesquisa em Linguística de Corpus (Lapelinc) da Universidade Estadual do Sudoeste da Bahia (UESB). O corpus é constituído por documentos notariais manuscritos dos séculos XVIII e XIX pertencentes ao Primeiro Cartório de notas do Fórum de Vitória da Conquista. Os textos notariais que compõe o corpus são de natureza variada: Cartas de alforria; Testamentos; Procurações; Matrículas de escravos; Escrituras de imóveis; Atas de eleições municipais. A metodologia de edição e anotação do corpus DOViC é embasada na metodologia criada no projeto do Corpus Tycho Brahe (SANTOS; NAMIUTI, 2016).

5 Anotação do Corpus Tycho Brahe

O Corpus Tycho Brahe usa a linguagem XML para adicionar informação acerca da estrutura dos textos, informações das interferências de edição (edição de grafia, modernização, junção e segmentação de palavras), e ainda informações linguísticas no nível morfossintático.

Todas as anotações dos textos do Corpus Tycho Brahe, bem como dos corpora que seguem sua metodologia, são do tipo *inline*, ou seja, os dados primários são mesclados às anotações ou metadados dentro do mesmo arquivo. O CTB pode ser considerado um

corpus multicamada, uma vez que anota informações independentes, como edição, estrutura do texto e informações linguísticas (nos níveis morfossintático e sintático). As camadas de edição e estrutura do texto são mantidas no mesmo arquivo da camada morfossintática. Já a camada de anotação sintática é armazenada "sozinha" em um arquivo separado. Sendo assim, podemos dizer que o sistema de anotação Tycho Brahe é *inline* e multicamada, com algumas camadas em arquivo compartilhado e outras não (ZELDES, 2019).

O sistema de anotação de edições do CTB foi concebido por Paixão de Sousa (2006) no âmbito do Projeto "Memórias do Texto". As anotações são realizadas com o intuito de preservar características originais dos textos, fundamentais para as pesquisas filológicas e que ora são perdidas na fase de preparação dos textos para o tratamento computacional. Este sistema de anotação define etiquetas XML para cada estrutura variante no texto, permitindo manter e recuperar os elementos originais e suas versões editadas.

Tomemos como exemplo desse sistema de anotação um trecho anotado do documento "Carta de Liberdade do Cabrinha Bernardo", pertencente ao corpus DOViC. A Erro: Origem da referência não encontrada mostra a anotação morfológica e de edições do segmento "daVictoria" escrito no documento original. A etiqueta <w> marca os limites de uma palavra. <o> registra a forma no documento original, grafada como "daVictoria". Os elementos <m> anotam as estruturas morfológicas, e o atributo "v" recebe o valor de uma etiqueta morfossintática definida para o português brasileiro pelo Projeto Tycho Brahe. O valor "P+D-F" indica um pronome (P) amalgamado (+) a um determinante (D) feminino (F). O valor "NPR" indica a categoria de Nome Próprio. Os elementos <e> correspondem à camada de anotação de edições. O atributo "t" corresponde ao tipo de interferência realizada pelo editor. Os tipos de edição previstos no sistema de anotação são: Modernização (t="mod"), edição de grafia (t="gra"), segmentação (t="seg"), junção (t="jun") e correção (t="cor"). Edições sobrepostas são possíveis para uma mesma palavra (PAIXÃO DE SOUSA, 2007). Assim, no exemplo, temos que "<e "t=seg">da Victoria</e>" indicam que houve uma segmentação da "palavra" original "daVictoria" para "da Victoria". Uma edição de modernização (indicada por "t=m") foi acrescida, anotando "Vitória" como correspondente modernizado da palavra original "Victoria".

Figura 1: Anotação morfológica e de edições de trecho de um documento do Corpus DOViC

```
<w id="184">
    <o>daVictoria</o>
    <m v="P+D-F">da</m>
    <e t="seg">da Victoria</e>
    <m v="NPR">Victoria</m>
    <e t="mod">da Vitória</e>
</w>
```

Fonte: **(SANTOS; NAMIUTI, 2016)**.

O CTB também possui anotação das estruturas sintáticas dos textos, mas emprega um outro formato diferente da linguagem XML, o formato Penn TreeBank (PTB) (GALVES, 2008). A representação diferente entre estas duas camadas de anotação (morfológica e sintática) caracteriza uma não homogeneidade na representação, o que é uma desvantagem para fins de interoperabilidade.

6 Problemas identificados na anotação do CTB

Além da diferença nos formatos, as anotações de edições e morfológica não possuem correlação com a anotação sintática, o que significa dizer que não há nenhum mecanismo que estabeleça uma ligação ou referência entre os elementos do texto nas anotações separadas. Uma palavra anotada em sua versão original e modernizada no arquivo de edições não possui referência com a mesma palavra na anotação sintática, que registra apenas a versão editada. Sendo assim, os dados não podem ser extraídos da anotação sintática de forma idêntica à correspondente no documento manuscrito original, o que afeta o requisito de fidedignidade. Além disso, as buscas sintáticas perdem flexibilidade com a não integração destas informações, uma vez que não é possível disponibilizar resultados das buscas com diferentes versões do texto (original e modernizada).

Reiterando o requisito da fidedignidade para corpora linguístico-históricos, o sistema de anotação de edições concebido por Paixão de Sousa (2006) busca preservar as características originais dos textos, trazendo um controle de edições que permite visualizar o texto em diferentes versões. A despeito das vantagens de utilização da linguagem XML, um padrão de interoperabilidade recomendado também para anotação

linguística pelas diretrizes TEI, aliado ao controle de edições e manutenção da originalidade, algumas decisões tomadas sobre a marcação da estrutura do texto e edições impactam na recuperação das informações, podendo afetar a fidedignidade do texto digital em relação ao texto original, pelo menos no que se refere à visualização das diferentes versões.

A concepção da marcação de quebra de linhas e quebra de páginas no sistema de anotação atual impõe algumas restrições no controle das edições, dificultando a correspondência entre visualização do texto transcrito e imagem do documento original manuscrito. No sistema de anotação do CTB, tanto a quebra de linha quanto a quebra de página são marcadas com a etiqueta <bk>. O atributo "t" recebe o valor "l" quando a quebra é de linha e "p" quando a quebra é de página. <bk> é um elemento XML vazio, ou seja, não contém texto como conteúdo entre as tags de abertura e fechamento. Quebras de linha ou de página (que por sua vez também implica numa quebra de linha) podem ocorrer no meio de uma palavra. Para essas circunstâncias, concebeu-se inserir a etiqueta <bk/> no meio da palavra, exatamente no local da quebra, caracterizando o elemento de palavra original <o> como um elemento XML misto (com conteúdo textual e outro elemento inserido) (W3C, 2015). A Erro: Origem da referência não encontrada mostra um trecho do documento "Carta de Liberdade do Cabrinha Bernardo" do corpus DOViC em que há uma quebra de linha no meio da palavra "cabrinha".

Figura 2: Trecho de documento do corpus DOViC com quebra de linha

```
<w id="37">
    <o>cabri-<bk t="l" id="bk_4"/> nha</o>
    <m v="N"/>
    <e t="jun">cabri-nha</e>
    <e t="gra">cabrinha</e>
</w>
```

Fonte: **(SANTOS; NAMIUTI, 2016)**

Como XML é um padrão para metadados, há também mecanismos padronizados de busca para este formato. O Corpus DOViC utiliza a linguagem XQuery como mecanismo de busca embutido no software WebSinc (COSTA, 2015) para recuperação das informações nos arquivos XML. XQuery é uma recomendação padrão pra buscas

em arquivos XML, e fornece a função “data” para extrair o conteúdo de um elemento (W3C, 2015).

Considerando o exemplo dado na Erro: Origem da referência não encontrada, que mostra uma edição de segmentação para separar duas palavras escritas juntas (sem espaço em branco entre elas) no original, podemos visualizar um problema de natureza semântica na representação. No sistema de anotação em questão, a etiqueta <w> pode conter um elemento <o>, um ou vários elementos <m>, e vários ou nenhum elemento <e>. A etiqueta <w> marca uma palavra. No entanto, nesta estrutura hierárquica, quando há uma segmentação, a “palavra” original “daVictoria” é dividida em duas palavras. Para o anotador de categorias gramaticais (anotação POS), as duas palavras são fornecidas separadamente como entrada, as quais serão anotadas como “P+D-F” e “NPR”. Sendo assim, consideramos propor que duas etiquetas <w> representariam melhor estas informações, e a etiqueta “m” poderia ser substituída por um atributo da etiqueta <w>.

7 Anotação do Corpus Tycho Brahe

Na tentativa de obter uma maior flexibilidade nas buscas linguísticas no corpus DOViC, além do reuso de recursos computacionais, o Lapelinc concebeu uma anotação sintática correspondente ao formato PSD na linguagem XML para o Corpus DOViC, buscando obter homogeneidade e padronização no sistema de anotação. Com esse intuito, foi implementado um algoritmo de conversão do arquivo de anotação sintática PSD para a linguagem XML no software WebSinc (COSTA, 2015). O mapeamento das informações PSD para XML está descrito em detalhes em Namiuti e Costa (2014). As potencialidades da anotação de edições em XML foram exploradas pelo software WebSinc nas buscas morfossintáticas, retornando resultados na versão original ou editada do texto, o que até então não era possível com outras ferramentas de busca.

A homogeneidade conseguida com o mapeamento de anotação sintática para XML trouxe ganhos na padronização e na reutilização de recursos. O mesmo mecanismo de busca foi utilizado nos diferentes tipos de buscas linguísticas. No entanto, mesmo com a conversão para a linguagem XML feita pelo WebSinc, a anotação sintática continua não integrada às anotações morfossintática e de edições, sendo este ainda um desafio a ser transposto em um momento posterior da nossa pesquisa.

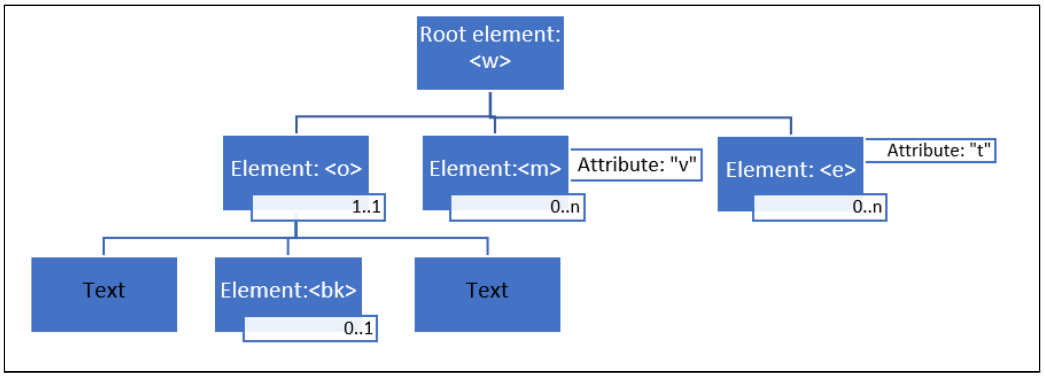
Já no sistema de anotação de edições, traçamos e apresentamos aqui uma proposta inicial alterando a estrutura hierárquica dos elementos utilizados, fazendo o mapeamento das informações existentes sugerindo eliminação ou alteração de etiquetas e atributos existentes, ou inserção de novas etiquetas e atributos. Motivados pelo requisito de conformidade com padrões de anotação, nos embasamos nas diretrizes TEI. A etiqueta recomendada para palavra pelas diretrizes é <w>, que já é utilizada no sistema de anotação atual, e portanto, consideramos que deve-se mantê-la. Todavia, propomos a eliminação da etiqueta <m>, levando a informação da categoria gramatical que era trazida no respectivo atributo "v" para o atributo "pos" da etiqueta <w>, seguindo a recomendação TEI de utilizar "pos" ou "msd". O elemento <o> deixa de ser filho de <w>, passando a ser irmão do mesmo, e ambos terão como pai um elemento que incialmente denominamos de "block". Elementos <e> continuam sendo filhos de <w>. A decisão de propor que <o> não seja filho de <w> se justifica por motivos que podem ser vistos no exemplo da Figura 1. O segmento "daVictoria" não é uma palavra, ou seja, um <w>, mas duas palavras, uma preposição e um nome próprio, que foram grafadas juntas no documento original.

A Figura 3 mostra a árvore parcial (a partir do elemento <w>[4]) da hierarquia no sistema de anotação atual e a Figura 3 mostra a estrutura hierárquica com as alterações propostas. Nas figuras, nas caixas de texto anexadas ao retângulo que representa nós do tipo elemento ou texto, indicamos a quantidade mínima e máxima de elementos que poderá estar presente na estrutura do esquema para cada ocorrência do elemento pai. Na árvore da Figura 3, por exemplo, temos que o elemento <w> pode conter zero o vários (indicado por "0..n") elementos de edição <e>. O número mínimo de ocorrências indica se o elemento é obrigatório (mínimo 1) ou opcional (mínimo 0). Optamos por representar os esquemas dessa forma gráfica hierárquica para não entrar em detalhes de recursos XML que permitem definir esquemas, como DTD (*Document Type Definition*)

[4] O elemento <w> não corresponde à raiz do documento de anotação XML do Corpus, mas está inserida dentro de outros elementos que abarcam a estrutura do texto, como sentenças (<s>), parágrafos (<p>) e seções (<sce>), entre outras. A árvore na figura é uma subárvore de toda a estrutura do documento, e <w> é a raiz desta subárvore representada nas figuras 314.

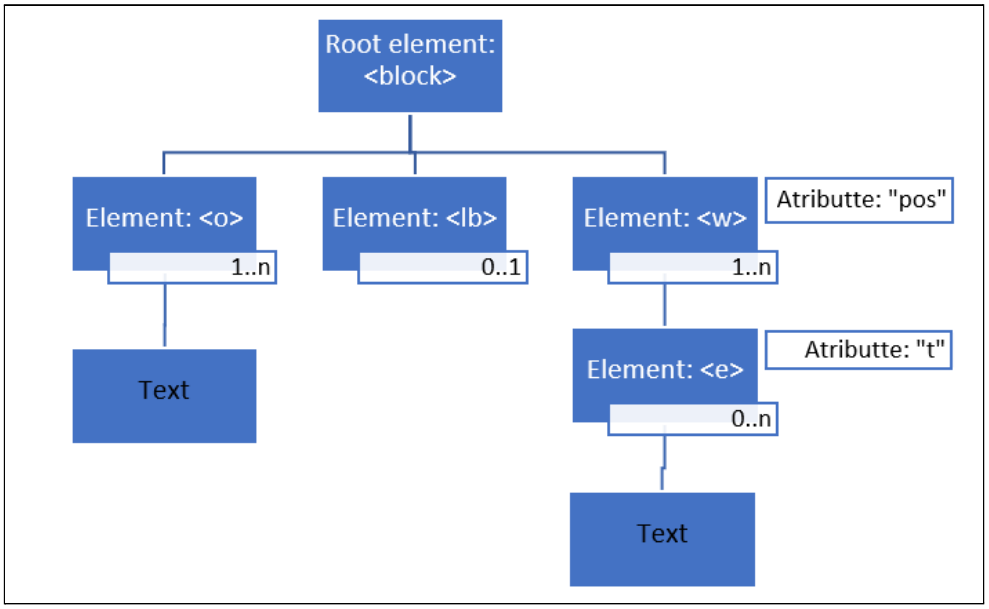
ou outro XML Schema[5]. O DTD do esquema de anotação de edições da metodologia do CTB pode ser consultada em Paixão de Sousa (2007).

Figura 3: Árvore parcial da estrutura da anotação atual



Fonte: Elaboração dos autores

Figura 4: Árvore parcial da estrutura da anotação proposta

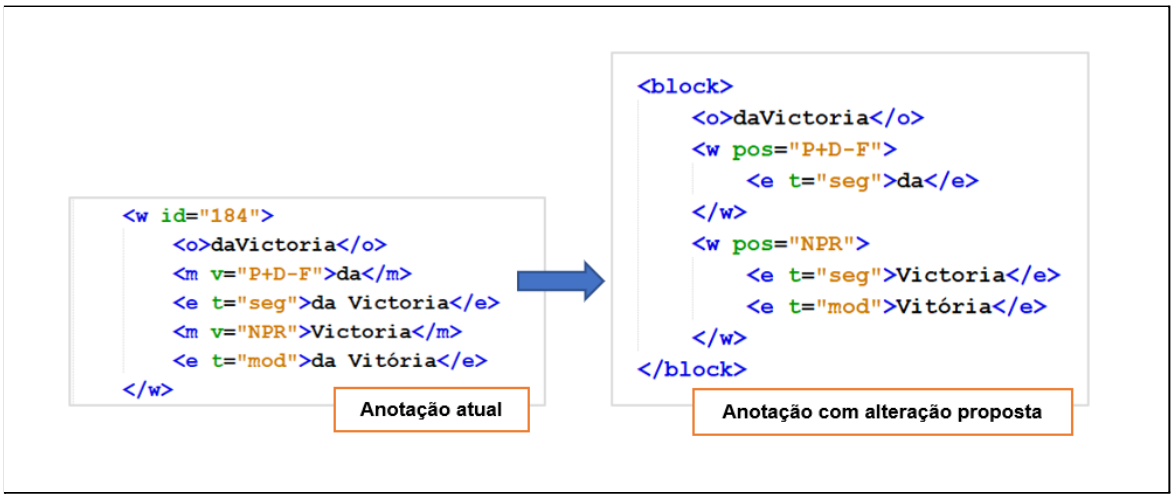


Fonte: Elaboração dos autores

[5] XML Schema (Esquema XML) é uma linguagem para expressar restrições sobre documentos XML. Existem várias linguagens de esquema em uso, mas as principais são Document Type Definitions (DTDs), Relax-NG, Schematron e W3C XSD (XML Schema Definitions) (W3C, 2015). DTD é a definição do tipo de documento, uma descrição que define a gramática válida para um documento XML, o que compreende a sintaxe e estrutura do documento (SILVA FILHO, 2004).

O resultado do trecho anotado na Figura 1 com as alterações propostas para o esquema de anotação aqui descritas é mostrado na parte direita da Figura 5. A parte esquerda da figura reproduz o mesmo exemplo dado anteriormente na Figura 1.

Figura 5: Resultado de trecho anotado com as alterações propostas



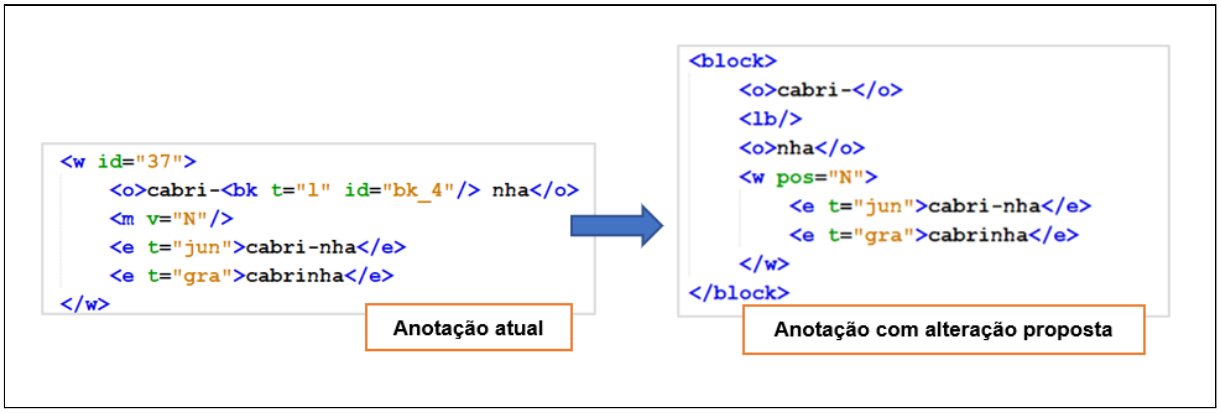
Fonte: Elaboração própria.

A marcação de quebra de linha inserida no meio de uma palavra no sistema de edições atual nos leva a reflexões sobre alternativas de representação, se considerarmos que o elemento <o>, em sendo um único elemento, melhor representa uma só "unidade de conteúdo". Dessa forma, é natural intuirmos que, quando a quebra de linha ocorrer no meio da palavra, poderiam ser utilizadas duas etiquetas ou elementos para marcar as duas partes da palavra. Retomando o exemplo da Figura 2, utilizaríamos então duas etiquetas para os dois segmentos "cabri-" e "nha", ao invés de uma única etiqueta <o> para ambos. Considerando a alteração da estrutura hierárquica já mostrada na figura 4, teríamos agora "o contrário" da segmentação do trecho da Figura 1, dois elementos <o> e um elemento <w> no mesmo bloco. Zeldes (2019) também sugere que as duas partes da palavra devem ser representadas separadamente em algum nível da anotação, quando ocorre esta situação de quebra de linha no meio da palavra, em se tratando de corpus histórico de documentos manuscritos.

Propomos ainda que a quebra de linha seja inserida entre os dois elementos <o>, usando a etiqueta recomendada pelas diretrizes TEI, <lb/> (elemento vazio para representar

uma quebra de linha (“line break”). Para quebra de páginas no meio de uma palavra, também propomos a recomendação do TEI, que corresponde à utilização de uma etiqueta <pb/> (“page break”). O resultado do trecho anotado na Figura 2 com as alterações propostas para o esquema de anotação aqui descritas é mostrado na parte direita da Figura 6. A parte esquerda da figura reproduz o mesmo exemplo dado anteriormente na Figura 2.

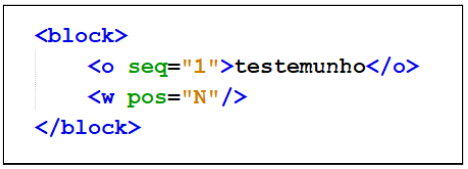
Figura 6: Resultado de trecho anotado com as alterações propostas



Fonte: Elaboração dos autores

Uma palavra que não tenha recebido nenhuma intervenção editorial será anotada apenas na forma original com a etiqueta <o>. O elemento <w> ficará vazio, já que não há nenhuma edição. O atributo “pos” será inserido depois da análise feita pelo etiquetador. Sendo assim, a Figura 7 mostra o resultado da anotação proposta para uma palavra de um texto do corpus DOViC que não sofreu edição. Se o texto ainda não tiver passado pelo processo de etiquetagem POS, a anotação poderá ser realizada com a etiqueta <w> sem o atributo “pos”.

Figura 7: Anotação de uma palavra do corpus DOViC com a alteração proposta



Fonte: Elaboração dos autores

8 Considerações finais

As alterações no esquema de anotação morfossintática e de edições do Corpus Tycho Brahe propostas nesse trabalho solucionam parcialmente problemas identificados no sistema de anotação no percurso das pesquisas com corpora embasados nessa metodologia. A alteração na estrutura hierárquica do esquema XML atrelada com mudanças de etiquetas e/ou atributos seguindo diretrizes de um padrão para anotação soluciona o problema de representação inadequada da quebra de linha e de certas intervenções editoriais, o que afetava o controle de edições e o requisito de fidedignidade postulado por Paixão de Sousa (2006) e Namiuti e Santos (2017). A solução proposta logra adequabilidade, uma vez que atende a esse requisito e mantém as informações antes representadas. A proposta também preserva a acurácia sugerindo alterações que não violam a sintaxe da linguagem XML. A utilização das diretrizes TEI como parâmetro para a proposta de correção dos problemas da anotação trazem os benefícios de conformidade e, consequentemente, de interoperabilidade.

Por se tratar de uma pesquisa em andamento, essa proposta trata-se uma solução parcial/inicial, que poderá ser modificada/refinada em qualquer das etapas seguintes do trabalho. A solução para o problema da falta de integração e homogeneidade com a camada sintática irá afetar fortemente essa proposta se a anotação resultante for do tipo *stand-off*. Uma vez que apresentamos esta proposta inicial, a próxima etapa consiste em testes experimentais com os softwares disponíveis para trabalho com corpora (de edição, de busca, anotadores, *parsers*, etc) e os que forem implementados nesta pesquisa, para verificação da viabilidade de uso e impacto de implantação da solução de anotação. Os resultados dos testes podem afetar a proposta, fazendo com esta que seja refinada e novamente aplicada aos testes experimentais, criando um ciclo iterativo de etapas. A pesquisa completa visa apresentar à comunidade de pesquisa um sistema de anotação multicamada para corpora históricos do Português, com os softwares desenvolvidos apropriadamente para a sua adoção.

Referências

ALUÍSIO, S. M.; ALMEIDA, G. M. D. B. O que é e como se constrói um corpus? Lições aprendidas na compilação de vários corpora para pesquisa linguística. **Caleidoscópio**, Brasília, v. 4, n. 3, p. 156-178, 2006. ISSN 2177-6202. Disponivel em: <http://revistas.unisinos.br/index.php/calidoscopio/article/view/6002>. Acesso em: 4 ago. 2019.

CARNEIRO, Z. et al. **CE-DOHS (Corpus Eletrônico de Documentos Históricos do Sertão)**. Feira de Santana: UEFS, 2012. Disponivel em: <http://www5.uefs.br/cedohs/view/home.html>. Acesso em: 22 nov. 2020.

COSTA, A. S. **WebSinC:** Uma Ferramenta Web para buscas sintáticas e morfossintáticas em corpora anotados - Estudo de Caso do Corpus DOViC – Bahia. Dissertação (Programa de Pós-Graduação em Linguística). Vitória da Conquista: UESB, 2015.

COSTA, A. S. **Um sistema de anotação de múltiplas camadas para o corpus digital DOViC. Projeto de Pesquisa de Doutorado (Programa de Pós-Graduação em Linguística)**. Vitória da Conquista: UESB, 2019.

ECKART, K. Chapter 3. Resource annotations. In: 5, C.-D. A. **CLARIN-D User Guide**. 1.1. ed. Stuttgart: Universität Stuttgart, 2020. Disponivel em: <https://media.dwds.de/clarin/userguide/text/>. Acesso em: 23 nov. 2020.
GALVES, C. **Tycho Brahe Parsed Corpus of Historical Portuguese. Syntatic Anotation System**. São Paulo: Unicamp, 2008. Disponivel em: <https://www.tycho.iel.unicamp.br/corpus/manual/syn-frm.html>. Acesso em: 2020 nov. 27.

GALVES, C.; ANDRADE, A. L. D.; FARIA, P. **Tycho brahe parsed corpus of historical portuguese**. [S.l.]: [s.n.], 2017. Disponivel em: <http://www.tycho.iel.unicamp.br/~tycho/corpus/en/index.html>.

GALVES, C.; BRITTO, H. **A Construção do Corpus Anotado do Português Histórico Tycho Brahe – o sistema de anotação morfológica**. Campinas: Unicamp, 2008. Disponivel em: <http://www.tycho.iel.unicamp.br/~tycho/pesquisa/artigos/GALVES_Cetal-Fase1a.pdf> . Acesso em: 13 ago. 2019.

IDE, N.; ROMARY, L.; CLERGERIE, E. D. L. International Standard for a Linguistic Annotation Framework. **WORKSHOP ON SOFTWARE ENGINEERING AND ARCHITECTURE OF LANGUAGE TECHNOLOGY SYSTEMS SEALTS**, 2003. Disponivel em: <https://www.cs.vassar.edu/~ide/papers/ide-romary-clergerie.pdf>. Acesso em: 23 fev. 2019.

ISO. **ISO TC 37/SC 4. Language Resource Management**. [S.l.]: [s.n.], 2001. Disponivel em: <https://www.iso.org/committee/297592.html>. Acesso em: 2020 nov. 19.

MANNING, C. D.; SCHUTZE, H. **Foundations of Statistical Natural Language Processing**. Massachusetts: The MIT Press, 2000.

MENGEL, A.; LEZIUS, W. An XML-based representation format for syntactically annotated corpora, 2000. Disponivel em: <http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=14E13F7984717A2C1EB5E6CB039C4C92?doi=10.1.1.26.6389&rep=rep1&type=pdf>. Acesso em: 10 fev. 2020.

NAMIUTI, C.; COSTA, A. S. Reflexões sobre anotação sintática e ferramentas de busca - Uso da linguagem XML para anotação sintática no corpus digital DOViC. **LETRAS & LETRAS**, v. 30, n. 2, dez. 2014. ISSN ISSN 1981-5239.

PAIXÃO DE SOUSA, M. C. Memória do Texto. **Revista Texto Digital**, Santa Catarina, 2006.

PAIXÃO DE SOUSA, M. C. **Sistema de Edições Eletrônicas do Corpus Tycho Brahe**. São Paulo: Unicamp, 2007. Disponivel em: <https://www.tycho.iel.unicamp.br/corpus/manual/prep/manual_frameset.html>.

PAIXÃO DE SOUSA, M. C.; MONTE, V. M. D. **Mulheres na América Portuguesa**. São Paulo: USP, 2017. Disponivel em: <http://map.prp.usp.br/>. Acesso em: 15 nov. 2020.

PAIXÃO DE SOUZA, M. C. Memórias do Texto. **Texto Digital**, n. 2, 2006. Disponivel em: <http://www.textodigital.ufsc.br/num02/paixao.htm>. Acesso em: 23 fev. 2019.

PAIXÃO DE SOUZA, M. C.; NAMIUTI, C. **Laboratório Virtual de Humanidades Digitais.** [S.l.]: Anteprojeto para Acordo de Cooperação Acadêmica Nacional, 2019. Disponivel em: <http://lavihd.fflch.usp.br/>. Acesso em: 20 nov. 2020.
SANTORINI, B. **Annotation manual for the Penn Historical Corpora and the PCEEC**, 2010. Disponivel em: <http://www.ling.upenn.edu/hist-corpora/annotation/index.html>. Acesso em: 23 fev. 2019.

SANTOS, J. V.; NAMIUTI, C. DE MANUSCRITOS HISTÓRICOS A CORPORA ANOTADOS: DO DOCUMENTO FISICO (DF) AO DOCUMENTO DIGITAL IMAGEM (DDI). **A Cor das Letras**, Feira de Santana, 2016.

SANTOS, J. V.; NAMIUTI, C. **DOViC - Documentos Oitocentistas de Vitória da Conquista. Memória Conquistense**. Vitória da Conquista: UESB/LAPELINC, 2016. Disponivel em: <http://memoriaconquistense.uesb.br/websinc>. Acesso em: 23 fev. 2019.

SANTOS, J. V.; NAMIUTI, C. **O objeto livro:** a complexidade da forma e o digital. Anais do X Congresso Internacional da ABRALIN. Niterói: UFF. 2017.

SILVA FILHO, A. M. D. **Programando com XML**. Rio de Janeiro: Campus, 2004.

TEI CONSORTIUM. **TEI P5:** Guidelines for Electronic Text. 3.6.0. ed. [S.l.]: [s.n.], 2019. Disponivel em: <https://tei-c.org/Vault/P5/3.6.0/doc/tei-p5-doc/en/html/>. Acesso em: 20 nov. 2019.

W3C. **XML Technology**. [S.l.]: [s.n.], 2015. Disponivel em: <https://www.w3.org/standards/xml/>. Acesso em: 2020 nov. 28.

ZELDES, A. **Multilayer Corpus Studies**. New York and London: Routledge, 2019.