Marcação semântica de entidades nomeadas em *Os Lusíadas*

Adriane Maria de Oliveira Queiroz
Universidade Federal da
Grande Dourados (UFGD)
Programa de Pós-Graduação
em Letras (PPGL)
Dourados, MS, Brasil
adrianeoliqueiroz@gmail.com

Bruno Oliveira Maroneze
Universidade Federal da
Grande Dourados (UFGD)
Faculdade de Comunicação,
Artes e Letras (FALE)
Dourados, MS, Brasil
brunomaroneze@ufgd.edu.br

Abstract

Este artigo apresenta a modelagem semântica de entidades nomeadas em *Os Lusíadas*, de Luís de Camões, com base no padrão TEI P5. Propõe-se um fluxo híbrido de anotação que combina NER (spaCy), dicionário de autoridade (gazetteer) e pós-edição filológica manual. São tipificados antropônimos, mitônimos e topônimos por meio dos elementos `<persName>` (nome de pessoa), `<placeName>` (nome de lugar) e `<rs>` (referencing string, para cadeias de referências), com especial atenção à marcação de epítetos. O estudo evidencia os limites de modelos treinados em corpora jornalísticos diante da sintaxe épica e da ortografia da edição de 1572, demonstrando a necessidade de uma abordagem híbrida. Conclui-se que o XML/TEI atua como ferramenta de modelagem do conhecimento literário.

1 Introdução

A preservação e o estudo de monumentos literários como *Os Lusíadas*, de Luís de Camões, têm passado por uma mudança de paradigma: a transição da digitalização meramente visual para a modelagem profunda de dados textuais. No entanto, a aplicação de técnicas de Processamento de Linguagem Natural (NLP) e Reconhecimento de Entidades Nomeadas (NER) em textos do século XVI impõe obstáculos severos. A densidade onomástica da épica camoniana, caracterizada por uma complexa rede de referências históricas e mitológicas, aliada à instabilidade ortográfica da edição de 1572 e à sintaxe latinizante, cria um cenário de *domain shift* (desvio de domínio), situação na qual modelos contemporâneos de NLP treinados em corpora jornalísticos apresentam desempenho subótimo.

Este trabalho, integrante do projeto "Lusíadas Digital", descreve a implementação de um fluxo de trabalho (*workflow*) híbrido e "human-in-the-loop" para a anotação semântica de antropônimos, mitônimos e topônimos em *Os Lusíadas*. A pesquisa não visa apenas a extração automática de dados, mas a construção de uma infraestrutura de conhecimento baseada no padrão TEI P5 (Text Encoding Initiative) (TEI Consortium, 2023). A inovação reside na integração harmônica entre a escala computacional e o rigor da tradição filológica luso-brasileira, utilizando o XML/TEI como uma ponte dialógica entre a crítica textual clássica e os métodos quantitativos das Humanidades Digitais. O tratamento lexical e onomástico do corpus dialoga com a tradição crítica camoniana e filológica, particularmente com os comentários e estudos de autores como Epiphanio da Silva Dias (Camões, 1916) e Aguiar e Silva (2010). O objetivo final é demonstrar como a colaboração entre algoritmos de inteligência artificial e a curadoria especializada permite converter o poema em um grafo de conhecimento interoperável e dinâmico.

2 Fundamentação Teórica e Modelo Semântico

A presente pesquisa situa-se na interseção entre Filologia Digital, Humanidades Digitais e Processamento de Linguagem Natural, partindo da premissa de que a modelagem textual não é apenas um procedimento técnico, mas uma operação interpretativa. A codificação em TEI P5 não representa somente a estrutura formal do texto, mas também explicita hipóteses críticas sobre identidade, referência e função narrativa.

A definição do modelo semântico em um projeto de Filologia Digital exige o que a ciência da computação denomina "compromisso ontológico": a decisão de quais categorias da realidade serão

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 293–297

representadas e como serão hierarquizadas. No caso d'*Os Lusíadas*, a escolha recaiu sobre o uso dos elementos <persName>, <placeName> e <rs>, todos pertencentes ao módulo namesdates da TEI P5. Esta escolha, contudo, não é a única possível e convida a uma reflexão sobre a granularidade e a finalidade da marcação.

Uma alternativa simplificada seria o uso da tag genérica <name>. Embora o elemento <name> reduza a complexidade da anotação automática, ele falha ao não distinguir semanticamente a natureza da entidade. Para um poema que transita entre o registro histórico e o maravilhoso pagão, a distinção entre <persName> e <placeName> é vital para futuras extrações de dados.

Dentro de <persName>, optou-se pela tipificação via atributo @type ("hist" para personagens históricos e "myth" para divindades). Discute-se na comunidade TEI se divindades deveriam possuir uma tag própria (como uma hipotética <mythName>), mas a prática consensual reforça que, funcionalmente, deuses atuam como actantes (Greimas, 1973) e pessoas no discurso épico. Em Semântica estrutural, Greimas propõe que a narrativa pode ser descrita a partir de funções estruturais abstratas (sujeito, objeto, destinador, destinatário, adjuvante e oponente) que independem da materialidade lexical do personagem. Essa distinção entre personagem empírico e função narrativa é fundamental para a modelagem digital, pois permite compreender que diferentes expressões linguísticas podem remeter à mesma instância actancial. Assim, a distinção por atributo em vez de elemento mantém a compatibilidade com ferramentas de análise de redes sociais (Social Network Analysis) (Moretti, 2011), que buscam interações entre "pessoas", independentemente de sua natureza metafísica.

O ponto mais sensível do modelo é a marcação de referências indiretas. Camões utiliza a antonomásia como recurso estilístico central (ex: "O forte Capitão" para Vasco da Gama). Existem duas vias de marcação aqui:

1. Marcar "O forte Capitão" como um nome de pessoa.

2. Marcar como uma "cadeia de referência".

A identificação automática de entidades nomeadas (NER) reconhece apenas parte desse fenômeno, já que modelos estatísticos tendem a privilegiar formas canônicas. A crítica textual, por sua vez, evidencia que expressões como "O forte Capitão" operam como marcadores identitários plenos, ainda que não apresentem um nome próprio explícito. Nesse sentido, a adoção do elemento <rs> no padrão TEI permite registrar cadeias referenciais que extrapolam a simples nomeação. A distinção entre <persName> e <rs> torna-se metodologicamente relevante: enquanto a primeira fixa uma entidade tipificada, a segunda preserva a dimensão retórica da enunciação.

Defendemos que o uso de <rs> é a abordagem superior. Marcar um epíteto como <persName> constitui um erro filológico, pois confunde o "nome de batismo" com a caracterização literária. O elemento <rs>, aliado ao atributo @ref, permite que o pesquisador mapeie a fama da personagem, ou seja, como ela é construída e referenciada indiretamente ao longo dos cantos, sem corromper a taxonomia dos nomes próprios. Essa escolha possibilita, por exemplo, o estudo estatístico da frequência com que Camões evita o nome próprio em favor do título épico.

Para os lugares, o uso de <placeName> com o atributo @type="real" ou "myth" resolve a dicotomia geográfica predominante no poema. Todavia, em versos de alta densidade erudita, o modelo enfrenta o desafio da polissemia. O verso "O Pado o sabe, o Lampetusa o sente" (I, 46) é o exemplo mais eloquente dessa complexidade, onde a marcação semântica exige o respaldo do aparato crítico tradicional para ser precisa.

A decisão de marcar "Lampetusa" apenas como um lugar geográfico (a ilha de Lampedusa), como é marcada automaticamente, ignoraria o "gesto interpretativo" inerente à camonologia. O *Dicionário d'Os Lusíadas* de Afrânio Peixoto e Pedro A. Pinto (1924) (Peixoto and Pinto, 1924), bem como o *Dicionário e Gramática de Os Lusíadas* de Júlio Nogueira (1960) (Nogueira, 1960), são categóricos ao definir Lampetusa como uma das Helíades, as irmãs de Faetonte, que choraram a morte do irmão às margens do rio Pado. O desafio de codificação amplia-se ao considerar a nota filológica de Augusto Epiphanio da Silva Dias (Camões, 1916), em sua versão comentada. Silva Dias observa que a escolha do nome por Camões reflete uma linhagem de fontes específicas: enquanto Ovídio nomeia Phaethusa e Lampetie, o nome Lampetusa ocorre nos manuscritos de Fulgêncio (Mit. i, 16) e nos comentários de Sérvio, sendo erroneamente atribuído a Ovídio por Boccaccio em Genealogiae (vii, 42).

Diante dessa estratigrafia de significados, a mar-

cação proposta nesta pesquisa refuta a *tag* única e simplista. Adotou-se uma anotação profunda no atributo `@ref`, vinculando o termo a múltiplos identificadores. No `<teiHeader>`, essas referências são "casadas" com as notas de Peixoto e Silva Dias, inseridas no elemento <note>. Assim, a hierarquia do XML não apenas identifica a palavra, mas documenta a erudição camoniana e a história da sua recepção crítica, transformando o arquivo TEI em uma ferramenta de interoperabilidade bibliográfica.

A literatura recente em Humanidades Digitais reforça que a codificação estruturada não é neutra, mas implica escolhas ontológicas. Ao atribuir tipos (@type) e referências (@ref), constrói-se uma camada semântica que pode ser explorada tanto para análises quantitativas quanto qualitativas. Desse modo, a edição digital deixa de ser apenas um repositório eletrônico e passa a funcionar como um laboratório interpretativo.

3 Metodologia e Desenvolvimento

Para essa pesquisa foram utilizados dois arquivos XML de base: a edição de 1572 (versão dextrógira), mantendo a ortografia arcaica e marcos codicológicos como quebras de linha (`<lb/>`) e assinaturas de cadernos (`<fw>`). E a edição modernizada do Projeto Gutenberg, uma versão normalizada com as regras ortográficas e gramaticais atuais, para facilitar a eficácia dos modelos de linguagem modernos. Ambas as versões foram submetidas ao mesmo processo de análise automática, com o objetivo de observar o comportamento das ferramentas de reconhecimento de entidades nomeadas em contextos textuais distintos, um com ortografia histórica e estrutura editorial preservada, e outro com ortografia modernizada. Essa comparação permitiu avaliar em que medida modelos contemporâneos de Processamento de Linguagem Natural conseguem lidar com textos literários clássicos em diferentes estágios de normalização. O processamento foi realizado em Python, estruturado em três etapas:

1. **Dicionário de Autoridade (Gazetteer):** Uma lista pré-definida de termos críticos garantiu a precisão de entidades frequentes.

2. **Modelo de Linguagem (NER):** Utilizou-se a biblioteca spaCy (modelo `pt_core_news_lg`) para identificar entidades não catalogadas.

3. **Algoritmo de Proteção de Estrutura:** Para a versão de 1572, desenvolveu-se um sistema de regex capaz de identificar nomes segmentados por tags de quebra de linha (ex: `Lusi<lb/>tana`), evitando a fragmentação do dado semântico.

A comparação entre os resultados obtidos nas duas versões evidenciou limitações no desempenho do modelo de linguagem para a identificação consistente das entidades nomeadas no corpus camoniano. Em razão disso, a marcação manual e a utilização do dicionário de autoridade mostraram-se estratégias mais eficazes e confiáveis para a anotação semântica do texto.

A utilização da biblioteca spaCy (modelo `pt_core_news_lg`) e de scripts em Python permitiu a extração célere de antropônimos e topônimos recorrentes. A automação garante a escalabilidade e a consistência terminológica, evitando omissões comuns em tarefas manuais exaustivas. No entanto, verificou-se um significativo *domain shift*. Modelos NER contemporâneos, treinados em dados jornalísticos, apresentam baixa performance diante da instabilidade ortográfica da edição de 1572 e da sintaxe épica. O algoritmo frequentemente falha na desambiguação contextual, classificando figuras mitológicas (ex: Marte) como locais geográficos (planeta) e ignorando epítetos complexos (ex: "o grão Macedônio"), que são tratados como substantivos comuns. Além disso, substantivos como "Fama" e "Mar" foram classificados erroneamente como pessoas devido à capitalização poética.

A intervenção manual foi aplicada para converter o texto digital em uma edição curada, funcionando como uma camada de pós-edição filológica, essencial para a resolução de casos de polissemia. Onde a Inteligência Artificial (IA) identifica apenas uma string, o pesquisador, subsidiado por Silva Dias (1913) e Afrânio Peixoto (1924), entre outros, identifica a densidade intertextual. O caso de "Lampetusa" (I, 46) é emblemático: enquanto o NER sugere um local, o editor humano codifica a referência mitológica às Helíades, utilizando o elemento `<rs>` para mapear a antonomásia. No entanto, a marcação manual é onerosa e dificilmente escalável para grandes volumes de dados, além de ser suscetível à subjetividade do anotador, o que pode gerar inconsistências estruturais sem o auxílio de esquemas de validação (como o RelaxNG).

A metodologia adotada utilizou a automação como um "primeiro estágio" de detecção estrutural e o dicionário de autoridade (Gazetteer) para fixar entidades inequívocas. O esforço humano concentrou-se no refinamento semântico de alto

nível, especificamente na tipificação de mitônimos e na estruturação de cadeias de referência. Essa simbiose entre o processamento algorítmico e a crítica textual permitiu que o XML final servisse tanto para análises quantitativas de frequência quanto para a recuperação qualitativa da tradição crítica camoniana. A hierarquia XML foi desenhada para garantir a interoperabilidade. O elemento `<teiHeader>` abriga a Prosopografia (relação de todas as pessoas mencionadas) e a Gazeta (relação de todos os lugares mencionados) do projeto, onde cada entidade possui um `xml:id` unívoco. No corpo do texto, o atributo `@ref` estabelece o vínculo semântico (*Linked Data*), permitindo a normalização de variantes (ex: "Lusitana", "Lusa" e "Portugal" convergem para `#portugal`).

4 Discussão

Os resultados da modelagem demonstram que a combinação entre métodos automáticos e curadoria humana produz um ganho significativo na representação semântica do texto épico. Enquanto o modelo NER identifica entidades explícitas com eficiência satisfatória, ele apresenta limitações na detecção de epítetos, antonomásias e formas alegóricas. A inserção de um *gazetteer* especializado mitiga parcialmente esse problema, mas não elimina a necessidade de intervenção crítica.

A estrutura TEI adotada permite consultas complexas que não seriam possíveis em uma edição linear. Por exemplo, torna-se viável recuperar todas as ocorrências de uma entidade independentemente de sua forma superficial, mapear a distribuição de mitônimos ao longo dos cantos ou analisar a frequência relativa de referências históricas e mitológicas. A presença sistemática do atributo `@ref` consolida essa interoperabilidade, aproximando a edição de princípios de *Linked Data*.

Entretanto, a formalização impõe limites. A categorização tripartida (antropônimos, mitônimos e topônimos) simplifica um universo referencial mais fluido, no qual certas figuras oscilam entre história e mito. Além disso, a tipificação actancial não foi plenamente automatizada, exigindo decisões interpretativas que podem variar conforme a tradição crítica adotada.

Apesar dessas restrições, o modelo proposto revela-se escalável e replicável para outros textos da tradição épica renascentista. A metodologia pode ser adaptada a diferentes corpora históricos, desde que acompanhada de uma etapa de revisão filológica rigorosa. O principal contributo reside na demonstração de que a edição digital, quando concebida como estrutura semântica relacional, amplia as possibilidades de investigação literária e historiográfica.

5 Considerações Finais

A metodologia aplicada neste estudo demonstrou que a marcação semântica de textos clássicos exige uma abordagem que transcenda a automação purista. Os resultados evidenciam que, embora modelos NER modernos (como o spaCy) ofereçam uma base sólida para a escalabilidade, eles são incapazes de capturar a densidade metafórica e as ambiguidades eruditas inerentes à poesia renascentista. A abordagem híbrida proposta — integrando *gazetteers* especializados e pós-edição humana — superou o isolamento algorítmico, especialmente na identificação de antonomásias e no tratamento de polissemias complexas, como o caso "Lampetusa", onde o dado geográfico e o mítico coexistem.

A principal contribuição deste trabalho para as Humanidades Digitais reside na validação do XML/TEI não apenas como um formato de arquivamento, mas como uma ferramenta de hermenêutica digital. Ao "casar" o texto poético com as autoridades lexicográficas de Afrânio Peixoto, Júlio Nogueira e Silva Dias, o corpus deixa de ser um silo de informação para tornar-se uma base de dados conectada (*Linked Data*). Esta arquitetura possibilita, em etapas futuras, a realização de análises de redes sociais (*Social Network Analysis*) para mapear a interação entre deuses e heróis, além da geração de cartografias digitais precisas das navegações lusitanas. Em última análise, a pesquisa reafirma que o futuro da Filologia Digital camoniana não reside na substituição do pesquisador pela máquina, mas na instrumentalização técnica do olhar crítico, garantindo que a erudição clássica seja preservada e potencializada no ecossistema digital.

References

Vítor Manuel de Aguiar e Silva. 2010. *Camões: Labirintos e Fascínios*. Cotovia, Lisboa.

Luís de Camões. 1916. *Os Lusíadas*. Companhia Portugueza Editora, Porto. Comentados por Augusto Epiphanio da Silva Dias. 2. ed. melhorada. Tomo I.

Algirdas Julien Greimas. 1973. *Semântica estrutural: pesquisa de método*. Cultrix, São Paulo.

Franco Moretti. 2011. Network theory, plot analysis. Pamphlet 2, Stanford Literary Lab.

Júlio Nogueira. 1960. *Dicionário e Gramática de "Os Lusíadas"*. Livraria Freitas Bastos S.A., Rio de Janeiro.

Afranio Peixoto and Pedro A. Pinto. 1924. *Dicionário d'Os Lusíadas de Luís de Camões*. Livraria Francisco Alves - Casa de Paulo Azevedo e Cia., Rio de Janeiro. Acesso em: 21 maio 2025.

TEI Consortium. 2023. Tei p5 guidelines.