O projeto *Edição Digital dos Vocabulários da Academia das Ciências:* o VOLP-1940

*Ana Salgado** & *Rute Costa***

*Academia das Ciências de Lisboa, Instituto de Lexicologia e Lexicografia da Língua Portuguesa
**NOVA CLUNL, Centro de Linguística da Universidade NOVA de Lisboa,

Abstract:

This paper presents the *Digital Edition of the Vocabularies of the Academy of Sciences* project, which aims to digitise the spelling vocabularies of the Lisbon Academy of Sciences (ACL) in order to create a digital lexicographic corpus bringing together the printed versions of all these lexicographical reference works – the 1940, 1947, 1970, and finally the 2012 editions. The first stage started with the *Vocabulário Ortográfico da Língua Portuguesa* [Orthographic Vocabulary of the Portuguese Language] (VOLP-1940), our case study. After digitising this vocabulary, the work described here focuses on the linguistic annotation of VOLP-1940 using eXtensible Markup Language (XML), an annotation metalanguage, and following the annotation directives of the Text Encoding Initiative (TEI), more specifically the application of TEI Lex-0, a new TEI sub-format. We aim to highlight the need for rigorous linguistic data processing in the creation of new lexical resources to increase the quality of their description and applicability.

Keywords: lexicography, vocabularies, Text Encoding Initiative (TEI), linguistic annotation, Digital Humanities

Palavras-chave: lexicografia, vocabulários, Iniciativa de Codificação Textual (TEI), anotação linguística, Humanidades Digitais

1. Considerações prévias

A Academia das Ciências de Lisboa (ACL)[1], por intermédio do Instituto de Lexicologia e Lexicografia da Língua Portuguesa (ILLLP)[2], manifestou a intenção de dar início à digitalização de obras lexicográficas académicas de referência. Este desígnio ditou a celebração de um protocolo entre a ACL e a Faculdade de Ciências Sociais e Humanas da Universidade NOVA de Lisboa (NOVA FCSH), por intermédio do Centro de Linguística da Universidade NOVA de Lisboa (NOVA CLUNL), grupo de Lexicologia, Lexicografia e Terminologia (LLT)[3], no qual são estabelecidas as bases de cooperação entre as duas instituições. O projeto, intitulado *Edição Digital dos Vocabulários da Academia das Ciências*, engloba a disponibilização *online* das edições impressas dos vocabulários académicos portugueses (1940, 1947, 1970, 2012) com futura conexão ao primeiro vocabulário ortográfico digital publicado em 2018[4] e a outras fontes externas.

O presente artigo centra-se, assim, no primeiro vocabulário académico – a *Edição Digital do Vocabulário Ortográfico da Língua Portuguesa* (VOLP-1940)[5], e tem como objetivo criar uma base textual informatizada anotada, de acesso aberto, em conformidade com os princípios FAIR (*Findable, Accessible, Interoperable,*

[1] http://www.acad-ciencias.pt/
[2] http://www.acad-ciencias.pt/academia/illlp-missao
[3] https://clunl.fcsh.unl.pt/grupos_clunl/lexicologia-lexicografia-terminologia/
[4] https://www.volp-acl.pt/
[5] https://clunl.fcsh.unl.pt/investigacao/projetos-curso/edicao-digital-do-vocabulario-ortografico-da-lingua-portuguesa-volp-1940/

Reusable)[6], de forma a garantir a sua conexão futura a outros sistemas e recursos existentes, em particular do mundo lusófono. Esta investigação pretende ainda suprir a lacuna que se verifica no panorama lexicográfico português no sentido de que são ainda escassos os recursos lexicográficos disponibilizados na *web* ou como fonte de acesso aberto (Williams, 2019, p. 83) e elaborados com base em normas e metodologias atuais que possibilitem a partilha e a harmonização de dados, assim como o seu alinhamento com recursos lexicais existentes. Uma consulta realizada ao *European Dictionary Portal*[7] permite-nos constatar a inexistência de dicionários patrimoniais (*legacy dictionaries*) portugueses convertidos em bases de dados, disponíveis para consulta e pesquisáveis *online.*

Uma das nossas ambições passa por nivelar Portugal relativamente a outros países europeus em que se têm verificado grandes esforços no sentido de disponibilizar em acesso aberto fontes de reconhecido valor patrimonial, já idealizadas como recursos pesquisáveis dinâmicos (*dynamic searchable resources*). Esta é uma tarefa de relevância linguística, patrimonial e histórica que seguramente irá contribuir para a fixação do léxico português na época em questão – até 1940 –, à volta da qual se tem construído e preservado a identidade de uma comunidade linguística e cultural.

Para efeitos de anotação dos dados lexicais, seguimos a TEI (*Text Encoding Initiative* ou Iniciativa de Codificação Textual)[8], um padrão internacional de marcação para a codificação de dados, especificamente a TEI Lex-0[9], um subformato simplificado da TEI para codificar dicionários no sentido amplo da aceção, cuja aplicação será alvo da nossa atenção neste artigo.

2. Enquadramento teórico

A presente contribuição enquadra-se no domínio da atuação da Lexicografia digital e assenta num conjunto de pressupostos teóricos sobre os quais pretendemos tecer algumas considerações.

O campo da Lexicografia, tendo por finalidade a elaboração de obras que são de uma grande variedade tipológica – dicionários, vocabulários, glossários, enciclopédias, etc. –, tem sofrido alterações significativas no que respeita à produção, investigação, publicação, difusão, preservação e partilha da informação. Esta mudança está diretamente relacionada com o avanço do campo das Humanidades Digitais, que rapidamente se tornou um centro catalisador na investigação académica. Se as Humanidades Digitais eram apenas associadas à computação nos seus primórdios – *Humanities Computing* (Terras *et al.*, 2003) –, hoje a sua definição está longe de ser consensual (Gold e Klein, 2012), uma vez que abrange uma grande variedade e multiplicidade de trabalhos de diferentes ramos do conhecimento que se caracterizam por recorrer a ferramentas e métodos digitais, mas que implicam, sobretudo, um novo olhar sobre as Humanidades em geral.

Dentro das Humanidades Digitais, podemos encontrar a própria Lexicografia e os dicionários, tomando aqui o termo dicionários no seu sentido mais amplo, integrando, assim o termo vocabulário no conceito do mesmo. Nos tempos atuais, é muito raro conceber recursos lexicográficos que não sejam digitais. A passagem do suporte em papel para o digital provocou um grande impacto no trabalho lexicográfico, pelo que ainda não é possível antever uma definição estável para um novo conceito de dicionário que emerge face aos avanços das Humanidades Digitais, sendo imprescindível atentar às necessidades reais dos utilizadores do século XXI. As tecnologias e ferramentas computacionais desenvolvidas nas últimas décadas têm ditado uma redefinição do trabalho lexicográfico e da forma de o conceber. Como salienta Fajardo (2018), «el concepto de diccionario digital no se ha materializado aún en un producto de características tan claramente definibles como eran los diccionarios de papel» (p. 256).

[6] https://www.go-fair.org/fair-principles/
[7] http://www.dictionaryportal.eu/en/ctlg/?objLang=pt
[8] https://tei-c.org/
[9] https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html#

Os dicionários são, antes de mais, objetos de consulta, pelo que a conversão de obras lexicográficas em recursos digitais deve ser devidamente equacionada para potenciar o acesso à informação linguística e lexicográfica aí armazenada até se converter num recurso genuinamente digital. Esta revolução digital (Trap-Jensen, 2018; L'Homme e Cormier, 2014) requer cada vez mais a aplicação de normas e *softwares* adaptados e capazes de garantir a publicação estruturada dos dados para diferentes sistemas, ou seja, garantir a interoperabilidade[10], um dos termos-chave da atual revolução. Se a digitalização de dicionários impressos assinalou a modificação de um paradigma, a disseminação da *web* obriga a repensar o conceito de obra lexicográfica. Mais do que nunca, é necessário saber tirar partido e explorar as possibilidades do ambiente digital (Trap-Jensen, 2018; Bergenholtz *et al.*, 2009), criando léxicos dinâmicos, mais robustos, enriquecidos com informações semânticas, conceptuais, estatísticas e em que os dados de diferentes recursos possam ser interligados (*linked data*)[11]. Apesar de um razoável número de trabalhos lexicográficos poder ser atualmente consultado *online*, estes recursos dicionarísticos acabam por ser estáticos, não tirando o verdadeiro proveito do ambiente digital. Como afirma Tasovac (2010, p. 1), «we cannot think of dictionaries any more without thinking about digital libraries and the status which electronic texts have in them».

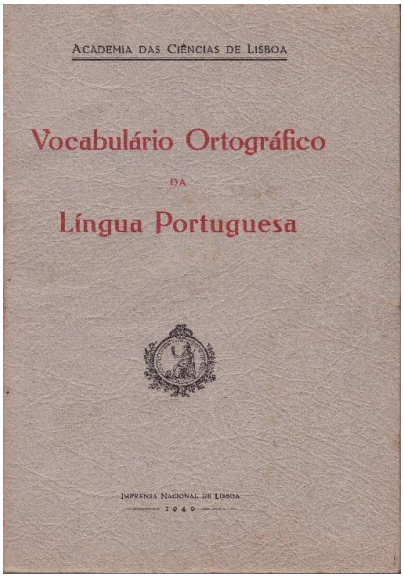
Atentos a esta nova realidade, propomo-nos aplicar estes novos princípios – métodos computacionais, padrões interoperáveis e tecnologias semânticas que facilitam a organização de grandes quantidades de dados lexicais – segundo uma metodologia rigorosa e tendo por base necessariamente o conhecimento linguístico e lexicográfico, muitas vezes ignorado ou subvalorizado na era da vertente mais tecnológica das Humanidades Digitais. A sistematização e classificação dos dados e metadados[12] requerem uma análise linguística profunda que estará subjacente em todas as fases do projeto agora em curso.

Os dicionários e os vocabulários patrimoniais que perfazem uma tradição lexicográfica europeia são uma importante herança cultural (*cultural heritage*), de valor histórico, científico e sociológico, para a sociedade. Facultando um olhar sobre o passado (Williams, 2019), a disponibilização e a preservação digital destes recursos lexicográficos são um requisito, não apenas como uma simples reprodução das obras impressas, mas como recursos dicionarísticos digitais dinâmicos, como o que pretendemos agora fazer com o nosso caso de estudo, o VOLP-1940.

3. O projeto VOLP-1940

No acervo lexicográfico português, o VOLP-1940, o primeiro vocabulário ortográfico com a chancela da ACL, foi publicado pela Imprensa Nacional de Lisboa, num só volume, com um total de 821 páginas (Figura 1).

[10] A norma ISO/IEC 2382: 2015 define interoperabilidade como: «capability to communicate, execute programs, or transfer data among various functional units in a manner that requires the user to have little or no knowledge of the unique characteristics of those units» [capacidade de comunicar, executar programas ou transferir dados entre várias unidades funcionais de uma maneira que exija que o utilizador tenha pouco ou nenhum conhecimento das características exclusivas dessas unidades].
[11] https://www.w3.org/standards/semanticweb/data.html
[12] https://groups.niso.org/apps/group_public/download.php/17446/Understanding%20Metadata.pdf



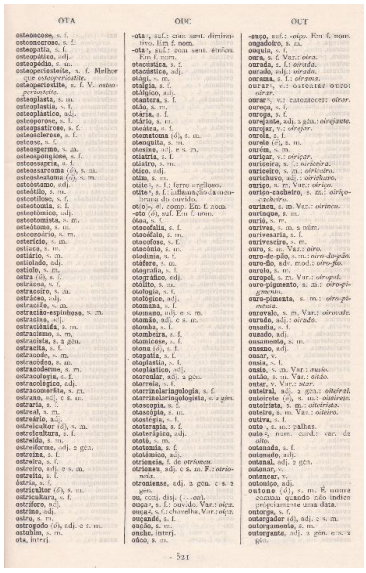


Figura 1: VOLP-1940 (ACL) e amostra da nomenclatura

Além de ser o primeiro de uma série de vocabulários subsequentes, esta obra lexicográfica tem um grande valor histórico e linguístico. O século XX foi uma centúria produtiva no panorama da lexicografia do português, tendo sido publicados vários vocabulários ortográficos, não só por instituições como a ACL ou a Academia Brasileira de Letras (ABL) como também vieram a lume edições particulares, da autoria de diferentes filólogos e linguistas, quer portugueses quer brasileiros, bem como versões resumidas. Anterior ao nosso caso de estudo, em 1912, Gonçalves Viana (1840–1914) publica o *Vocabulário Ortográfico e Remissivo da Língua Portuguesa: com mais de 100 000 vocabulos, conforme a ortografia oficial*. Antenor Nascentes, um ano após da publicação do primeiro vocabulário da ACL, em 1941, publica, no Brasil, *Vocabulário Ortográfico do Idioma Nacional.* Dois anos depois, em 1943, é a vez da ABL publicar o *Pequeno Vocabulário Ortográfico da Língua Portuguesa*, voltando apenas a publicar um novo vocabulário em 1981, da autoria de Antônio Houaiss, com sucessivas reedições até 2009. Mas é em 1966 que se publica o vocabulário que ainda hoje, mesmo com a aplicação de novas regras de escrita, continua a ser uma obra de referência para muitos dos profissionais da língua. Referimo-nos ao *Vocabulário da Língua Portuguesa*, de Rebelo Gonçalves, publicado pela Coimbra Editora em 1966. Já no século XXI, em 2010, e inaugurando a disponibilização de obras deste cariz *online*, o Instituto de Linguística Teórica e Computacional (ILTEC) publica o *Vocabulário Ortográfico do Português*[13], sob coordenação de Margarita Correia com o apoio e financiamento do Fundo da Língua Portuguesa. Este mesmo vocabulário será a base de trabalho para o *Vocabulário Ortográfico Comum da Língua Portuguesa*[14] (VOC), um instrumento para uma política da língua, construído pelos países que têm o português como língua oficial. À exceção dos vocabulários académicos brasileiro[15] e português, que já se encontram *online*, e do vocabulário do ILTEC que deu origem ao VOC, nenhum outro vocabulário foi objeto de conversão para o digital, o que reforça o que acima referimos sobre a escassez de recursos lexicográficos de valor patrimonial disponíveis *online.*

[13] http://www.portaldalinguaportuguesa.org/?action=vop&page=info
[14] https://voc.cplp.org/
[15] https://www.academia.org.br/nossa-lingua/busca-no-vocabulario

O VOLP-1940 constitui uma parte fundamental do património cultural português, uma vez que serviu como ferramenta para discutir uma nova medida ortográfica entre a ACL e a Academia Brasileira de Letras, que veio a resultar na convenção ortográfica luso-brasileira de 1945[16], vulgo Acordo Ortográfico de 1945, em vigor até 2011. Em termos de ortografia, o VOLP segue a base da reforma ortográfica de 1911[17], recorrendo ainda a outras duas bases acessórias: a reforma de 1920 segundo a Portaria n.º 2:533, de 29 de novembro de 1920, que alterou as disposições de 1911, e o *Acordo Ortográfico Luso-Brasileiro* (1931), celebrado entre a academia portuguesa e a brasileira em 30 de abril de 1931, aprovado e mandado executar em Portugal pela Portaria n.º 7:117, de 27 de maio do mesmo ano, e no Brasil pelos Decretos n.ºs 20:108 e 23:028, respetivamente de 15 de junho de 1933 e de 2 de agosto de 1933. Apresenta este vocabulário uma «revisão metódica das disposições» (p. X) desta tentativa de acordo.

A nomenclatura «abrange apenas a língua portuguesa moderna, isto é, o período linguístico que decorre do século XVI até à época actual» [entenda-se 1940] (p. XII), registando unidades lexicais que deram entrada na língua depois de 1500, deixando de fora unidades «pertencentes ao período arcaico do idioma» (*idem*). No entanto, arcaísmos ainda usados à época podem ser encontrados nesta obra.

3.1. A macroestrutura e microestrutura

Em termos macroestruturais, o VOLP-1940 é constituído por três secções: (i) vocabulário comum (pp. 3–713) do «léxico geral da língua descontados os nomes próprios», incluindo elementos de composição (p. IX); (ii) vocabulário onomástico (pp. 717–809), «nomes próprios de várias categorias» (p. IX), tais como antropónimos, topónimos e patronímicos, assim como etnónimos, hierónimos, mitónimos, cronónimos e bibliónimos; (iii) apêndice de um «Registo de abreviaturas» de uso corrente em finais da década de 30 do século XX (pp. 813–819), «portuguesas e ainda de outras não portuguesas que são empregadas na nossa escrita […] as abreviaturas de maior importância para os usos correntes e de maior curiosidade geral para os dois países de língua portuguesa» (p. IX). A obra apresenta uma dedicatória: «Às Nações Portuguesa e Brasileira oferece e consagra, no Duplo Centenário da Fundação e da Restauração de Portugal, a Academia das Ciências de Lisboa.» Segue-se a «Introdução», prefaciada «por Francisco Rebelo Gonçalves […], um dos dois grandes filólogos-ortógrafos do século xx em Portugal» (Anselmo, 2011). Como salienta Anselmo (2011), os «Comentários ortográficos», da lavra de Rebelo Gonçalves, enriquecem o tomo ao apresentar uma justificação das opções gráficas no registo de certos vocábulos duvidosos.

O nosso trabalho, ainda numa fase incipiente, apenas se centra na secção (i) vocabulário comum. As restantes partes serão objeto de estudo e análise quando esta estiver concluída.

Mas como definir, afinal, vocabulário ortográfico? Um vocabulário ortográfico é uma lista de palavras na sua forma gráfica oficial com indicação da categoria morfossintática, podendo ainda incluir informações adicionais, como ortoépia, especificidades de flexão, regras de escrita, entre outras. É «um instrumento fundamental para a gestão da ortografia da língua» (Academia Brasileira de Letras, 2017).

No que diz respeito à sua microestrutura, com base na Figura 2, observamos que um artigo lexicográfico, regra geral, pode incluir os seguintes elementos: lema; ortoépia; categoria; significado.

[16] Aprovado pelo Decreto n.º 35 228, de 8 de dezembro de 1945. Alterado pelo Decreto-Lei n.º 32/73, de 6 de fevereiro. Consultar: http://www.portaldalinguaportuguesa.org/?action=acordo&version=1945.
[17] http://www.portaldalinguaportuguesa.org/?action=acordo&version=1911

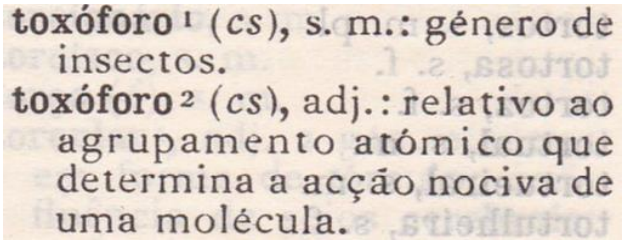

toxóforo [1] (*cs*), s. m.: género de insectos.
toxóforo [2] (*cs*), adj.: relativo ao agrupamento atómico que determina a acção nociva de uma molécula.

Figura 2: Entrada **toxóforo** do VOLP-1940

Na Figura 2, ilustramos um caso de homonímia, «toxóforo», enquanto lema desdobrado. Segundo a tradição lexicográfica portuguesa, o lema ou a «forma da unidade lexical que é usada para a representar no artigo lexicográfico» (Correia, 2009, p. 132) corresponde à forma do singular do nome ou adjetivo e à forma masculina quando há flexão em género em palavras variáveis e, no caso de verbos, corresponde à forma do infinitivo impessoal. Assim acontece no VOLP-1940 em que estas unidades lexicais vêm impressas em negrito redondo como forma de realce. No caso particular de palavras homónimas, como se pode observar na Figura 2, um algarismo em sobrescrito ([1], [2]) do lado direito do lema identifica este tipo de vocábulos e distingue-os.

A ortoépia ou indicação normativa da pronúncia de uma unidade lexical apenas se assinala em palavras de pronúncia duvidosa, como é, por exemplo, o caso do som da letra *x* que pode corresponder a /cs/ – na Figura 2 corresponde a (*cs*) –, /ss/ ou /z/. O timbre das vogais tónicas fechadas *e* e *o*, quando não acentuadas graficamente (por exemplo, «gaveta» /ê/ ou «bochornoso» /ô/) também pode ser facultado, ou quando uma determinada vogal tónica possa ser frequentemente pronunciada de forma incorreta (por exemplo, «boiz» /í/ ou «padeira» /à/). Essa indicação, de natureza normativa, é dada entre parênteses curvos após o lema.

A categoria, ou seja, a «indicação da classe de palavras a que pertence a entrada a ser descrita» (Correia, 2009, p. 129) surge após o lema ou a ortoépia quando é assinalada e é indicada de forma abreviada com letras minúsculas.

O significado do lema é apresentado, em poucos casos, essencialmente para desambiguar casos de homonímia, aos quais é acrescido, como acima referido, um número sobrescrito do lado direito do lema.

Os elementos de composição também constam da nomenclatura do VOLP-1940. O lema, neste caso, vem seguido de um hífen e após a indicação «el. comp.» (elemento de composição), à qual se segue um texto descritivo do emprego desse elemento, apresentando, no final, exemplos ilustrativos da aplicação da regra enunciada (Figura 3).

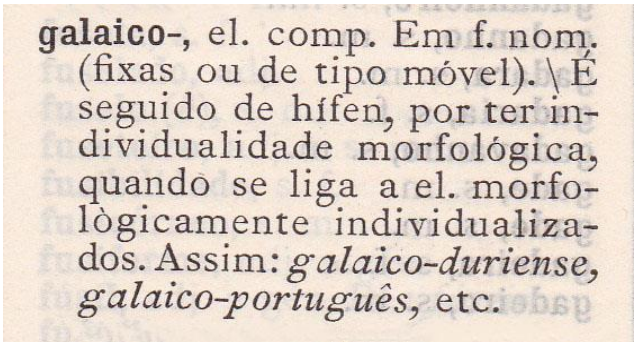

galaico-, el. comp. Em f. nom. (fixas ou de tipo móvel). \ É seguido de hífen, por ter individualidade morfológica, quando se liga a el. morfològicamente individualizados. Assim: *galaico-duriense*, *galaico-português*, etc.

Figura 3: Entrada **galaico-** do VOLP-1940

Verificam-se ainda registos de variantes ortográficas – Rebelo Gonçalves apelidava-as de «variações vocabulares» –, por exemplo, «ouriço» e «oiriço», como se pode observar na Figura 4. As variantes, em princípio, não figuram na nomenclatura, ou seja, «oiriço» não surge na lista alfabética, podendo apenas ser encontrada no artigo «ouriço». Há algumas exceções a este critério que são explicadas na «Introdução» (p. XVIII), como, por exemplo, «cousa» e «coisa», sempre que a variante é mais usual do que a forma básica.

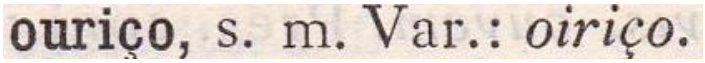

ouriço, s. m. Var.: *oiriço*.

Figura 4: Entrada **ouriço** do VOLP-1940

O VOLP-1940 também fornece indicações morfológicas, nomeadamente o plural de nomes terminados em -ão, cuja formação, por vezes, suscita dúvida por se puderem formar em -ãos, -ões e -ães, este último em número reduzido. É o caso, por exemplo, da entrada da Figura 5, «corrimão», que apresenta duas formas de plural.

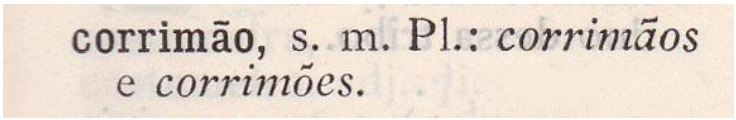

corrimão, s. m. Pl.: *corrimãos* e *corrimões*.

Figura 5: Entrada **corrimão** do VOLP-1940

Constata-se também o registo de informações sobre o uso de vocábulos, praticamente exclusivo, em locuções. Se um determinado vocábulo, por exemplo, só se usa em determinada locução, essa indicação é fornecida como se verifica na Figura 6, ou seja, surge como entrada do que se considera ser a palavra nuclear dessa mesma locução.

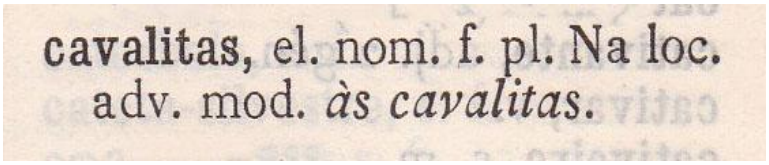

cavalitas, el. nom. f. pl. Na loc. adv. mod. *às cavalitas*.

Figura 6: Entrada **cavalitas** do VOLP-1940

Uma outra indicação normativa diz respeito a construções que iniciam pela fórmula «Melhor que». As formas consideradas preferenciais são aquelas que se consideram mais próximas do seu étimo ou mais corretas por determinados motivos. É o caso de «colocintina» e «coloquintina», termos químicos hoje obsoletos (Figura 7).

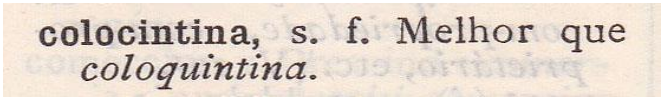

colocintina, s. f. Melhor que *coloquintina*.

Figura 7: Entrada **colocintina** do VOLP-1940

Estes são os componentes essenciais e mais relevantes do VOLP-1940 já identificados. Esta análise é determinante na fase de anotação linguística discutida na secção 4.

3.2. A conversão do VOLP para suporte digital

Como referem Marquilhas e Hendrikx (2016), as grandes vantagens da informática refletem-se na «prevenção de erro humano em transcrições e edições e prevenção de abandono de tarefas demasiado gigantescas para a capacidade humana» (p. 252). Estas vantagens, grandes benefícios daquilo que hoje designamos por Humanidades Digitais, aplicam-se na perfeição ao trabalho lexicográfico *per se*. Basta folhear algumas páginas do VOLP-1940 para detetar algumas gralhas tipográficas (por exemplo, «S.f.», quando o correto seria «s.f.»). Estes casos, que hoje seriam rapidamente colmatados com o trabalho feito em computador, à época, representavam um enorme esforço humano no que respeita à sistematização e uniformização dos dados ou, porventura, poderiam exigir demasiadas provas de revisão, o que nem sempre era viável. Por outro lado, é sabido que o trabalho hercúleo de fazer dicionários nem sempre chegou a bom porto (cite-se o próprio caso da ACL que só em 2001 conseguiu, pela primeira vez, publicar um dicionário completo, de A a Z).

Como referido no texto introdutório, é nosso propósito: i) criar um novo recurso lexicográfico *online*, acessível a toda a comunidade científica, bem como ao público em geral; ii) melhorar a consistência dos metadados originais, seguindo uma anotação linguística rigorosa conforme as recomendações TEI, garantindo simultaneamente a acessibilidade aos dados e sua reutilização; iii) descrever a anotação linguística para um posterior enriquecimento semântico da base de dados; iv) acrescentar novos metadados, nomeadamente a inclusão de marcas de domínios, informação que será recuperada de outras obras lexicográficas que contêm essa anotação, e fazer a ligação entre várias unidades sinónimas que constem da nomenclatura da obra.

Em termos metodológicos, a nossa atenção recai sobre: i) a metodologia usada para criar o VOLP-1940 retrodigitalizado, com identificação dos seus componentes estruturais após OCR; ii) a organização das informações lexicais e sua representação em TEI.

4. Metodologia

A digitalização da obra resultou numa série de ficheiros em imagem do documento original que foram convertidos em texto corrido, usando um programa comercial de reconhecimento de caracteres (OCR, *Optical Character Recognition*), o *Omnipage Pro*. O texto foi, posteriormente, exportado para um programa de edição (documento do *Microsoft Word*) com o intuito de corrigir gralhas e inconsistências geradas pelo OCR. Prosseguiu-se com a identificação das convenções lexicográficas do VOLP-1940 (por exemplo, a vírgula usada após cada lema ou o uso de abreviaturas listadas nas páginas iniciais da obra em papel) para iniciar um trabalho de uma possível anotação automatizada de toda a obra.

A introdução e a lista geral de abreviaturas das páginas iniciais da edição impressa foram lidas, analisadas e anotadas em TEI para publicação em ambiente digital.

Na fase seguinte, após a preparação do formato de codificação para um formato compatível com XML, iremos recorrer a algumas ferramentas e/ou tecnologias para garantir um tratamento uniforme e rigoroso de todos os metadados do vocabulário. Como referido anteriormente, a codificação será baseada nos padrões TEI de acordo com o subformato TEI Lex-0.

A terceira fase, embora tenha sido iniciada em simultâneo com a primeira, é de natureza puramente linguística e também de cariz computacional. Aproveitando a distinção entre unidades mono e polilexicais (Salgado *et al.*, 2019) – ponto explorado na subsecção 5.2.2. –, essa informação será recuperada e inserida após cada lema. Ainda que seja uma informação não visível para o utilizador final, estes dados serão úteis para incrementar as possibilidades do sistema de pesquisa. Os dados morfossintáticos, isto é, a anotação das classes de palavras e seu respetivo género no caso de nomes e adjetivos será alvo de uma verificação final e a sua consistência validada para efeitos de demonstração estatística aquando da publicação da obra. Por fim, analisaremos questões semânticas, nomeadamente os casos de homonímia. No que concerne ao enriquecimento da base textual, temos como meta a introdução de marcas de uso, nomeadamente marcas de domínio. Esta atribuição está diretamente relacionada com

a perspetiva futura de uma ontologia de etiquetas de domínio (Costa *et al.* 2020) e com a preocupação de alinhamento desses metadados com uma qualquer outra obra lexicográfica da ACL.

No que concerne à ortografia original dos documentos, esta será preservada. No entanto, disponibilizar hoje um recurso ao público e considerar a prevalência dos mecanismos de pesquisa exige a modernização das grafias (*Simões et al.*, 2012), principalmente ao nível do lema. A ortografia original do lema será alinhada com as ortografias mais atuais. Na execução desta tarefa, criaremos posteriormente uma correspondência entre as grafias segundo a convenção ortográfica luso-brasileira de 1945 e o *Acordo Ortográfico da Língua Portuguesa* de 1990[18], aproveitando um trabalho anteriormente desenvolvido para o *Vocabulário Ortográfico da Língua Portuguesa* (VOLP-ACL)[19] da ACL. O resultado permitirá que o utilizador final possa pesquisar as grafias atuais, com as quais está familiarizado, e encontre a entrada correspondente à grafia antiga, mantendo o recurso fiel ao original.

A base de dados do VOLP-1940 será alojada no LeXmart (Simões *et al.*, 2019)[20], uma plataforma *web*, de código aberto, criada para a edição e publicação de recursos lexicais.

5. Aplicação da TEI

5.1. TEI

A codificação do VOLP-1940 segue as diretrizes[21] da TEI. Apesar de a TEI não ter o *status* legal de norma (Stührenberg, 2012), tornou-se uma norma internacional *de facto* para a codificação de diferentes géneros de documentos, desde manuscritos, poemas, dicionários, receitas culinárias, anotação de *corpora*, entre muitos outros. Criada em 1987 por um consórcio de diversas instituições – TEI Consortium – com o objetivo de desenvolver um formato normalizado para a edição eletrónica de conteúdos textuais em múltiplos formatos, a TEI apresenta uma metalinguagem que compreende um vocabulário (um conjunto de elementos e atributos) e uma gramática (um esquema) para anotar, estruturar e validar documentos e cujas sintaxe e semântica específicas em linguagem XML (*eXtensible Markup Language*) a torna num método de análise textual para processamento digital.

A versão atual das *TEI Guidelines* (TEI P5) foi publicada em 2 de novembro de 2007 e continua a ser alvo de atualizações constantes.[22] Optámos por seguir este formato normalizado por ser comummente usado na edição e preservação digital de documentos[23] e ainda por ser interoperável. Por outro lado, o mesmo tem sido amplamente testado em projetos autorais com bons resultados (Costa *et al.*, 2020; Salgado *et al.*, 2019).

A TEI P5 tem um módulo específico inteiramente dedicado a dicionários, o capítulo 9[24]. Toma-se o vocábulo «dicionários» na sua aceção mais geral, ou seja, englobando não só dicionários, mas vocabulários, enciclopédias, glossários, etc., como previamente referido. É reconhecida pela comunidade científica (Salgado *et al.*, 2019) a complexidade dos recursos lexicográficos, quer pela sua diversificada componente estrutural, quer porque diferentes recursos optam por diferentes critérios no que respeita à representação e tratamento da informação lexicográfica, daí a importância de seguirmos uma norma que assegure o trabalho do PLN (processamento de língua natural), tal como o alinhamento de entradas, isto é, a correspondência entre entradas, ou sentidos, e a identificação de pares de sentidos que, entre um ou mais recursos lexicais, apresentem o mesmo significado.

[18] https://dre.pt/application/file/a/403254
[19] https://www.volp-acl.pt/
[20] http://www.lexmart.eu/
[21] https://tei-c.org/Guidelines/
[22] À data de publicação deste artigo, a última atualização foi realizada em 19 de agosto de 2020 (versão 4.1.0).
[23] Alguns exemplos de aplicação: cf. BASnum, Nénufar, ARTFL, VICAV, ICLTT's Dictionaries.
[24] https://tei-c.org/release/doc/tei-p5-doc/en/html/DI.html

5.2. TEI Lex-0

Como as diretrizes da TEI se caracterizam pela sua extrema flexibilidade de anotação – várias possibilidades de codificação para os mesmos componentes, o que representa um obstáculo à interoperabilidade entre diferentes recursos –, um novo subformato, mais restrito, está a ser desenvolvido especificamente para ser aplicado a recursos lexicais: o TEI Lex-0 (Tasovac *et al.*, 2018; Romary e Tasovac, 2018; Bański *et al.*, 2017). A preparação deste formato teve início em 2016 e hoje é liderado por um grupo de trabalho do DARIAH[25], formado por especialistas em recursos lexicais. O objetivo da TEI Lex-0 é definir uma estrutura de anotação clara e versátil, mas não demasiadamente permissiva, para facilitar a interoperabilidade de recursos lexicais codificados heterogeneamente. A TEI Lex-0 deve ser considerada como «a format that existing TEI dictionaries can be unequivocally transformed to in order to be queried, visualised, or mined uniformly».[26] Como o esquema deste formato ainda não está fechado, temos contribuído ativamente para o desenvolvimento do mesmo com a criação de *issues* no GitHub.[27]

5.2.1. TEI header

Cada documento codificado em XML-TEI é dividido em duas partes: um cabeçalho (designado `TeiHeader`) e o conteúdo real do documento (o `text`).

Para o VOLP-1940, criámos um `teiHeader` no qual se anota a metainformação do documento de trabalho, isto é, os dados bibliográficos quer do documento original, quer do documento eletrónico, para futuramente serem usados por motores de pesquisa.

O `teiHeader` é constituído pelo *File description* (`fileDesc`), que é obrigatório e apresenta a descrição bibliográfica completa do documento. Sendo um elemento obrigatório, o `teiHeader` do VOLP-1940 foi elaborado e apresenta: (a) o elemento `titleStmt` que contém dados bibliográficos do documento eletrónico, como seja o seu título; (b) o elemento `publicationStmt` que contém dados relacionados com a publicação e distribuição do documento eletrónico; (c) o elemento `sourceDesc` que fornece uma descrição bibliográfica da fonte original. Note-se que existem dois elementos `title`, um, que corresponde ao título do documento eletrónico (em `titleStmt`) e outro, que corresponde ao título do documento original (em `sourceDesc`).

```
<teiHeader>
   <fileDesc>
      <titleStmt>
         <title>VOLP-1940</title>
      </titleStmt>
      <publicationStmt>
         <publisher>Academia das Ciências de Lisboa / Imprensa Nacional de
Lisboa</publisher>
      </publicationStmt>
      <sourceDesc>
         <bibl>
            <title>Vocabulário Ortográfico da Língua Portuguesa</title>
            <extent>1 volume</extent>
            <extent>821 pp.</extent>
            <author>Academia das Ciências</author>
            <publisher>Imprensa Nacional de Lisboa </publisher>
            <date>1940</date>
```

[25] https://www.dariah.eu/activities/working-groups/lexical-resources/
[26] https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html#index.xml-body.1_div.1
[27] https://github.com/DARIAH-ERIC/lexicalresources/projects/1

```
            </bibl>
         </sourceDesc>
      </fileDesc>
   </teiHeader>
```

Exemplo (1): Elemento `teiHeader` do VOLP-1940

5.2.2. Estrutura básica de uma entrada do VOLP-1940 em TEI Lex-0

Tratando-se de um vocabulário ortográfico, o VOLP-1940 é constituído por artigos lexicográficos. Como já referido, estes iniciam por um lema (a entrada), seguido da informação morfossintática sobre essa unidade. Eis a estrutura básica e regular de uma entrada do VOLP-1940:

```
<entry xml:id="..." xml:lang="pt" type="...">
   <form type="lemma">
      <orth>...</orth>
   </form>
   <gramGrp>
      <gram type="pos">...</gram>
      <gram type="gen">...</gram>
   </gramGrp>
</entry>
```

Exemplo (2): Estrutura básica e regular de uma entrada do VOLP-1940

Essa estrutura básica pode ser observada no seguinte exemplo:

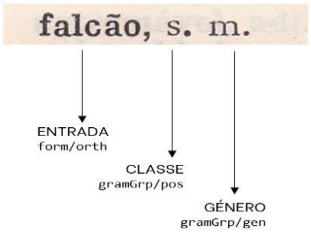


Figura 8: Entrada **falcão** do VOLP-1940 e respetiva descodificação dos componentes

A unidade monolexical **falcão**, como representada na Figura 8, apresenta algumas características tipográficas tradicionais: entrada em negrito delimitada por uma vírgula, classe da palavra, indicada de forma abreviada, seguida pela indicação de género por se tratar de um nome («s. m. [substantivo masculino]»), uma opção geralmente usada em edições impressas para economizar espaço. Ao aplicar TEI Lex-0 a este exemplo, o resultado é o seguinte:

```
<entry xml:lang="pt" xml:id="falcao" type="monolexicalUnit">
   <form type="lemma">
      <orth>falcão</orth>
   </form>
   <pc>,</pc>
   <gramGrp>
      <gram type="pos" norm="NOUN">s.</gram>
      <gram type="gen">m.</gram>
   </gramGrp>
</entry>
```

Exemplo (3): Codificação da entrada **falcão** do VOLP-1940 em TEI Lex-0

O elemento `entry` engloba todo o artigo lexicográfico. O elemento `form` anota as informações relativas ao lema. Este elemento também especifica o seu atributo `type` como `"lema"`, e a forma ortográfica é fornecida no elemento `orth`. É importante salientar que, em TEI Lex-0, o elemento `entry` requer os atributos `@xml:id`[28], o identificador da entrada e `@xml:lang`, o código de idioma apropriado conforme a IETF BCP 47[29] que, por sua vez, se baseia na ISO 639[30]. Sendo uma entrada de um vocabulário, estamos a usar `form type=lema`.

Importa salientar que, em Salgado *et al.* (2019), identificámos quatro tipos diferentes de entradas no trabalho de conversão para TEI Lex-0 do *Dicionário da Língua Portuguesa Contemporânea* (DLPC, 2001), classificação essa que será replicada neste trabalho: unidades monolexicais – que, no Exemplo (3), corresponde ao atributo `type "monolexicalUnit"` –, unidades polilexicais, afixos e abreviações. Segundo os autores, as entradas polilexicais podem ser de dois tipos diferentes: (i) compostos[31]; (ii) todos os tipos de combinações lexicais, como colocações ou frasemas. No âmbito deste vocabulário em particular e, de um modo mais geral na tradição ortográfica portuguesa, as unidades polilexicais que apresentam um elevado grau de unidade formal e semântica e formadas por compostos morfológicos (por exemplo, «austro-alemão»), compostos morfossintáticos (por exemplo, «surdo-mudo», «peixe-boi») ou por compostos sintagmáticos (por exemplo, «água-de-colónia» (Rio-Torto, 2013), são registadas como entrada, ou seja, a um nível macroestrutural. Importa ainda aqui referir o caso das palavras derivadas por prefixação que eram escritas com hífen segundo as bases ortográficas da época, regras agora alteradas com o novo acordo (por exemplo, «anti-escravismo ou «ultra-romântico»). Na tradição lexicográfica portuguesa, as palavras hifenizadas são registadas em entrada, pelo que as unidades lexicais acima enunciadas são consideradas lemas no presente vocabulário. O facto de termos adotado esta classificação (unidades monolexicais/polilexicais), e de as termos anotado, permitirá, futuramente, localizar de forma rápida e automática os fenómenos que se pretendem observar. Esta anotação não será explícita, isto é, a informação não será visível para o utilizador final, sendo, por isso, recomendável, tal como se demonstra, acrescentar essa informação como um atributo `@type` em `entry`.

As propriedades morfossintáticas de uma entrada lexical devem ser especificadas em `entry`/`gramGrp`. Anotámos as classes morfológicas de palavras usando `@type="pos"`, marcando também o género como `@type "gen"`. Estamos também a equacionar o uso do atributo `@norm` para os valores da *Universal Dependencies Part-of-Speech*[32] para efeitos de interoperabilidade futura com outros recursos lexicográficos nacionais e internacionais. Para garantir a precisão dessa conversão, uma lista de possibilidades completas para o conteúdo dessa etiqueta será calculada, e a anotação será adicionada manualmente. Na Tabela 1, apresentamos uma amostra do levantamento efetuado:

VOLP-1940	*Universal Dependencies Part-of-Speech*	*Universal POS tags*
CLASSE ABERTA DE PALAVRAS / *OPEN CLASS WORDS*		
adjetivo (adj.)	*adjective*	ADJ
advérbio (adv.)	*adverb*	ADV
interjeição (inter.)	*interjection*	INTJ
substantivo (s.)	*noun*	NOUN
verbo (v.)	*verb*	VERB

[28] O formato XML não permite o uso de caracteres acentuados nos identificadores de elementos.
[29] https://tools.ietf.org/html/bcp47
[30] https://www.iso.org/iso-639-language-codes.html
[31] «By compounds we mean every lexical unit formed by two or more elements with autonomy within the language that together form a new lexical unit with a new meaning. », *in* Salgado *et al.* (2019).
[32] As etiquetas POS universais são marcas de classes de palavras usadas na *Universal Dependencies* (UD), que é um projeto que está a desenvolver uma anotação *treebank* multilingue consistente para vários idiomas. O esquema de anotação baseia-se na evolução de dependências (universais) de Stanford, etiquetas universais de classes de palavras do Google e na Interset para etiquetas morfossintáticas. *Vd*. https://universaldependencies.org/u/pos/.

CLASSE FECHADA DE PALAVRAS / *CLOSED CLASS WORDS*		
artigo (art.)	*determiner*	DET
conjunção (conj.)	*coordinating conjunction* *subordinating conjunction*	CCONJ SCONJ
numeral (num.)	*numeral*	NUM
preposição (prep.)	*adposition*	ADP
pronome (pron.)	*pronoun*	PRON

Tabela 1: Amostra de correspondências entre as classes de palavras do VOLP-1940 e a *Universal Dependencies Part-of-Speech*.

No caso particular de palavras homónimas, como é o caso do exemplo ilustrado na Figura 2, «toxóforo», o lema surge desdobrado. Em TEI Lex-0, evitando possíveis ambiguidades estruturais, o elemento `superEntry` deixou de ser permitido, e usamos `entry` de forma sistemática. Para assinalar o índice numérico, o elemento `lbl` preserva o algarismo do original. O atributo `n` de `entry` revelar-se-á, por sua vez, importante para o processamento posterior da entrada por ferramentas computacionais, uma vez que `lbl` tem um uso mais generalizado.

```
<entry xml:lang="pt" xml:id="toxoforo_1" n="1" "type="monolexicalUnit">
   <form type="lemma">
      <orth>toxóforo</orth>
      <lbl>1</lbl>
      <pc>(</pc>
      <pron extend="part">cs</pron>
      <lbl>)</lbl>
   </form>
   <pc>,</pc>
   <gramGrp>
      <gram type="pos" norm="ADJ">adj.</gram>
   </gramGrp>
   <pc>:</pc>
   <sense>
      <def>género de insectos</def>
      <pc>.</pc>
   </sense>
</entry>
```

Exemplo (4): Codificação da entrada **toxóforo**[1] do VOLP-1940 em TEI Lex-0

De somenos importância, refira-se que para todo e qualquer sinal de pontuação, como é o caso da vírgula que serve como delimitador de campo a seguir ao lema, dos parênteses curvos para delimitar a ortoépia, do ponto final que encerra o artigo lexicográfico, ou, por vezes, dos dois-pontos que introduzem informação a seguir ao lema, usaremos sempre o elemento `<pc>` para preservar a informação original na anotação.

Como referido anteriormente, as entradas do VOLP-1940, por vezes, fornecem indicações ortoépicas para esclarecer a pronúncia, como é o caso do som da letra *x*, como se demonstra na Figura 9.

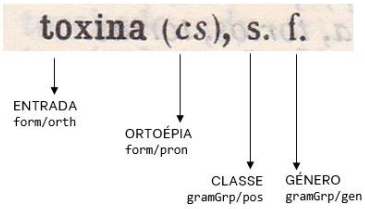


Figura 9: Entrada **toxina** do VOLP-1940 e respetiva descodificação dos componentes

Recorremos ao elemento `form/pron`, seguido do atributo `extend` e o valor `"part"`, que indica que a pronúncia indicada é referente a parte do lema. Futuramente, e como o elemento `pron` tem sido usado para fornecer indicações de transcrição fonética noutras obras lexicográficas da ACL, definiremos um valor específico dependente do atributo `notation` para assinalar que se trata de ortoépia. Veja-se o Exemplo (5):

```
<entry xml:lang="pt" xml:id="toxina" type="monolexicalUnit">
   <form type="lemma">
      <orth>toxina</orth>
      <pc>(</pc>
      <pron extend="part">cs</pron>
      <lbl>)</lbl>
   </form>
   <pc>,</pc>
   <gramGrp>
      <gram type="pos" norm="NOUN">s.</gram>
      <gram type="gen">f.</gram>
   </gramGrp>
</entry>
```

Exemplo (5): Codificação da entrada **toxina** do VOLP-1940 em TEI Lex-0

5.2.3. O enriquecimento do VOLP-1940

Apesar de pretendermos ser fiéis ao original, é nosso intuito enriquecer a base textual com informação que fornecerá ao utilizador uma maior capacidade de resposta na hora de consulta: a possibilidade de encontrar variantes ortográficas associadas à palavra pesquisada, variantes lexicais, informação terminológica (marcas de domínio), sinónimos, entre outras. De seguida, passaremos a descrever alguns desses melhoramentos.

5.2.3.1. Variantes ortográficas

As variantes ortográficas, que, de uma forma geral não constam da nomenclatura do VOLP-1940, serão adicionadas ao lema para possibilitar a pesquisa por ambas as unidades.

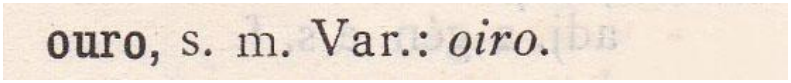


Figura 10: Entrada **ouro** do VOLP-1940

```
<entry xml:lang="pt" xml:id="ouro" type="monolexicalUnit">
   <form type="lemma">
      <orth>ouro</orth>
   </form>
   <form type="variant" xml:id="oiro" xml:lang="pt">
      <orth>oiro</orth>
   </form>
   <pc>,</pc>
   <gramGrp>
      <gram type="pos" norm="NOUN">s.</gram>
      <gram type="gen">m.</gram>
   </gramGrp>
   <note>Var.: oiro</note>
   <pc>.</pc>
</entry>
```

Exemplo (6): Codificação da entrada **ouro** do VOLP-1940 em TEI Lex-0

Como se pode observar no Exemplo (6), o acrescento das variantes ortográficas será anotado em `form type="variant"`, ou seja, adicionando informação implícita – que não visível para o utilizador –, sendo útil para pesquisa futura por variante. O elemento `note`, de momento, está a ser usado para albergar qualquer outro tipo de informação, que não necessita de diferenciação, para preservar o conteúdo original.

5.2.3.2. Variedades geográficas

Um dos pontos que pretendemos ver refletido neste trabalho é a diversidade da língua portuguesa em termos de variedades geográficas. Consideraremos, para tal, a variedade europeia (a de português de Portugal) – o nosso ponto de partida, considerando por omissão que as unidades registadas no VOLP-1940 são de português de Portugal –, a variedade brasileira (português do Brasil) e também as variedades africanas (tendo em consideração os países africanos de expressão portuguesa). Em TEI Lex-0, regra geral, a variação pode ser codificada como forma [`@type="variant"`] e incorporada no elemento pai para o qual uma característica subordinada exibe variação. No exemplo a seguir, adicionamos a variante «miçanga», típica no português do Brasil, à entrada do VOLP-1940 «missanga», típica de Portugal. Em termos de codificação, a variante brasileira corresponde a `form type="variant"`, que será um elemento implícito, e a indicação da área geográfica é dada em `<usg type="geographic">`. Uma vez mais, este acrescento possibilitará a consulta da base por qualquer utilizador de expressão portuguesa.

```
<entry xml:lang="pt" xml:id="missanga" type="monolexicalUnit">
   <form type="lemma">
    <orth>missanga</orth>
   </form>
   <form type="variant" xml:id="micanga" xml:lang="pt">
    <usg type="geographic">Brasil</usg>
     <orth>miçanga</orth>
   </form>
   <pc>,</pc>
   <gramGrp>
    <gram type="pos" norm="NOUN">s.</gram>
    <gram type="gen">f.</gram>
```

```
  </gramGrp>
 </entry>
```

Exemplo (7): Codificação da entrada **missanga** do VOLP-1940 em TEI Lex-0

5.2.3.3. Outras informações

Existem outras informações que poderão ser reaproveitadas de obras académicas anteriores também para melhorar a resposta da pesquisa. A ligação com outras obras está prevista, como previamente enunciado. Pretende-se, assim, tirar proveito de todos os componentes que podem vir a enriquecer os resultados. Cite-se, por exemplo, o caso das marcas de domínios, que podem ser extraídas do dicionário publicado em 2001 (DLPC). Como se pode observar no Exemplo (8), a etiqueta de domínio «Zool.» (Zoologia) é inserida em `usg type="domain"`. Essa informação, uma vez mais, estará implícita para mecanismos de pesquisa, não sendo, contudo, visível para o utilizador, para preservar o conteúdo original. Ainda no que respeita às marcas de domínios, pretendemos, de futuro, integrar a ontologia de domínios que está a ser elaborada para aplicar ao novo dicionário da ACL (Costa *et al.*, 2020).

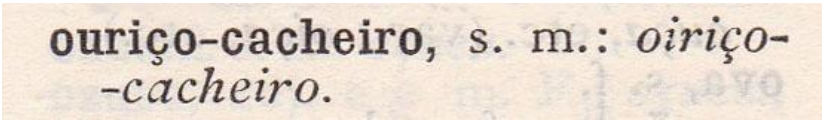
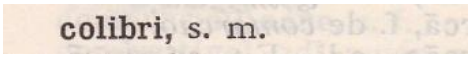

ouriço-cacheiro, s. m.: oiriço--cacheiro.

Figura 11: Entrada **ouriço-cacheiro** do VOLP-1940

```
<entry xml:lang="pt" xml:id="ouriço-cacheiro" type="polylexicalUnit">
   <form type="lemma">
      <orth>ouriço-cacheiro</orth>
      <form type="variant" xml:id="oirico-cacheiro" xml:lang="pt">
         <orth>oiriço-cacheiro</orth>
      </form>
      <pc>,</pc>
      <gramGrp>
         <gram type="pos" norm="NOUN">s.</gram>
         <gram type="gen">m.</gram>
      </gramGrp>
      <usg type="domain">Zool.</usg>
      <note>oiriço-cacheiro</note>
</entry>
```

Exemplo (8): Codificação da entrada **ouriço-cacheiro** do VOLP-1940 enriquecida com novos componentes em TEI Lex-0

Para o enriquecimento linguístico da base de dados, tendo como ponto de partida o DLPC (2001) também poderão ser acrescentadas ao banco de dados relações semânticas (por exemplo, sinónimos, hipónimos), referências cruzadas (por exemplo, remissões). Como representação deste aproveitamento, servimo-nos das entradas «colibri» no VOLP-1940 e VOLP-2018 (Figura 12).



VOLP-1940

nome masculino

VOLP-2018

colibri [kɔlibrí]. *s. m.* (Do caribe *kolibris*, pelo cast. *colibrí*). *Zool.* Designação vulgar de várias aves da família dos troquilídeos (*Trochilus*, Lin.), de tamanho reduzido, plumagem de cores vivas e brilhantes, voo muito veloz, frequentes na América tropical, também conhecidas por *beija-flor*, *chupa-flor*, *chupa-mel* e *pica-flor*.

DLPC 2001

Figura 12: Entrada **colibri** no VOLP-1940, DLPC (2001), VOLP-2018

Entre o VOLP-1940 e o VOLP-2018, a única diferença é a mudança do termo «substantivo» (s.) para «nome». O facto de o VOLP-2018 ser uma edição digital justifica o uso do termo por extenso, e não na sua forma abreviada, como era prática comum nas edições em papel por uma economia de espaço. No entanto, se confrontarmos estas duas entradas com a entrada correspondente no DLPC 2001, verificamos, como seria de esperar, que o dicionário apresenta mais informações – dada a sua natureza discursiva – das quais pretendemos aproveitar a marca de uso (*Zool.*, de Zoologia), bem como os sinónimos *beija-flor*, *chupa-flor*, *chupa-mel* e *pica-flor*. Assim, essas informações podem ser reaproveitadas para a base do VOLP-1940.

```
<entry xml:lang="pt" xml:id="colibri" type="monolexicalUnit">
   <form type="lemma">
      <orth>colibri</orth>
   </form>
   <pc>,</pc>
   <gramGrp>
      <gram type="pos" norm="NOUN">s.</gram>
      <gram type="gen">m.</gram>
   </gramGrp>
   <usg type="domain">Zool.</usg>
   <xr type="synonymy"><ref target="#beija-flor" type="entry">beija-flor</ref></xr>
   <xr type="synonymy"><ref target="#chupa-flor" type="entry">chupa-flor</ref></xr>
   <xr type="synonymy"><ref target="#chupa-mel" type="entry">chupa-mel</ref></xr>
   <xr type="synonymy"><ref target="#pica-flor" type="entry">pica-flor</ref></xr>
</entry>
```

Exemplo (9): Codificação da entrada **colibri** do VOLP-1940 em TEI Lex-0

Os dados de marcas de domínios, bem como os sinónimos, serão acrescentados automaticamente na anotação dos ficheiros do VOLP-1940, com posterior revisão manual e validação. No caso dos sinónimos, em TEI Lex-0, é usado o elemento `<xr type="synonymy">` para anotação dos mesmos, enquanto `ref` estabelece uma referência que aponta para a entrada sinónima que se encontra na mesma fonte, sendo por isso precedida de `#`.

5.2.3.4. Modernização ortográfica

No que respeita à aplicação da nova ortografia, ou seja, a aplicação do *Acordo Ortográfico da Língua Portuguesa* (1990), a correspondência entre as grafias segundo a norma seguida no VOLP-1940 e as do novo acordo ortográfico também se prevê implementar de acordo com a proposta demonstrada em Bański *et al.*, 2017. O elemento `orth` terá como atributo `notBefore="2011”,` que assinala o ano em que o novo acordo entrou em vigor, seguido da referência explícita a essa reforma em `usg type="time"` (Figuras 13 e 14).

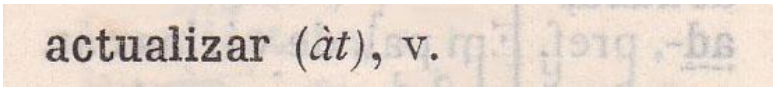


Figura 13: Entrada **actualizar** do VOLP-1940

```
<entry xml:lang="pt" xml:id="actualizar" type="monolexicalUnit">
   <form type="lemma">
      <orth>actualizar</orth>
   </form>
   <pc>,</pc>
   <form type="variant">
      <orth notBefore="2011" xml:lang="pt-PT">atualizar</orth>
      <usg type="time">Acordo Ortográfico de 1990</usg>
   </form>
[…]
</entry>
```

Exemplo (10): Codificação da entrada **actualizar** do VOLP-1940 em TEI Lex-0

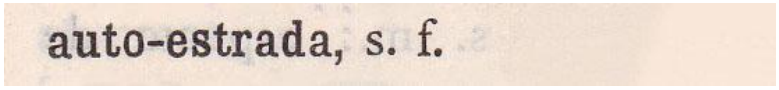


Figura 14: Entrada **auto-estrada** do VOLP-1940

```
<entry xml:lang="pt" xml:id="auto-estrada" type="polylexicalUnit">
   <form type="lemma">
      <orth>auto-estrada</orth>
   </form>
   <pc>,</pc>
   <form type="variant">
      <orth notBefore="2011" xml:lang="pt-PT">autoestrada</orth>
      <usg type="time">Acordo Ortográfico de 1990</usg>
   </form>
[…]
</entry>
```

Exemplo (11): Codificação da entrada **auto-estrada** do VOLP-1940 em TEI Lex-0

6. Considerações finais e trabalho futuro

O presente artigo apresenta os primeiros avanços de um projeto ambicioso, já em curso, para a criação de novo recurso lexicográfico português *online* que reunirá as versões impressas dos vocabulários do ACL (1940, 1947, 1970, 2012). Nesta primeira fase, visamos a criação de uma edição digital do primeiro vocabulário ortográfico académico português: o VOLP-1940. Após a digitalização da obra, uma amostra de entradas foi anotada, recorrendo à metalinguagem XML e conforme as diretrizes de anotação da TEI. Os resultados demonstram que a codificação, embora mais pormenorizada em TEI Lex-0, é mais estrutural e rigorosa, permitindo que os sistemas processem melhor os dados anotados.

Muitos dos princípios agora definidos e adotados serão usados como guia para anotação das restantes entradas e para aplicação às obras subsequentes, uma vez que partilham várias das convenções tipográficas agora identificadas. Com este processo de retrodigitalização de obras lexicográficas de referência e a aplicação da respetiva metodologia, pretendemos representar a crescente sinergia entre lexicógrafos, terminólogos, linguistas computacionais e humanistas digitais que tanto defendemos. Segundo a nossa perspetiva, a organização da informação metalinguística tem de ser obrigatoriamente aliada a um rigoroso processamento

linguístico dos dados, com dados lexicais de alta qualidade, estruturados e regidos por princípios teóricos e metodológicos que respondam às necessidades do século XXI, interoperáveis com soluções computacionais, e no qual, na nossa perspetiva, os trabalhos lexicográfico e terminológico podem ser complementares (Costa, 2013).

No contexto português, esta investigação vem preencher uma lacuna em relação a obras lexicográficas retrodigitalizadas *on-line* pesquisáveis, baseadas em padrões e metodologias atuais que promovem a partilha e a harmonização de dados. De futuro, é nossa intenção ainda garantir a conexão com outros sistemas e recursos lexicais académicos como exteriores. No final do projeto, esperamos ter codificado um vocabulário com um valor patrimonial significativo, compatível com os padrões mais avançados para edições digitais académicas e de acesso aberto. As versões serão pesquisáveis por meio de uma *interface* de pesquisa avançada.

Acreditamos que este projeto irá contribuir significativamente para a análise e anotação de recursos lexicais portugueses através do uso de processos assistidos por computador, permitindo repensar a forma de conceber novos produtos lexicográficos genuinamente digitais, e não como simples reprodução de edições em papel, que responderão de maneira mais eficaz às necessidades dos utilizadores finais.

Agradecimentos

Este artigo tem o apoio do financiamento nacional português, através da FCT – Fundação para a Ciência e Tecnologia –, como parte do projeto estratégico do Centro de Linguística da Universidade NOVA de Lisboa (UID/LIN/03213/2020), e do programa de investigação e inovação Horizonte 2020 da União Europeia, sob a referência n.º 731015 (ELEXIS).

Referências:

Academia Brasileira de Letras (2017) *Vocabulário Ortográfico da Língua Portuguesa*, Bechara, E. (coord.), 6.ª edição. Rio de Janeiro: Academia Brasileira de Letras https://voc.cplp.org/index.php?action=von&csl=br).

Academia das Ciências de Lisboa (1940) *Vocabulário Ortográfico da Língua Portuguesa* (VOLP-1940) (1940). Lisboa: Imprensa Nacional.

Academia das Ciências de Lisboa (2001) *Dicionário da Língua Portuguesa Contemporânea*. João Malaca Casteleiro (coord.), 2 vols. Lisboa: Academia das Ciências de Lisboa & Editorial Verbo.

Anselmo, A. (2011) Vocabulários da Língua Portuguesa editados em Portugal (1866–1970). In *Vocabulário Ortográfico Actualizado da Língua Portuguesa*. Academia das Ciências de Lisboa. Lisboa: Imprensa Nacional Casa da Moeda.

Bański, P., Bowers, J., Erjavec, T. (2017) TEI-Lex0 guidelines for the encoding of dictionary information on written and spoken forms. *Electronic Lexicography in the 21st Century: Proceedings of eLex 2017 Conference*, pp. 485–494.

Bergenholtz, S., Nielsen, S. Tarp. (eds.) (2009) *Lexicography at a Crossroads. Dictionaries and Encyclopedias Today, Lexicographical Tools Tomorrow*, pp. 165–194. Bern: Peter Lang AG.

Correia, M. (2009) *Os Dicionários Portugueses*. Coleção: O Essencial Sobre Língua Portuguesa. Lisboa: Editorial Caminho.

Costa, R. (2013) Terminology and specialised lexicography: two complementary domains. *Lexicographica*, 29, Issue 1 Internationales Jahrbuch für Lexicographie. Ed. by Gouws, Rufus Hjalmar / Heid, Ulrich / Schierholz, Stefan J. / Schweickard, Wolfgand / Wiegand, Heribert Ernst. Berlin, New York: De Gruyter, pp. 29 – 42.

Costa, R., Carvalho, S., Salgado, A., Simões, A., Tasovac, T. (2020, no prelo) Ontologie des labels de domaines

appliquée aux dictionnaires de langue générale. In *La lexicographie en tant que méthodologie de recherche en linguistique*. *Langue(s) et Parole – Revue de Philologie Française et Romane 5*.

Fajardo, A. (2018) Lexicografía histórica con corpus y recursos digitales: aspectos metodológicos. In D. Corbella, A. Fajardo y J. Langenbacher-Liebgott (eds.). *Historia del léxico español y Humanidades digitales*. Berlim: Peter Lang.

Gold, K. M., Klein L. F. (eds.) (2015) *Debates in the Digital Humanities*. Mineápolis: University of Minnesota Press.

L'Homme, M.-C., Cormier, M. C. (2014) Dictionaries and the Digital Revolution: A Focus on Users and Lexical Databases. In *International Journal of Lexicography*, Volume 27, Issue 4, dezembro 2014, pp. 331–340.

Marquilhas, R., Hendrickx, I. (2016) Avanços nas humanidades digitais. In A. M. Martins e E. Carrilho (eds.) *Manual de Linguística Portuguesa*, MRL Series. De Gruyter, pp. 252–277.

Rio-Torto, G. (coord.) (2013) *Gramática Derivacional do Português*, Imprensa da Universidade de Coimbra.

Romary, L., Tasovac, T. (2018) TEI Lex-0: a target format for TEI-encoded dictionaries and lexical resources. In *Proceedings of the 8th Conference of Japanese Association for Digital Humanities*, pp. 274–275.

Salgado, A., Costa, R., Tasovac, T., Simões, A. (2019) TEI Lex-0 In Action: Improving the Encoding of the Dictionary of the Academia das Ciências de Lisboa. In I. Kosem *et al.* (eds.), Electronic lexicography in the 21st century. P*roceedings of the eLex 2019 conference*, pp. 417–433, 1–3 outubro 2019, Sintra, Portugal. Brno: Lexical Computing CZ, s.r.o. Retrieved from: https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_23.pdf.

Simões, A., Salgado, A. Costa, R. & Almeida, J. J. (2019) LeXmart: A Smart Tool for Lexicographers. In I. Kosem et al. (eds.), *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference*, pp. 453–466, 1-3 October 2019, Sintra, Portugal. Brno: Lexical Computing CZ, s.r.o. Retrieved from: https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_25.pdf.

Simões, A., Sanromán, Á. I., Almeida, J. J. (2012) Dicionário-Aberto: A Source of Resources for the Portuguese Language Processing. In H. Caseli, A. Villavicencio, A. Teixeira, F. Perdigão (eds.), *Computational Processing of the Portuguese Language. PROPOR 2012. Lecture Notes in Computer Science*, vol 7243. Berlim, Heidelberg: Springer.

Stührenberg, M. (2012) The TEI and Current Standards for Structuring Linguistic Data. *Journal of the Text Encoding Initiative* [Online], issue 3, novembro 2012. URL: http://journals.openedition.org/jtei/523.

Tasovac, T. (2010) Reimagining the Dictionary, or Why Lexicography Needs Digital Humanities. *Digital Humanities 2010*. Disponível em: http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/pdf/ab-883.pdf.

Tasovac, T., Romary, L., Bánski, P., Bowers, J., Does, J. de, Depuydt, K., Erjavec, T., Geyken, A., Herold, A., Hildenbrandt, V., Khemakhem, M., Petrović, S., Salgado, A., e Witt, A. (2018) *TEI Lex-0: A baseline encoding for lexicographic data*. Version 0.8.5. DARIAH Working Group on Lexical Resources. https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html.

TEI Consortium (ed.): *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. TEI Consortium, http://www.tei-c.org/Guidelines/P5/.

Terras, M., Nyhan, J., Vahouette, E. (eds.) (2013) *Defining Digital Humanities: A Reader*. Londres: Ashgate.

Trap-Jensen, L. (2018) Lexicography between NLP and Linguistics: Aspects of Theory and Practice. In *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*. Ljubljana: Ljubljana University Press, Faculty of Arts, pp. 25–37.

Williams, G. (2019) The problem of interlanguage diachronic and synchronic markup. In A. Villalva e G. Williams (eds.), *The Landscape of Lexicography*. Lisboa–Aveiro: Centro de Linguística da Universidade de Lisboa–Universidade de Aveiro.