laborhistórico

Revista LaborHistórico
ISSN 2359-6910
v.9, n.1, e52387, 2023
DOI: 10.24206/lh.v9i1e52387

Artigo

A elaboração de um dicionário terminológico histórico com recursos digitais

Bruno Maroneze[1,2]
E-mail: maronezebruno@yahoo.com.br

Graça Rio-Torto[2]
E-mail: gracart@gmail.com

[1] Universidade Federal da Grande Dourados, Dourados, Mato Grosso do Sul, Brasil.
[2] Universidade de Coimbra, Faculdade de Letras, Centro de Estudos de Linguística Geral e Aplicada (CELGA-ILTEC), Coimbra, Portugal.

Resumo

O presente artigo visa descrever a elaboração de um projeto em Humanidades Digitais intitulado *Dicionário Histórico de Termos da Biologia*. Foi elaborada uma versão "piloto", contando com 50 (cinquenta) termos, extraídos da obra *Diccionario dos Termos Technicos de Historia Natural*, de Domingos Vandelli, publicada em 1788. Após selecionados os termos, foram elaborados as definições e os textos das informações histórico-etimológicas para cada verbete. Em seguida, construiu-se uma estrutura computacional que permite a consulta *online* e a atualização de maneira eficaz. Por fim, são apresentadas as perspectivas futuras de ampliação e contínua atualização da obra. São também disponibilizados os endereços eletrônicos para acesso ao dicionário, ao código-fonte e aos dados.

Palavras-chave

Lexicografia, Terminologia, Lexicografia computacional, Lexicografia Diacrônica, Humanidades Digitais.

Recebido: 12/05/2022
Aceito: 12/04/2023

Como citar:
MARONEZE, Bruno; RIO-TORTO, Graça. A elaboração de um dicionário terminológico histórico com recursos digitais. *Revista LaborHistórico*, v.9, n.1, e52387, 2023. doi: https://doi.org/10.24206/lh.v9i1e52387

Abstract

This article aims to describe the elaboration process of a Digital Humanities project entitled *Dicionário Histórico de Termos da Biologia* (Historical Dictionary of Biology Terms). Initially, we made a pilot version of the dictionary, containing fifty terms, extracted from the work *Diccionario dos Termos Technicos de Historia Natural* (Dictionary of technical terms of Natural History), by Domingos Vandelli, published in 1788. After the selection of the terms, we wrote the definitions and the texts of the historical-etymological information for each entry. After that, we built a computational structure that allows online consultation and updating in a straightforward way. We present the future perspectives of enlargement and continuous updating of the dictionary, together with the electronic addresses of the dictionary, the source code and the data.

Keywords

Lexicography, Terminology, Computational Lexicography, Historical Lexicography, Digital Humanities.

Introdução

A elaboração e disponibilização de recursos linguísticos, tanto para a consulta por seres humanos quanto para a criação de ferramentas computacionais, inscreve-se no campo das chamadas Humanidades Digitais. Nesse sentido, o projeto aqui descrito, centrado no dicionário eletrônico intitulado *Dicionário Histórico de Termos da Biologia*, visa compilar e apresentar dados históricos sobre o léxico científico da língua portuguesa. É o resultado de um estágio de pós-doutoramento desenvolvido na Universidade de Coimbra e contou também com a colaboração de estudantes de graduação. A obra *Diccionario dos termos technicos de Historia Natural*, de 1788 de Domingos Vandelli (1735-1816), italiano radicado em Portugal e grande renovador da ciência da sua época, constituiu o *corpus* inicial para a elaboração do "piloto" deste dicionário.

Na seção 1, detalhamos os objetivos gerais do projeto. Nas seções 2 e 3, descrevemos as decisões lexicográficas que foram tomadas na elaboração de um "piloto" (primeira versão) desse dicionário. Na seção 4, encontram-se as decisões computacionais tomadas, bem como uma descrição da estrutura computacional da obra. As perspectivas de ampliação e atualização da obra são descritas na seção 5. Nas considerações finais, apresentamos os endereços eletrônicos para acesso à consulta, aos dados e ao código-fonte do dicionário.

Numa época de inequívoca efervescência das Humanidades Digitais, e da sua enorme mais-valia, nomeadamente em numerosos setores dos estudos sobre a linguagem, a concepção e estruturação do dicionário eletrônico aqui descrito não é alheia a toda a vasta plêiade de questões teóricas que a construção de uma base de dados convoca. Como se apresenta nas seções seguintes, trata-se de uma base de dados assentada num acervo lexicográfico de 1788 (VANDELLI, 1788) que levanta questões prementes de interdisciplinaridade e de fronteira teórica com a natureza mais ou menos terminológica dos itens lexicais que elenca (cf. FUERTES-OLIVERA, 2017; KARPOVA; KARTASHKOVA, 2009, entre outros). Limitamo-nos nesta apresentação do projeto, a aflorar algumas dessas questões que se nos colocam na elaboração e tratamento dos materiais que constarão do dicionário em elaboração.

Objetivos do Dicionário Histórico de Termos da Biologia

O dicionário eletrônico em elaboração, que intitulamos *Dicionário Histórico de Termos da Biologia*, foi concebido com o objetivo de reunir dados que subsidiem estudos interdisciplinares, envolvendo as áreas da Linguística (especialmente os estudos da Morfologia, da Lexicologia e da Terminologia), da História (em particular, a História da Ciência) e das ciências biológicas.

O estudo diacrônico do léxico científico tem como uma de suas principais contribuições o entendimento da dinâmica dos conceitos no decorrer da história. Por meio do estudo de datações e etimologias, é possível identificar a transmissão de determinado conceito através de diferentes culturas, contribuindo assim para compreender a dinâmica da história das ciências. Nas palavras de Gutiérrez Rodilla (2016, p. 117):

> La reconstrucción de la historia del léxico científico es algo insoslayable para poder conocer, por un lado, la historia de los conceptos que maneja la ciencia, la historia de su discurso. Pero, por otro lado, es imprescindible también para completar la historia del léxico de una lengua, que no puede construirse tan solo sobre las palabras que pueblan los textos comunes y literarios. De aquí que, en nuestra opinión, el estudio de la historia del léxico científico debiera tener reservado un lugar fundamental entre las tareas de la historia de la ciencia y la de la lengua.[1]

[1] "A reconstrução da história do léxico científico é algo incontornável para poder reconhecer, por um lado, a história dos conceitos que a ciência emprega, a história de seu discurso. Mas, por outro lado, também é imprescindível para completar a história do léxico de uma língua, que não pode ser construída apenas sobre as palavras que povoam os textos comuns e literários. Por isso,

Nesse sentido, faz-se necessária a convergência entre os estudos de História da Ciência e de História da Língua, para que seja possível

> rastrear la penetración de las distintas doctrinas, calibrar el éxito o fracaso cosechados por un término o grupo de términos ligados a una determinada teoría y la extensión semántica de cada una de esas voces; o comprobar los posibles cambios semánticos sufridos con el paso del tiempo.[2] (GUTIÉRREZ RODILLA, 2016, p. 119)

Para além dos estudos de História da Ciência, o estudo diacrônico do vocabulário científico também traz contribuições para a Morfologia da língua portuguesa, na medida em que os dados sobre os empregos dos termos em épocas históricas anteriores jogam luz na história de seus elementos formadores (prefixos, sufixos, elementos de composição, tais como *-logia*, *-grafia*, *mega-*, *pseudo-* etc.); assim, é possível identificar os sentidos e as condições de uso desses elementos ao longo do tempo.

O estudo diacrônico do vocabulário científico feito sob o enfoque da Terminologia tem sido denominado de Terminologia Diacrônica. Curti-Contessoto (2022) entende que, nesse âmbito, o que caracteriza a diacronia é "o estudo das transformações que acontecem nos conceitos e nos termos" (CURTI-CONTESSOTO, 2022, p. 116). Assim, é possível observar as transformações tanto no nível dos conceitos (o "significado") quanto no nível dos termos (o "significante"). Um dos objetivos do *Dicionário Histórico de Termos da Biologia* é poder embasar esses estudos, apresentando dados concretos que mostram essas transformações. Um exemplo é o emprego do termo "inseto" para se referir a aranhas, camarões e caranguejos, que é o uso desse termo no texto do século XVIII analisado (cf. seção 2, a seguir). Atualmente, o conceito de "inseto" é mais restrito, referindo-se apenas a um grupo de artrópodes, tendo havido uma mudança no conceito. Ao mesmo tempo, o conceito que era denominado pelo termo "inseto" é hoje denominado pelo termo "artrópode", tendo havido, assim, também uma mudança de significante.

Com esses objetivos em mente, procedeu-se à delimitação do público-alvo, da macroestrutura e da microestrutura, conceitos tradicionais da Lexicografia (cf., por exemplo, Welker, 2004). Definiu-se o público-alvo como formado por profissionais das áreas de História da Ciência e História da Língua Portuguesa; secundariamente, por professores de Biologia e por interessados em etimologia. Essa decisão é importante

---

em nossa opinião, o estudo da história do léxico científico deve ter um lugar reservado entre as tarefas da história da ciência e da história da língua." (tradução nossa)

[2] "Rastrear a penetração das distintas doutrinas, calibrar o êxito ou o fracasso colhidos por um termo ou um grupo de termos ligados a uma determinada teoria e a extensão semântica de cada uma dessas vozes; ou comprovar as possíveis mudanças semânticas sofridas com o passar do tempo." (tradução nossa)

porque as definições serão redigidas não com o objetivo de serem cientificamente precisas em relação ao conhecimento científico contemporâneo, mas sim levando em consideração os conhecimentos que se tinha em momentos históricos anteriores, bem como as mudanças ocorridas. A macroestrutura e a microestrutura do dicionário são descritas, respectivamente, nas seções 2 e 3, a seguir.

A seleção do corpus e a elaboração da macroestrutura

A prática lexicográfica atual (WELKER, 2004, p. 87-91) recomenda que os dicionários sejam baseados em *corpus*, tanto na seleção das entradas (macroestrutura) quanto na apresentação de exemplos e outras informações (microestrutura). Assim, para a delimitação da macroestrutura da versão "piloto" (um rascunho) do dicionário, efetuou-se primeiro a seleção do *corpus*.

Escolheu-se como ponto de partida o século XVIII, pelas seguintes razões: a) nesse período, ocorreu uma aceleração na produção científica da Europa (e, portanto, também do mundo luso-brasileiro), impulsionada por diversos avanços filosóficos, científicos e tecnológicos; b) antes desse período, a produção científica era escrita sobretudo em latim, sendo, portanto, difícil encontrar textos em língua portuguesa que pudessem compor o *corpus* do dicionário; c) muitas obras científicas em português publicadas nesse período são facilmente encontráveis em repositórios *online* e bibliotecas virtuais; d) evitou-se trabalhar com textos de períodos muito recuados devido ao fato de as ferramentas computacionais disponíveis para análise de *corpus* serem voltadas para a língua portuguesa contemporânea, apresentando muitos erros com textos muito antigos (em especial anteriores ao século XVII).

A área das ciências biológicas foi uma área de grande desenvolvimento no século XVIII, especialmente devido à influência do célebre autor sueco Carl von Linné (Lineu, em português), que revolucionou o pensamento científico nessa área; particularmente no mundo luso-brasileiro, o mais importante divulgador das ideias de Lineu foi Domingos Vandelli (1735-1816), italiano radicado em Portugal[3], e cuja obra *Diccionario dos termos technicos de Historia Natural*, de 1788 (cf. CABRAL, 2018; MARONEZE, 2019; PEREIRA, 2017) constituiu o *corpus* inicial para a elaboração do "piloto" deste dicionário.

Dessa forma, definido o *corpus* inicial (a obra de Vandelli de 1788), passou-se à transcrição do texto do formato PDF (disponível no portal Google Livros) para o

[3] No âmbito da reforma pombalina da Universidade de Coimbra, Vandelli foi responsável pela implantação do Jardim Botânico, do Laboratório Químico e do Museu de História Natural da Universidade de Coimbra. Foi membro de várias academias científicas, tendo participado ativamente na criação da Real Academia das Ciências de Lisboa.

formato TXT (que é um formato legível por qualquer *software*). Simultaneamente, passou-se também à seleção de uma amostra de 50 (cinquenta) termos, que viriam a se constituir nos primeiros verbetes do dicionário. Os termos foram selecionados com o objetivo de se constituírem numa amostra diversificada, com seguintes critérios: a) os termos foram selecionados apenas do domínio da Botânica, visto ser essa a principal especialidade de Vandelli e a área que, aparentemente, recebeu mais atenção na sua obra; b) procurou-se selecionar tanto substantivos (*antera*, *bráctea* etc.) como adjetivos (*bífido*, *conivente*, *deflexo* etc.); b) procurou-se selecionar tanto termos simples (*gema*, *gomo*, *estame* etc.) quanto morfologicamente complexos (*bulboso*, *foliáceo* etc.), com a inclusão de um termo formado por mais de uma palavra (*jardim botânico*); c) por fim, procurou-se diversificar as origens, incluindo termos de origem grega (*antera*, *epiderme*, *hermafrodita* etc.), latina (*crena*, *cutícula*, *estípula* etc.) e também termos de longa existência na língua portuguesa (*disco*, *gema*, *pimpolho* etc.).

A microestrutura dos verbetes

Em relação à microestrutura, definiu-se que cada verbete do dicionário apresentará a classe gramatical do termo e a sua definição (ou definições, no caso de termos polissêmicos). Como já mencionado na seção 1, por ser uma obra de caráter histórico, a definição apresentada deve ser redigida conforme o significado e o emprego do termo na época e na obra em questão, ainda que seja diferente do significado atual.

Tendo em vista que o objetivo principal da obra é apresentar informações históricas sobre cada termo, definiu-se que cada verbete deve incluir uma "discussão histórico-etimológica", que trará dados sobre a etimologia e eventuais alterações de significado por que o termo tenha passado, bem como um resumo das informações apresentadas em outras obras lexicográficas (eventualmente contestando-as). Além disso, o verbete incluirá variantes ortográficas, visto que, em períodos anteriores à padronização ortográfica do século XX, era comum encontrar diversas grafias diferentes para o mesmo termo, muitas vezes na mesma obra. Por fim, o verbete deve incluir também abonações extraídas do *corpus*, para que o consulente possa observar o emprego do termo nos textos de cada época.

Uma informação muito importante para os estudos históricos é a data da primeira atestação da ocorrência do termo na língua portuguesa, o chamado *terminus a quo* (VIARO, 2011). Numa situação ideal hipotética, em que se disporia de um *corpus* contendo todos os textos sobre Biologia em português, essa informação tornar-se-ia redundante, visto que as abonações já seriam suficientes para mostrar ao consulente as datas mais recuadas no tempo. Porém, na situação real em que se tem um *corpus* mais ou menos limitado, pode ser importante incluir a data da primeira atestação, quando se sabe ser anterior aos textos do *corpus*. Assim, para o "piloto"

deste dicionário, optou-se por incluir no verbete a informação de atestações anteriores à obra de Vandelli, especialmente quando disponíveis em outros artigos ou obras lexicográficas.

Definida a microestrutura do dicionário, passou-se a redigir verbetes para os 50 termos selecionados. Esses verbetes foram inicialmente redigidos usando um editor de textos tradicional (no caso, o *Google Docs*). Como exemplo, apresentam-se os verbetes *jardim botânico* e *pistilo*:

jardim botânico sm.

Definição: Jardim onde se cultivam plantas para fins de estudo, em geral aberto à visitação pública.

Atestações mais antigas (1718: "Jardim Real Botanico"): "O Duque Regente naõ querendo encarregar-se da nomeaçaõ de outro, deyxou a eleyçaõ ao Duque de Maine, & Marechal de Ville-Roy, & Duqueza de Ventadour, os quaes nomearaõ a Mons. Dodart, que foy Medico do Duque de Borgonha, & dos Principes seus filhos, & agora o era da Princeza viuva de Conti; & como este emprego tem de renda 54U. libras, reservou S. A. Real para o seu Medico Mons. Chivax a direcçaõ do Jardim Real Botanico, ou das plantas medicinaes exquisitas, que rende seis mil florins. (Extraído de: Gazeta de Lisboa Occidental, n. 19, maio de 1718, p. 151. Disponível em: http://hemerotecadigital.cm-lisboa.pt/Periodicos/GazetadeLisboa/1718/Maio/Maio_item1/P15.html)

(1735): "*A morte dos enemigos de* Arcagathus*, a inefficacia dos* encantos magicos*, as honras que havia feito á Faculdade* Attalus*, o ultimo Rey de* Pergamus*, o qual fez a o* Povo Romano *seu herdeiro, e foy tam grande promotor da Sciencia Medica, que chegou a cultivar hum* jardim botanico *em seu proprio palacio, para fazer experiencias nos condenados por criminozos, para beneficio do resto de seus Vassalos* (Extraído de: Jacob de Castro Sarmento, Materia Medica Physico-Historico-Mechanica, Reyno Mineral, 1735, p. xi. Disponível em: https://books.google.pt/books?id=B1fi4PvVN8wC)

(1788): "O primeiro conhecimento adquire-se com o estudo da botanica, o segundo com experiencias e reflexões fisicas, o terceiro, e quarto com hum jardim botanico, no qual he necessario cultivar os vegetaes de todos os climas, e terrenos." (Extraído de: Diccionario dos termos technicos de Historia Natural, 1788, p. 294)

Discussão histórico-etimológica: O Dicionário Houaiss (HOUAISS; VILLAR, s/d) data essa expressão do ano de 1852, mas não informa a fonte. A expressão é bem mais antiga, como mostra o contexto de 1735. Em 1718, a expressão que aparece é "Jardim Real Botanico", com o elemento "Real" intercalado, o que revela que a expressão ainda estava em vias de se consolidar na forma que tem nos dias atuais.

É possível que essa expressão seja um decalque de uma expressão semelhante de outra língua europeia. O francês *jardin botanique* já aparece em 1673 ("Recherche des Antiquités et Curiosités de la ville de Lyon", disponível em https://books.google.pt/books?id=btFTAAAAcAAJ), ainda que a data indicada pelo Trésor de la Langue Française seja 1732. O latim *hortus botanicus* é ainda mais antigo, aparecendo na obra "Critica Sacra", de Edward Legh, 1639 (disponível em https://books.google.pt/books?id=0IRmAAAAcAAJ). Dessa forma, é razoável supor que a expressão portuguesa tenha sido uma tradução de uma expressão equivalente em outra língua.

pistilo sm.

Definição: Parte da flor, em geral entre as anteras, pela qual entra o pólen para a fecundação.

Atestações mais antigas (1782): CUNHA, Antônio Geraldo da. Dicionário etimológico Nova Fronteira da língua portuguesa. Rio de Janeiro, 1982. 2. ed., 1986. [informado pelo Dicionário Houaiss (HOUAISS; VILLAR, s/d)]

(1788): "1. As ordens *Polygamia æqualis* consta de muitas flores pequenas com estames, e pistillos." (Extraído de: Diccionario dos termos technicos de Historia Natural, 1788, p. 191)

"133. *Pistillum*. *fig*. 143. 149. *d*. 150. *a*. 152. He huma parte da flor posta o meio, pela qual entra o *pollen*, ou a aura seminal no *germe*, ou ovario para a fecundaçaõ; està cercado dos *filamentos*, e está posto geralmente entre as *antheras*, e consta de *germe*, *stilo* e *estigma*.

Os pistilos saõ differentes." (Extraído de: Diccionario dos termos technicos de Historia Natural, 1788, p. 267-8)

Formas variantes: pistillo (Vandelli)

Discussão histórico-etimológica: O Dicionário Houaiss (HOUAISS; VILLAR, s/d) informa que a palavra derivaria do "lat. *pistīllum* ou *pistīllus,i* no sentido de 'mão de pilão'"; no entanto, é pouco provável que seja uma palavra herdada, visto que a datação é tardia. Assim, não se pode falar que o sentido latino de "mão de pilão" tenha se transformado no sentido de "parte da flor" em português. O emprego da forma latina *pistillum* no latim científico (como atesta o dicionário de Vandelli) deixa claro que o étimo da forma portuguesa é o latim científico, já no sentido corrente de "parte da flor", e a alteração de sentido se deve a um emprego metafórico já ocorrido no latim científico.

A estrutura computacional do dicionário

Com os verbetes já redigidos, procurou-se uma forma de apresentá-los num formato *online*, que ao mesmo tempo possibilitasse a atualização de maneira simples, sem que fosse necessário reconstruir todo o dicionário. Para essa etapa, por sugestão da pesquisadora Ligeia Lugli (comunicação pessoal), foi empregada a linguagem de programação R (https://www.r-project.org/), que dispõe de algumas bibliotecas[4] para o trabalho com *corpus* e para a construção de *websites*. Sendo assim, foram empregadas as seguintes bibliotecas:

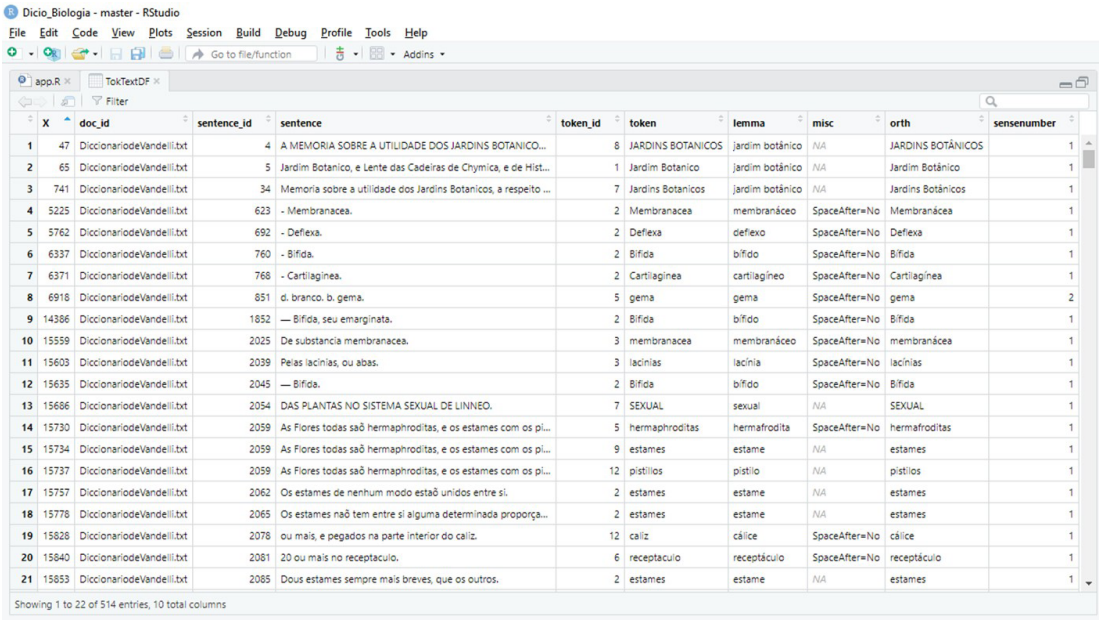
a. Shiny (https://cran.r-project.org/web/packages/shiny): permite a criação de uma interface *online* (na forma de um *website*) que apresenta dados a partir de decisões tomadas pelo usuário. Assim, o usuário pode, por exemplo, selecionar uma entrada do dicionário e os comandos do Shiny possibilitam a apresentação do verbete completo na tela;

[4] Bibliotecas de linguagens de programação são conjuntos de comandos já prontos, que facilitam o desempenho de certas tarefas. São criadas por programadores que percebem a necessidade de resolver um problema específico recorrente e, assim, permitem que os demais programadores não tenham que resolver esse mesmo problema novamente.

b. Stringr (https://cran.r-project.org/web/packages/stringr): contém diversos comandos destinados à manipulação de textos, como comandos para inserção de trechos num texto, localização de trechos etc.;

c. Udpipe (https://cran.r-project.org/web/packages/udpipe): é uma biblioteca voltada especificamente para análise de *corpus*. Apresenta, entre outras funcionalidades, etiquetadores morfológicos e sintáticos e recursos para lematização.

Com o emprego desses recursos computacionais, escreveu-se um código em R que gera um arquivo em formato de tabela com as informações extraídas do *corpus*, conforme se observa na figura 1 a seguir:


Dicio_Biologia - master - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

app.R × TokTextDF ×

Filter

	X	doc_id	sentence_id	sentence	token_id	token	lemma	misc	orth	sensenumber
1	47	DiccionariodeVandelli.txt	4	A MEMORIA SOBRE A UTILIDADE DOS JARDINS BOTANICO...	8	JARDINS BOTANICOS	jardim botânico	NA	JARDINS BOTÂNICOS	1
2	65	DiccionariodeVandelli.txt	5	Jardim Botanico, e Lente das Cadeiras de Chymica, e de Hist...	1	Jardim Botanico	jardim botânico	NA	Jardim Botânico	1
3	741	DiccionariodeVandelli.txt	34	Memoria sobre a utilidade dos Jardins Botanicos, a respeito ...	7	Jardins Botanicos	jardim botânico	NA	Jardins Botânicos	1
4	5225	DiccionariodeVandelli.txt	623	- Membranacea.	2	Membranacea	membranáceo	SpaceAfter=No	Membranácea	1
5	5762	DiccionariodeVandelli.txt	692	- Deflexa.	2	Deflexa	deflexo	SpaceAfter=No	Deflexa	1
6	6337	DiccionariodeVandelli.txt	760	- Bifida.	2	Bifida	bífido	SpaceAfter=No	Bífida	1
7	6371	DiccionariodeVandelli.txt	768	- Cartilaginea.	2	Cartilaginea	cartilagíneo	SpaceAfter=No	Cartilagínea	1
8	6918	DiccionariodeVandelli.txt	851	d. branco. b. gema.	5	gema	gema	SpaceAfter=No	gema	2
9	14386	DiccionariodeVandelli.txt	1852	— Bifida, seu emarginata.	2	Bifida	bífido	SpaceAfter=No	Bífida	1
10	15559	DiccionariodeVandelli.txt	2025	De substancia membranacea.	3	membranacea	membranáceo	SpaceAfter=No	membranácea	1
11	15603	DiccionariodeVandelli.txt	2039	Pelas lacinias, ou abas.	3	lacinias	lacínia	SpaceAfter=No	lacínias	1
12	15635	DiccionariodeVandelli.txt	2045	— Bifida.	2	Bifida	bífido	SpaceAfter=No	Bífida	1
13	15686	DiccionariodeVandelli.txt	2054	DAS PLANTAS NO SISTEMA SEXUAL DE LINNEO.	7	SEXUAL	sexual	NA	SEXUAL	1
14	15730	DiccionariodeVandelli.txt	2059	As Flores todas saõ hermaphroditas, e os estames com os pi...	5	hermaphroditas	hermafrodita	SpaceAfter=No	hermafroditas	1
15	15734	DiccionariodeVandelli.txt	2059	As Flores todas saõ hermaphroditas, e os estames com os pi...	9	estames	estame	NA	estames	1
16	15737	DiccionariodeVandelli.txt	2059	As Flores todas saõ hermaphroditas, e os estames com os pi...	12	pistilos	pistilo	NA	pistilos	1
17	15757	DiccionariodeVandelli.txt	2062	Os estames de nenhum modo estaõ unidos entre si.	2	estames	estame	NA	estames	1
18	15778	DiccionariodeVandelli.txt	2065	Os estames naõ tem entre si alguma determinada proporça...	2	estames	estame	NA	estames	1
19	15828	DiccionariodeVandelli.txt	2078	ou mais, e pegados na parte interior do caliz.	12	caliz	cálice	SpaceAfter=No	cálice	1
20	15840	DiccionariodeVandelli.txt	2081	20 ou mais no receptaculo.	6	receptaculo	receptáculo	SpaceAfter=No	receptáculo	1
21	15853	DiccionariodeVandelli.txt	2085	Dous estames sempre mais breves, que os outros.	2	estames	estame	NA	estames	1

Showing 1 to 22 of 514 entries, 10 total columns

Figura 1 – Corpus etiquetado em formato de tabela
Fonte: elaborado pelos autores

Essa tabela (aqui denominada "tabela do *corpus* etiquetado"), formulada a partir do texto que compõe o *corpus* (o *Diccionario...* de Vandelli), contém informações para cada um dos termos selecionados para o "piloto" do dicionário: o texto do *corpus* onde ocorre o termo (coluna *doc_id*), a sentença onde o termo ocorre (coluna *sentence*), o termo grafado conforme ocorre no *corpus* (coluna *token*), a forma lematizada do termo (coluna *lemma*), a ocorrência do termo na ortografia padronizada atual (coluna *orth*) e o número da acepção relacionada à ocorrência do termo (coluna *sensenumber*). A biblioteca *Udpipe* faz a segmentação das sentenças automaticamente e separa cada uma das ocorrências do *corpus* em uma linha separada da tabela (processo

computacional chamado de *tokenização*). A inclusão do lema, da forma na ortografia atual e do número de cada acepção[5] foi feita manualmente.

Além dessa tabela, foram criadas outras três:

a. Uma tabela contendo os dados dos verbetes propriamente ditos (aqui denominada "tabela principal"): entrada, classe gramatical, variantes, discussão histórico-etimológica, nome do autor do verbete;

b. Uma tabela contendo as definições[6] (aqui denominada "tabela de definições");

c. Uma tabela contendo informações sobre cada obra do *corpus* (autor, título, ano de publicação) (aqui denominada "tabela de obras").

A etapa seguinte foi a elaboração da interface gráfica propriamente dita do dicionário, que contou com a ajuda da pesquisadora Ligeia Lugli e foi baseada em um de seus trabalhos anteriores (Lugli *et al.*, 2019-2022). A figura 2, a seguir, mostra o verbete *cartilagíneo*, com os seus elementos, explicados abaixo:

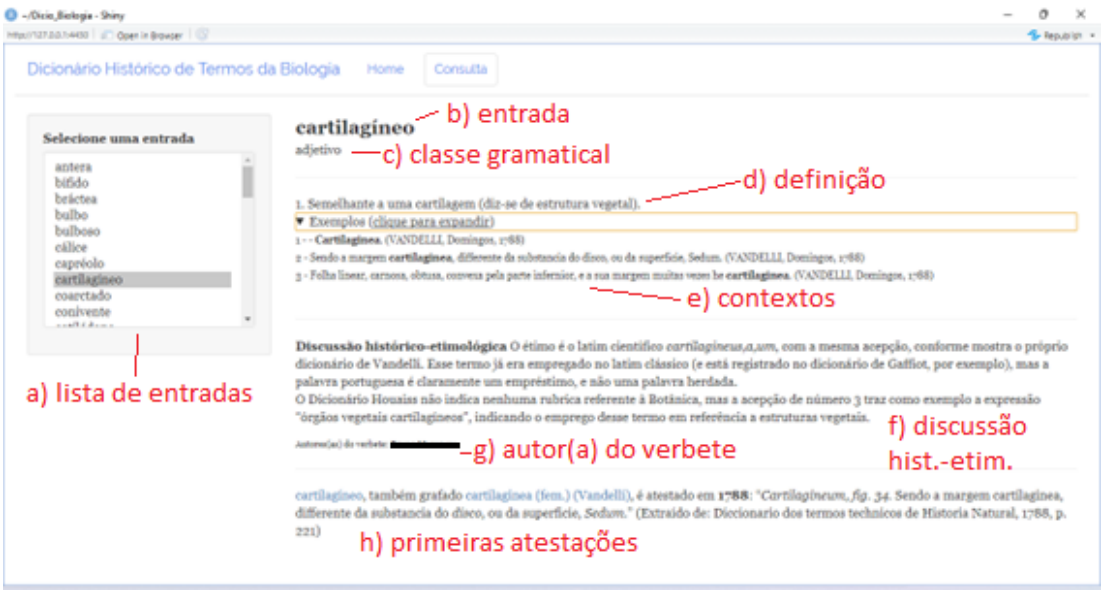


Figura 2 – Interface do dicionário
Fonte: elaborado pelos autores

[5] A identificação das diferentes acepções de um termo é conhecida por "desambiguação" (cf., por exemplo, Nóbrega, 2013). Observe-se que, para o termo "gema", foram identificadas duas acepções: 1. Protuberância no caule ou ramos de uma planta, de onde se originam ramos, folhas ou outras estruturas; 2. Porção interna do ovo das aves. Na figura 1, encontra-se o contexto para a acepção 2.

[6] As definições precisam estar numa tabela separada da dos dados dos verbetes porque, no caso de termos polissêmicos, há mais de uma definição para cada termo, e não é possível dividir apenas uma célula da tabela.

a. A lista de entradas, posicionada à esquerda, é extraída da tabela principal;

b. A entrada, posicionada em destaque no alto da página, é extraída da tabela principal;

c. A classe gramatical, abaixo da entrada, também é extraída da tabela principal;

d. As definições são extraídas da tabela de definições. O *website* identifica automaticamente a quais definições corresponde o verbete selecionado;

e. Os contextos são extraídos da tabela do *corpus* etiquetado. Como podem ocorrer vários contextos para um único termo, optou-se por empregar o recurso da expansão com o clique do usuário, ou seja, o usuário só tem acesso aos contextos após clicar em "Exemplos". No caso de termos polissêmicos, os contextos são separados para cada acepção (cf. a figura 3, abaixo). Após cada contexto, aparecem as informações de autor e data, que são extraídas da tabela de obras;

f. A discussão histórico-etimológica é extraída da tabela principal;

g. O(a) autor(a) do verbete também é extraído da tabela principal;

h. As informações sobre as primeiras atestações, juntamente com as formas variantes ortográficas, também são extraídas da tabela principal. Espera-se que, no futuro, com a expansão do *corpus* para outras obras de períodos anteriores, essas informações deixem de ser necessárias.

Na figura 3, abaixo, no verbete *gema*, observa-se que cada acepção apresenta seus próprios contextos. A primeira acepção ("Protuberância no caule ou ramos de uma planta..." tem quatro ocorrências no *corpus*, enquanto a segunda ("Porção interna do ovo das aves") tem apenas uma.

Figura 3 – Verbete *gema*
Fonte: elaborado pelos autores

Dessa forma, com a interface do dicionário pronta, é possível inserir mais textos no *corpus* e mais verbetes, apenas com a atualização das tabelas, o que pode ser feito a partir de um *software* de planilha eletrônica.

Os "subprodutos" gerados com a elaboração do dicionário

A apresentação dos verbetes em formato de um dicionário eletrônico tem grande importância para o público em geral, que poderá consultar facilmente as informações. Mas a importância da pesquisa não se esgota na divulgação para o público; tão importante quanto o dicionário eletrônico é o acesso aos dados e ao código-fonte, para que possam ser reutilizados em aplicações futuras.

Inicialmente, entende-se que o trabalho de transcrição das obras que compõem o *corpus* deve ser feito de maneira a possibilitar a sua reutilização por parte de outros pesquisadores. Uma transcrição feita automaticamente, por meio de *softwares* do tipo OCR (*optical character recognizer*), costuma trazer muitos erros de reconhecimento, o que é agravado pelo fato de os textos se encontrarem em ortografia e tipografia diferentes das atuais. Portanto, é indispensável a revisão da transcrição por olhares humanos. Essa revisão pode ser feita simultaneamente à divulgação da obra, de modo que, à medida que os textos vão sendo corrigidos, o dicionário vai sendo atualizado (em vez de primeiro corrigir tudo para apenas depois divulgar o trabalho finalizado).

Dessa forma, a transcrição dos textos do *corpus* é entendida como um "subproduto" da elaboração do dicionário, mas que pode ser tão ou mais importante, para alguns (como historiadores da ciência), do que o próprio dicionário. Daí a importância de se empregar, na transcrição, critérios explícitos e bem delimitados, bem como um formato eletrônico que permita a fácil reutilização. Atualmente, o *corpus* está sendo transcrito num formato TXT simples, mas considera-se a possibilidade de empregar futuramente o formato XML no padrão *Text Encoding Initiative*[7] (TEI - https://tei-c.org/), que visa padronizar a etiquetagem de textos para projetos de Humanidades Digitais.

Para além da simples transcrição do *corpus*, outro "subproduto" com potencial para reutilização é a tabela do *corpus* etiquetado, por trazer informações como classe gramatical, lematização e desambiguação. Quaisquer pesquisas futuras desenvolvidas utilizando os mesmos textos não precisarão refazer os mesmos passos já feitos para

[7] De acordo com o próprio *website*, o *Text Encoding Initiative* (TEI) "is a consortium which collectively develops and maintains a standard for the representation of texts in digital form" ("é um consórcio que desenvolve e mantém coletivamente um padrão para a representação de textos em formato digital"). Apresenta diretrizes para vários formatos de texto, como poema, texto teatral, transcrição de fala, manuscritos etc.

este dicionário. Por fim, o próprio código-fonte pode ser facilmente reutilizado para outros projetos similares, como outros dicionários eletrônicos ou outro tipo de disponibilização de dados.

Assim, além do dicionário eletrônico propriamente dito, também o acesso livre ao código-fonte e aos dados foi pensado para que outros pesquisadores possam usá-los em futuras aplicações. Desde que citada a fonte, os pesquisadores poderão reutilizar os dados da forma que lhes for adequada.

O futuro: a ampliação e atualização contínua do dicionário

O *Dicionário Histórico de Termos da Biologia* já se encontra disponível para consulta *online* (https://dicbio.fflch.usp.br/), bem como o seu código-fonte e todas as tabelas que o "alimentam" (https://github.com/brunomaroneze/dicbio).

Para testar o código-fonte, foi posteriormente incluída[8] no *corpus* a obra "Anatomia do corpo humano", de Bernardo Santucci[9] (1739), tendo sido possível verificar que a inserção de um novo texto ocorre de maneira simples e direta; assim, pretende-se dar seguimento ao projeto, com a inclusão de mais textos no *corpus*.

A questão da compilação do *corpus* traz ainda outros dois problemas: o primeiro diz respeito à disponibilidade das obras, tendo em vista que poucas obras científicas em português do século XVIII são de fácil acesso por meios eletrônicos. Dentre elas, destacam-se o *Compendio de Botanica*, de Félix Avelar Brotero (1788); e as *Memorias da Academia Real das Sciencias de Lisboa* (cujo tomo I foi publicado em 1797), ambas disponíveis *online*; mas diversas outras ainda precisam ser identificadas e digitalizadas. Já o segundo problema diz respeito à dificuldade de conversão do texto para o formato TXT, especialmente porque a tipografia e a ortografia da época não são legíveis pela maioria dos *softwares* de reconhecimento óptico de caracteres (OCR), exigindo, portanto, um maior cuidado na transcrição.

Também se faz necessária a ampliação do número de verbetes. Do ponto de vista computacional, essa ampliação também ocorre de forma simples e direta, bastando incluir novos itens nas tabelas; a dificuldade, aqui, se dá em relação à exigência de rigor científico na pesquisa sobre a história e etimologia de cada um dos termos.

---

[8] A inclusão dessa obra foi possível graças ao trabalho de transcrição do texto efetuado pelos estudantes de graduação Amarildo Braga de Oliveira, Letícia Tranquile da Silva, Rafaela Lima Domingos e Raissa Silveira Buss.

[9] Bernardo Santucci foi, segundo se lê na própria folha de rosto de sua obra, médico formado em Bolonha e lente régio da cadeira de Anatomia do Hospital Real de Lisboa. Segundo Fleck e Dillmann (2021), foi contratado pelo poder régio português em 1732 e permaneceu no cargo de lente até 1747.

No entanto, é possível prever que tanto a ampliação do *corpus* quanto a inserção de novos verbetes ocasionem futuramente o problema computacional do uso excessivo de memória, causando lentidão na consulta. Com apenas dois textos e 50 verbetes, esse problema ainda não se coloca; porém, à medida que o dicionário se amplie, talvez seja necessário otimizar o processo computacional.

A etiquetagem do *corpus* feita com o auxílio de ferramentas computacionais não é isenta de erros; em especial, devido aos textos não estarem na ortografia padronizada atual, correções manuais precisam ser feitas. O processo de lematização também foi feito manualmente, bem como a separação das acepções (desambiguação). Para resolver essas questões, pretende-se elaborar uma ferramenta computacional para o trabalho colaborativo, também de acesso *online*, que permita que os diversos membros de uma equipe de trabalho façam as alterações necessárias na tabela. Como já apresentado (seção 5), a tabela pode ser reutilizada em outras aplicações, o que se constitui em mais um motivo para que essa etiquetagem seja feita com rigor.

Com um *corpus* adequadamente etiquetado e lematizado, é possível incluir, na interface do dicionário, a apresentação de dados estatísticos sobre cada termo. Dados como a frequência absoluta, a frequência de cada uma das formas flexionadas ou a frequência de ocorrência em cada texto do *corpus* podem ser implementados no futuro.

A análise minuciosa da obra de Vandelli e, posteriormente, também da obra de Santucci, revelou um problema que não havia sido previsto inicialmente: o grande número de termos, frases e às vezes parágrafos inteiros escritos em latim, mesclados ao texto em português. Trata-se de uma característica presente na imensa maioria dos textos científicos do período, e que certamente ocorrerá com mais intensidade à medida que o *corpus* for ampliado. Do ponto de vista computacional, pode ser interessante que esses trechos em latim sejam etiquetados, de modo a fazer o dicionário ignorá-los quando necessário (em contagens estatísticas, por exemplo). O formato XML no padrão TEI, já mencionado, talvez seja uma solução para essa questão.

Por fim, é importante mencionar que qualquer projeto de longo prazo que envolva recursos computacionais deve estar sempre atento à "evolução natural" da tecnologia. Dessa forma, num futuro próximo ou distante, talvez seja necessário rever inteiramente todo o código-fonte, por exemplo, ou convertê-lo para outra linguagem de programação. Por isso, é importante que tanto o *corpus* quanto os dados do dicionário estejam em formatos facilmente legíveis por qualquer *software*.

Considerações finais

Pretendeu-se, neste artigo, detalhar os procedimentos que foram empregados na elaboração de um projeto em Humanidades Digitais intitulado *Dicionário Histórico de Termos da Biologia*. O projeto constitui-se numa obra de acesso *online* que será continuamente atualizada e enriquecida. Atualmente, a obra contém apenas 50

verbetes, com novos verbetes já sendo elaborados. Neste artigo, foram detalhadas as decisões lexicográficas e computacionais envolvidas na elaboração do "piloto" do dicionário, bem como as propostas de ampliação e atualização.

Como já mencionado, o acesso à obra se dá pelo endereço <https://dicbio.fflch.usp.br/>, e os dados e o código-fonte estão disponíveis no endereço <https://github.com/brunomaroneze/dicbio>. A licença de uso é a Licença Creative Commons BY-NC-AS 4.0 (https://creativecommons.org/licenses/by-nc-sa/4.0/), que permite o compartilhamento e o uso dos dados, desde que devidamente creditados e que não sejam usados para fins comerciais.

Post-scriptum

No período compreendido entre a submissão do artigo e o aceite para publicação, implementaram-se algumas mudanças no dicionário:

O *layout* do dicionário foi aprimorado, para facilitar a visualização das informações;

Novos verbetes foram incorporados, num total de 66 até o presente momento (abril de 2023);

O *corpus* foi ampliado com a inclusão das obras de Santucci (1739) e Brotero (1788); além disso, está sendo testado o formato XML, conforme apontado no item 6.

Todos os dados e códigos-fonte seguem acessíveis nos endereços informados.

Referências

BROTERO, Félix de Avelar. *Compendio de Botanica*, ou Noçoens Elementares desta Sciencia, segundo os melhores Escritores modernos, expostas na lingua Portugueza. Paris: Vende-se em Lisboa, em caza de Paulo Martin, Mercador de Livros, 1788. Disponível em: https://digitalis-dsp.uc.pt/botanica/UCFCTBt-B-78-1-15_2/UCFCTBt-B-78-1-15_2_item1/index.html. Acesso em: 09 maio 2022.

CABRAL, João. *A história natural de Portugal em Domingos Vandelli*. Lisboa: Edições Colibri, 2018.

CURTI-CONTESSOTO, Beatriz. Em busca de uma Terminologia Diacrônica sistematizada: alguns conceitos básicos em foco. *Trabalhos em Linguística Aplicada*, n. 61.1, pp. 109-124, jan./abr. 2022.

FLECK, Eliane Cristina Deckmann; DILLMANN, Mauro. "Esta receita é maravilhosa": saberes e práticas curativas na literatura médica publicada em Portugal na primeira metade do século XVIII. *História*, vol. 40, 2021. Disponível em: http://old.scielo.br/scielo.php?-pid=S0101-90742021000100433&script=sci_arttext. Acesso em: 09 maio 2022.

FUERTES-OLIVERA, Pedro A. (ed.). *The Routledge Handbook of Lexicography*. London. Routledge, 2017.

GUTIÉRREZ RODILLA, Bertha. Reflexiones historiográficas sobre el léxico científico y los repertorios lexicográficos. *In*: GARRIGA, Cecilio e PÉREZ, José Ignacio (eds.). Lengua de la ciencia y historiografia. Anexos de *Revista de Lexicografía*, 35, A Coruña: Universidade da Coruña, Servizo de Publicacións, 2016.

HOUAISS, Antônio; VILLAR, Mauro de Salles. *Grande Dicionário Houaiss da Língua Portuguesa*. s/d. Disponível em: https://houaiss.uol.com.br/. Acesso em: 28 mar. 2023.

KARPOVA, Olga; KARTASHKOVA, Faina (eds.). *Lexicography and Terminology: A Worldwide Outlook*. Cambridge, Cambridge Scholars Publishing, 2009.

LUGLI, Ligeia *et al. Visual Dictionary and Thesaurus of Buddhist Sanskrit*. 2019-2022. Disponível em: https://mangalamresearch.shinyapps.io/VisualDictionaryOfBuddhistSanskrit/. Acesso em: 09 maio 2022.

MARONEZE, Bruno. Termos neológicos em sincronias pretéritas: um estudo do *Diccionario dos Termos Technicos de Historia Natural* de Vandelli. *In*: GIL, Beatriz Daruj *et al.* (orgs.). *Saberes lexicais*. São Paulo: FFLCH-USP, 2019, p. 96-109. Disponível em: http://www.livrosabertos.sibi.usp.br/portaldelivrosUSP/catalog/book/389. Acesso em: 09 maio 2022.

MEMORIAS da Academia Real das Sciencias de Lisboa. Tomo I. Desde 1780 até 1788. Lisboa: na Typografia da Academia, 1797. Disponível em: https://www.google.com.br/books/edition/Memorias_da_Academia_Real_das_Sciencias/peQAAAAAYAAJ?hl=pt-BR&gbpv=0. Acesso em: 09 maio 2022.

NÓBREGA, Fernando Antônio Asevedo. *Desambiguação lexical de sentidos para o português por meio de uma abordagem multilíngue mono e multidocumento*. Dissertação (mestrado em Ciências de Computação e Matemática Computacional). Instituto de Ciências Matemáticas e de Computação, USP, São Carlos, 2013. Disponível em: <https://www.teses.usp.br/teses/disponiveis/55/55134/tde-28082013-145948/publico/FernandoNobrega_revisaada.pdf>. Acesso em: 11 maio 2022.

PEREIRA, Rui Abel. A afirmação do português como língua de ciência: o caso da Botânica. *Filologia e Linguística Portuguesa*, v. 19, n. 1, p. 113-126, 2017. Disponível em: https://www.revistas.usp.br/flp/article/view/121487. Acesso em: 09 maio 2022.

SANTUCCI, Bernardo. *Anatomia do corpo humano*: recopilada com doutrinas medicas, chimicas, filosoficas, mathematicas: com indices e estampas representantes todas as partes do corpo humano: dividida em tres livros. Lisboa: na Officina de Antonio Pedrozo Galram, 1739. Disponível em: https://books.google.com.br/books/about/Anatomia_do_corpo_humano.html?id=D83JL7ybBeUC&redir_esc=y. Acesso em: 09 maio 2022.

VANDELLI, Domingos. *Diccionario dos termos technicos de Historia Natural* extrahidos das Obras de Linnéo, com a sua explicaçaõ, e estampas abertas em cobre, para facilitar a intelligencia dos mesmos; e a Memoria sobre a utilidade dos jardins botanicos; que oferece a Raynha D. Maria I. Nossa Senhora / Domingos Vandelli Director do Real Jardim Botanico, e Lente das Cadeiras de Chymica, e de Historia Natural na Universidade de Coimbra. Coimbra: na Real Officina da Univesidade, 1788. Disponível em: https://purl.pt/13958. Acesso em: 09 maio 2022.

VIARO, Mário Eduardo. *Etimologia*. São Paulo: Contexto, 2011.

WELKER, Herbert Andreas. *Dicionários*: uma pequena introdução à Lexicografia. 2. ed. Brasília: Thesaurus, 2004.