Estratégias de modelização de dicionários latim-português como *Linked Open Data*

Lucas Consolin Dezotti[1]

[1]Departamento de Letras Clássicas e Vernáculas
Universidade Federal da Paraíba (UFPB) – João Pessoa, PB – Brasil

lucas.dezotti@academico.ufpb.br

Abstract. *This paper outlines the process of incorporating data from three bilingual Latin-Portuguese dictionaries into the LiLa Knowledge Base of interoperable linguistic resources for Latin. It focuses on how lexical and lexicographical information is modelled using the Lexicon Model for Ontologies (OntoLex-lemon) and describes the strategies employed to preserve the original text given its historical significance. The process of linking the dictionary entries to the LiLa Lemma Bank is also detailed. The outcome is the first set of interoperable Latin-Portuguese lexical resources linked to the (meta)data of other Latin linguistic resources in the LiLa Knowledge Base.*

Resumo. *Este artigo apresenta o processo de incorporação dos dados de três dicionários bilíngues latim-português à LiLa Knowledge Base, uma base de recursos linguísticos interoperáveis voltados para o latim. O foco recai sobre como a informação extraída dos dicionários é representada através do modelo OntoLex-Lemon e descreve as estratégias utilizadas para preservar o texto original, dado seu interesse histórico. O processo de interligação das entradas ao LiLa Lemma Bank é descrito em detalhes. O resultado é a primeira coleção de recursos lexicais latim-português dotada de interoperabilidade com os (meta)dados dos demais recursos linguísticos conectados à LiLa Knowledge Base.*

1. Introdução

A revolução digital trouxe novas formas tanto de produzir quanto de acessar e utilizar dados lexicográficos [Tarp 2019], com a Lexicografia passando a incluir entre seus objetivos a construção de bases de dados e infraestruturas centralizadas que favoreçam o reuso de dados para a produção de dicionários e ferramentas lexicais [Steurs et al. 2020]. A conversão de conteúdo textual em dados estruturados é um processo fundamental para que a informação possa ser buscada e processada por algoritmos [Hanks 2022, p.482–3], especialmente quando promove interoperabilidade entre a base de dados e suas eventuais aplicações. Uma forma de tornar recursos compartilhados interoperáveis é seguir o paradigma *Linked Open Data* (LOD) [Berners-Lee 2006], que estabelece como princípios:

- estar disponível na Web sob licença aberta;
- conter dados estruturados e legíveis por máquina;
- utilizar formatos abertos (não-proprietários);
- usar os padrões da W3C para identificar os dados (RDF e SPARQL);
- usar URI como nomes de coisas para interligar os dados a outros recursos e prover contexto (de preferência URL para permitir aos usuários explorar tais conexões).

A despeito do suposto tradicionalismo associado ao estudo das Línguas Clássicas, a comunidade de pesquisa sobre o latim já tem a sua disposição um grande conjunto de recursos compartilhados em formato LOD. Trata-se da LiLa Knowledge Base [Passarotti et al. 2020], uma base de conhecimento interligado em que a informação extraída de uma série de recursos linguísticos dedicados ao latim é descrita por vocabulários de representação do conhecimento, garantindo a interoperabilidade semântica entre eles e potencializando sua utilização como fonte de conhecimento. Todavia, no que tange à lexicografia bilíngue latim-português, o melhor recurso digital tem sido o *Corpus Lexicográfico do Português* (CLP) [Verdelho and Silvestre 2002], uma grande coleção de dicionários históricos (sécs. XVI–XIX) em formato texto‘.

Este artigo descreve o processo de inclusão do conteúdo de três dicionários históricos latim-português [Cardoso 1570, Velez 1744, Fonseca 1798] na LiLa Knowledge Base, a partir do texto base disponível no CLP . Para tanto, as seções a seguir descrevem os métodos utilizados na segmentação e modelização da informação lexical como LOD, com particular atenção às estratégias empregadas para preservar integralmente o conteúdo textual dos verbetes de modo compatível com o vocabulário das ontologias, bem como as técnicas de mapeamento e interligação das entradas dos dicionários à coleção de lemas da LiLa Knowledge Base. Por fim, a discussão dos resultados enseja a apresentação de exemplos de uso do novo recurso e perspectivas de pesquisa futura.

2. Segmentação do conteúdo

O verbete lexicográfico contemporâneo é um tipo de texto relativamente bem estruturado, com informações organizadas e identificadas por padrões lógicos e visuais amplamente reconhecidos [Hartmann 2001, Welker 2004]. Do ponto de vista histórico, sabe-se que a estrutura atualmente estabelecida foi obra de séculos de desenvolvimento da lexicografia moderna a partir dos glossários medievais [Merrilees 1996], cuja estrutura binária mínima, formada simplesmente de palavra-entrada (lema) e explicação semântica (definição sinonímica), foi sendo gradativamente ampliada pela inclusão de uma variedade de informações suplementares, de natureza metalinguística, em torno dos elementos básicos, seja no espaço entre o lema e a definição, seja no espaço após a definição – isto é, em posição pós-lemática (PL) ou pós-definicional (PD), respectivamente –, porém sem uma associação estável entre estas e o tipo de informação, desenvolvida apenas em estágios posteriores.

Os primeiros dicionários bilíngues latim-português pertencem justamente a essa fase de transição. Com efeito, a análise posicional de uma amostra dos três dicionários selecionados neste estudo (Tabela 1) evidencia uma diversidade de tipos de informação lexicográfica em ambas as posições: categorizações gramaticais (e.g. “frequẽtatiuo’ i.e. tipo derivacional, ‘f.’ i.e. gênero feminino); etimologias (e.g. ‘a dico, is’ e ‘ab Ex & Solvo’ indicam elementos de formação); padrões sintáticos (e.g. ‘aliquem aliqua re’ indica a forma dos complementos verbais); e indicação de autoridade (‘Hor.’ associa o poeta latino Horácio ao significado fornecido). Note-se, ademais, que um mesmo tipo de informação pode ocorrer em posições diferentes, como é o caso da informação etimológica, em posição pós-definicional no primeiro verbete e pós-lemática no segundo. A mesma amostra revela que o esquema de análise posicional é uma forma eficaz de segmentação do texto extraído dos dicionários em tela, uma vez que provê um padrão estrutural único que pode ser aplicado em larga escala e de modo semi-automatizado nos verbetes, a partir da identificação no texto de sequências de caracteres que atuam como delimitadores (e.g.

Tabela 1. Amostra de verbetes segmentados segundo a análise posicional.

FONTE	LEMA	PÓS-LEM.	DEFINIÇÃO	PÓS-DEF.
[Cardoso 1570]	Dicto, as	frequẽtatiuo	Dizer	a dico, is
[Velez 1744]	Exolvo, is	ab Ex & Solvo	Desatar, expedir, livrar	aliquem aliqua re
[Fonseca 1798]	Area, ae	f. Hor.	Área, chão de edificio	

marcas de pontuação, formas desinenciais, termos e expressões metalinguísticas). Na ausência de uma lista prévia dessas marcas, sua identificação ocorre durante a checagem manual dos resultados, a cada passo do processo.

3. Representação da informação

A representação da informação é uma das quatro atividades que constituem o processo informacional – junto com descrição, análise e classificação; de caráter investigativo, e não apenas objetivo ou descritivo, envolve uma análise conceitual prévia cujo resultado deve ser traduzido, isto é, convertido em determinado conjunto de termos de indexação para fins de recuperação da informação [Andrade and Neves 2017, p.103]. O potencial da representação de contribuir com o desenvolvimento científico através da organização e produção de conhecimento, depende do estabelecimento de padrões comuns de criação, anotação e compartilhamento de recursos. Um conceito-chave nesse sentido é o de *interoperabilidade*, definido como "dimensão da capacidade de diferentes sistemas, organizações e/ou indivíduos para trabalhar em conjunto em busca de um objetivo comum" [Ide and Pustejovsky 2010, s.p.]. Dois tipos de interoperabilidade se distinguem no âmbito dos sistemas computacionais: a sintática e a semântica. A interoperabilidade sintática se baseia em especificações de formatos de dados e protocolos de comunicação que garantem a intercomunicação e a troca de dados, porém sem garantias de consenso sobre a interpretação desses dados. A interoperabilidade semântica, por sua vez, resulta do uso de um modelo comum de referência para a troca de informações, que permite que dois sistemas sejam capazes de interpretar automaticamente e de maneira precisa a informação compartilhada. Na base desse tipo de modelo está a definição de um vocabulário comum para os termos de indexação, isto é, uma ontologia.

Em se tratando da representação de informação lexical, o modelo *Lexicon Model for Ontologies* ou, simplesmente, *OntoLex-lemon* [Cimiano et al. 2016] tem sido reconhecido como padrão *de facto* para a modelização de recursos lexicais como *Linked Data* [Mambrini et al. 2021, Tiberius et al. 2021]. O propósito do modelo *OntoLex-lemon* é prover os principais elementos necessários para a construção de léxicos computacionais. Quando combinado com elementos de outras ontologias, permite construir representações padronizadas e ao mesmo tempo adaptadas aos dados a serem representados, sem perda de interoperabilidade. O modelo é formado por módulos que recobrem diferentes aspectos da informação lexical. O módulo fundamental, denominado *ontolex*, fornece instrumentos para representar a estrutura básica de uma entrada lexical (Figura 1). Para a representação da informação extraída dos dicionários de latim, foram selecionadas desse módulo três classes de objetos: LexicalEntry, LexicalSense e Form. A classe LexicalEntry[1] representa a unidade de análise do léxico, definida como "conjunto de formas gramaticalmente

[1] `http://www.w3.org/ns/lemon/ontolex#LexicalEntry`.

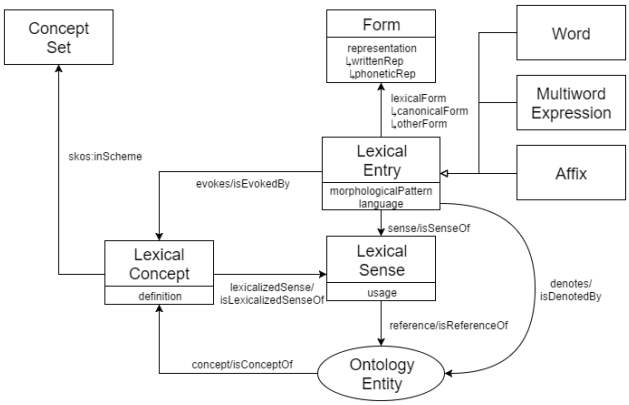

Figura 1. Diagrama representativo do módulo *ontolex* [Cimiano et al. 2016].

relacionadas e de significados associados a todas essas formas"; por formas gramaticalmente relacionadas se entende a presença simultânea de três características: (1) pertencer à mesma classe de palavras; (2) seguir o mesmo paradigma flexional; e (3) possuir a mesma etimologia. Os indivíduos da classe LexicalEntry são, portanto, entradas correspondentes a itens lexicais unívocos. A classe Form[2] serve para representar cada uma das formas associadas a um certo item lexical; as formas são normalmente associadas a uma representação escrita, veiculada no modelo *ontolex* como um valor literal da propriedade *writtenRepresentation*.[3] A classe LexicalSense,[4] por sua vez, representa um significado linguístico de um item lexical em referência a certo elemento de uma ontologia; em outras palavras, trata-se da forma lexicalizada que um conceito recebe em determinada língua, usualmente representada pela propriedade *definition*,[5] emprestada à ontologia SKOS. A classe LexicalSense pode abrigar ainda certas propriedades adicionais, relativas a condições específicas sob as quais a associação do significado com o item lexical é considerada válida, como contexto, registro, domínio, especificações formais como número, tempo verbal, entre outras. A associação entre os indivíduos das três classes selecionadas, por fim, é realizada através de propriedades específicas: a propriedade *lexicalForm* conecta uma entrada lexical (i.e. um indivíduo da classe LexicalEntry) às suas respectivas formas associadas (i.e. indivíduos da classe Form), ao passo que a propriedade *sense* conecta as entradas às suas respectivas acepções (i.e. indivíduos da classe LexicalSense).

A tríade formada pelas classes LexicalEntry, Form e LexicalSense é suficiente para modelar um léxico em seus elementos fundamentais. Porém, o vocabulário fechado da ontologias traz dificuldades para a inclusão das informações que se apresentam nas posições pós-lemática e pós-definicional, especialmente quando há interesse na preservação *ipsis litteris* do texto original dos dicionários para a pesquisa histórica. Uma estratégia para contornar esse problema é utilizar propriedades compatíveis com a representação de informações textuais, particularmente duas propriedades *note*, igualmente genéricas, pertencentes a diferentes ontologias (*Lexinfo*[6] e SKOS[7]), selecionadas para representar respectivamente

[2] `http://www.w3.org/ns/lemon/ontolex#Form`.
[3] `http://www.w3.org/ns/lemon/ontolex#writtenRep`.
[4] `http://www.w3.org/ns/lemon/ontolex#LexicalSense`.
[5] Cf. `http://www.w3.org/2004/02/skos/core#definition`.
[6] `http://www.lexinfo.net/ontology/3.0/lexinfo#note`.
[7] `http://www.w3.org/2004/02/skos/core#note`.

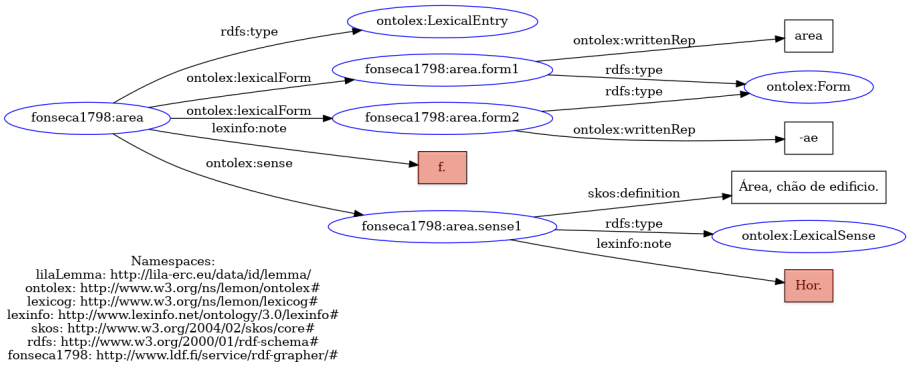

Figura 2. Visualização da representação RDF de parte da entrada lexical *Area, ae* [Fonseca 1798].

as posições pós-lemática e pós-definicional. Sua integração ao modelo *Ontolex-Lemon* permite incluir integralmente a informação metalinguística suplementar ao lema e à definição, bem como associar essa informação a um dos elementos principais através de sua atribuição à respectiva classe. A Figura 2 traz a visualização gráfica da parte inicial da entrada lexical *Area, ae* [Fonseca 1798] representada segundo o modelo *Ontolex*, com destaque para a informação metalinguística; nota-se que, se no original as abreviaturas se apresentam conjugadas ("f. Hor."), na versão modelizada cada uma está associada ao seu respectivo elemento, como valor da propriedade *lexinfo:note* – a saber, a informação gramatical 'f.' à classe LexicalEntry, a informação de autoria 'Hor.' à classe LexicalSense.

4. Mapeamento e interligação das entradas

A perspectiva de prover interoperabilidade para o conteúdo de dicionários latim-português requer, além de sua modelização, a interligação desse conteúdo entre si e com outros recursos compartilhados. Em se tratando de recursos voltados para o latim, a LiLa Knowledge Base organiza-se como uma base de dados interoperável de recursos linguísticos que tem como espinha dorsal uma coleção de lemas denominada Lila Lemma Bank [Mambrini and Passarotti 2023]. Na medida em que a lematização é uma camada de anotação compartilhada por recursos lexicais e textuais, entradas lexicais e *tokens* podem ser interligados através dos lemas, pela atribuição de um identificador único comum (URI) [Passarotti et al. 2020]. Logo, para publicar um novo recurso lexical na LiLa Knowledge Base, é preciso mapear as formas de entrada dos dicionários com as formas correspondentes no Lemma Bank, de modo a obter os respectivos URI.

O pareamento é feito por um processo semi-automatizado de *string matching* com base em uma abordagem progressiva em três etapas: (a) compara-se uma sequência formada pelo lema e a classe gramatical; os resultados são classificados em pares unívocos (1:1), pares ambíguos (1:N) ou sem pareamento (1:0); (b) uma segunda rodada de comparação é realizada sobre os não pareados (1:0) na etapa anterior, desta vez comparando apenas o lema; os resultados obtidos são classificados como no passo anterior; (c) para os lemas não pareados (1:0) remanescentes é feito um levantamento de possíveis candidatos por meio de cálculo de similaridade (*edit distance*). O trabalho de desambiguação dos pares ambíguos e candidatos a pareamento é feito manualmente, o que demanda um tempo considerável mas traz vantagens para ambos os recursos envolvidos: para o léxico, possibilita a detecção e eliminação de erros tipográficos da fonte; para o Lemma Bank,

Tabela 2. Resultados quantitativos do pareamento com o LiLa Lemma Bank.

	Cardoso		Velez		Fonseca	
lemas totais	27.752	100%	4.723	100%	32.723	100%
pareamento unívoco (1:1)	18.467	67%	4.093	87%	28.546	87%
pareamento ambíguo (1:N)	893	3%	368	8%	1.096	3%
sem pareamento (1:0)	8.392	30%	262	6%	3.081	9%
1:0 com um candidato	789	3%	151	3%	1.903	6%
1:0 com muitos candidatos	1.153	4%	104	2%	1.088	3%
1:0 sem candidatos	6.450	23%	7	< 1%	90	< 1%
novos lemas	1.467	5%	140	3%	1.886	6%
novas grafias	656	2%	60	1%	706	2%
erros tipográficos	153	< 1%	30	< 1%	489	1%

fornece novas entradas e grafias alternativas ainda não contempladas pela base. Os URI obtidos são então atribuídos às respectivas entradas lexicais por meio da propriedade *canonicalForm*. Um panorama quantitativo dos resultados do processo de pareamento dos três dicionários é fornecido na Tabela 2. Po fim, o conjunto de dados de cada dicionário é organizado em um arquivo Turtle RDF,[8] e inserido na LiLa Knowledge Base, agregando-se a todo o banco de recursos linguísticos que constitui essa base de dados.

5. Conclusão

A coleção de dicionários latim-português em formato LOD está disponível na página do projeto LiLa;[9] seu uso para pesquisa se dá através de um terminal SPARQL[10] e em dois repositórios em que está disponível para download.[11]

O modo de estruturar e representar dados lexicográficos discutido neste artigo, que combina análise posicional e modelo *Ontolex-Lemon*, foi aplicado com sucesso ao conteúdo dos três dicionários latim-português selecionados; um quarto dicionário, este do século XX, está em fase de finalização e trará definições e equivalentes em português contemporâneo. O método promete um alto potencial de reuso, ligado tanto a um grau de generalização suficiente para se adaptar a qualquer texto dicionarístico, antigo ou moderno; também serve de preparação para a representação individualizada das informações por meio de ontologias específicas (e.g. catálogos de autoria para nomes de escritores, descritores morfológicos para informações gramaticais, etc.). Da forma como está, o uso do recurso por latinistas, linguistas e lexicógrafos depende do conhecimento de linguagem SPARQL. Esse acesso poderá ser facilitado com a produção de *scripts* prontos para uso.

6. Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

[8] cf. `https://www.w3.org/TR/turtle/`.

[9] `https://lila-erc.eu/data-page/`.

[10] `https://lila-erc.eu/sparql/`.

[11] `https://github.com/CIRCSE/Latin-Portuguese-dictionaries` e `https://github.com/lucascdz/latinolusitanicum-db`.

Referências

Andrade, W. O. and Neves, D. A. B. (2017). Análise documental e representação da informação. In Fujita, M. S. L., de Brito Neves, D. A., and Dal'Evedove, P. R., editors, *Leitura documentária*, pages 93–112. Oficina Universitária/Cultura Acadêmica, Marília/São Paulo.

Berners-Lee, T. (2006). Linked data. <https://www.w3.org/designissues/linkeddata.html>.

Cardoso, J. (1570). *Dictionarium latino lusitanicum & vice versa lusitanico latinu[m]*. João de Barreira, Coimbra.

Cimiano, P., McCrae, J. P., and Buitelaar, P. (2016). Lexicon model for ontologies: Final community group report. Technical report, Ontology-Lexicon Community Group under the World Wide Web Consortium (W3C), Cambridge, MA.

Fonseca, P. J. (1798). *Parvum lexicum latinum lusitana interpretatione adjecta*. Typographia Regia, Lisboa.

Hanks, P. (2022). Lexicography. In Mitkov, R., editor, *The Oxford Handbook of Computational Linguistics*. Oxford University Press, Oxford.

Hartmann, R. R. K. (2001). *Structural and typological perspectives*, pages 57–79. Routledge, London/New York, 2nd. edition.

Ide, N. and Pustejovsky, J. (2010). What does interoperability mean, anyway. In *Proceedings of the Second International Conference on Global Interoperability for Language Resources (ICGL 2010)*, Hong Kong.

Mambrini, F., Litta, E., Passarotti, M., and Ruffolo, P. (2021). Linking the Lewis & Short dictionary to the LiLa Knowledge Base of interoperable linguistic resources for Latin. In Fersini, E., Passarotti, M., and Patti, V., editors, *Proceedings of the Eighth Italian Conference on Computational Linguistics*, AIxIA Series, Aachen, Germany. CEUR Workshop Proceedings.

Mambrini, F. and Passarotti, M. C. (2023). The lila lemma bank: A knowledge base of latin canonical forms. *Journal of Open Humanities Data*, 9(1).

Merrilees, B. (1996). The shape of the medieval dictionary entry. *Digital Studies/le Champ Numérique*, 4.

Passarotti, M., Mambrini, F., Franzini, G., Cecchini, F. M., Litta, E., Moretti, G., Ruffolo, P., and Sprugnoli, R. (2020). Interlinking through lemmas. the lexical collection of the LiLa Knowledge Base of linguistic resources for Latin. *Studi e Saggi Linguistici (SSL)*, 58(1):177–212.

Steurs, F., Schoonheim, T., Heylen, K., and Vandeghinste, V. (2020). The future of academic lexicography—a white paper.

Tarp, S. (2019). La ventana al futuro: despidiéndose de los diccionarios para abrazar la lexicografía. *RILEX. Revista sobre investigaciones léxicas*, 2(2):5–36.

Tiberius, C., Krek, S., Depuydt, K., Gantar, P., Kallas, J., Kosem, I., and Rundell, M. (2021). Towards the ELEXIS data model: defining a common vocabulary for lexicographic resources. In Kosem, I. and Cukr, M., editors, *Electronic lexicography in the*

21st century: Post-editing lexicography, Brno, Czech Republic. Lexical Computing CZ.

Velez, A. (1744). Index totius artis. In *Emmanuelis Alvari S. J. De Institutione Grammatica Libri Tres*, pages 366–661. Typographia Academiae, Eborae.

Verdelho, T. S. and Silvestre, J. P. M. (2002). Corpus lexicográfico do português.

Welker, H. A. (2004). *Dicionários: uma pequena introdução à lexicografia*. Thesaurus Editora, Brasília.