9 SKOS as a key element for linking lexicography to digital humanities

Rute Costa, Ana Salgado, and Bruno Almeida

Introduction

The humanities, where traditionally dictionaries fit into, have undergone significant changes in recent years regarding the production, research, publication, dissemination, preservation and sharing of information. Nowadays, the concept of "digital humanities" is characterised by associating the field of traditional humanities with computational methods, encompassing computation for the humanities, computational linguistics (Hockey 2004; Gold and Klein 2019) and ontologies, among others. Like digital humanities and lexicography, information science is contributing to the effort of sharing information, transferring its reference objects, thesauri, into digital versions in SKOS format. In the introduction to ISO 25964-1, 2011 one can read: "Today's thesauri are mostly electronic tools, having moved on from the paper-based era when thesaurus standards were first developed." (6). Historical dictionaries are going in the same direction – from paper to digital – requiring standards and tailored software to ensure effective interoperability.

The research is relevant because we believe that this project will contribute significantly to the analysis and annotation of Portuguese lexical resources using computer-assisted processes. It will allow us to rethink how to design new lexicographical products that are not merely a simple reproduction of paper editions, which will respond more effectively to the needs of the end-users. While digitisation signalled the modification of a paradigm, the spread of the Web has shaped a new concept for lexicographic works. Today we can create dynamic, more robust lexicons enriched with semantic, conceptual and statistical information and take advantage of Linked Data, highlighting the notion of content models and data mining by joining digital humanities and lexicography. Generating or re-digitising lexicographic products has linguistic, heritage and historical relevance, contributing to the establishment of the lexicon of a language at a given time, around which the identity of a linguistic and cultural community is built and preserved.

VOLP-1940, an ongoing project, is the first of a series of orthographic vocabularies published by the Lisbon Academy of Sciences (ACL) to be

DOI: 10.4324/9781003131816-17

digitised to create a lexicographical corpus. The digitisation of VOLP-1940 aims to allow its computational processing by creating a lexicographical resource encoded in Text Encoding Initiative (TEI P5), with structured information in Simple Knowledge Organisation System (SKOS), and in line with the Findable, Accessible, Interoperable, Reusable (FAIR) principles (see FAIR Principles). This will serve to guarantee its future connection to other systems and resources, in particular in the Portuguese-speaking world. This research also aims to fill a gap in Portuguese lexicography, given that legacy dictionaries are still rare online (Williams 2019, 83). These resources need to be encoded and published on the web, based on current standards and methodologies that enable data sharing and harmonisation as well as their alignment with existing lexical resources.

This chapter falls within the domain of the application of digital lexicography in the context of a scholarly editing project and is based on a set of methodological and theoretical assumptions for which we will make some considerations. We will focus on the organisation of linguistic information of a lexicographical nature within the field of digital humanities, emphasising the development of a cross-disciplinary methodology that combines lexicography and information science. Our goal is to attest the strong relationship between lexicographic practice, dictionaries and digital humanities, where we include information science (Robinson, Priego, and Bawden 2015).

We aim to build a digital lexicographical corpus bringing together the publicly available printed versions of ACL vocabularies (1940, 1947, 1970, 2012), and improving multiple search functionalities, as a source of scientific research and cultural heritage, especially on the evolution of Portuguese language and culture. Underlying this goal, we have a central research question: how could digital humanities integrate annotated dictionaries in a wider community, contributing and intervening in collaborative information organisation, search and retrieval in digital cultural heritage collections? The second main question related to the previous one concerns the standards: how could we join efforts to make different standards coming from different communities, such as SKOS and TEI, becoming more effective, contributing to the operationalisation of vocabularies?

This rest of this chapter is organised as follows. The section titled Background provides an overview on the relation between lexicography and information science as part of the digital humanities and existing standards. The section titled Case Study is dedicated to the *Vocabulário Ortográfico da Língua Portuguesa* (VOLP-1940; Orthographic Vocabulary of the Portuguese Language; see Academia das Ciências de Lisboa, 1940). After presenting our lexicographical case study, we describe the structure of the vocabulary, focusing on the macrostructural and microstructural main components and continue with a proposal of modelling in SKOS(-XL) and encoding in TEI Lex-0. Finally, we highlight our future work and present concluding remarks.

Background

Lexicography and information science as part of digital humanities: A brief overview

The field of lexicography, currently defined as the "total of all activities directed at the preparation of a lexicographic reference work" (Wiegand et al. 2020, 224 Wiegand et al. 2020, 224), aims to produce a great variety of resources, namely, dictionaries, vocabularies, glossaries and encyclopaedias. However, while a variety of lexicographical works were still being published on paper at the start of the 21st century, this scenario has changed radically over the past two decades. This is especially due to the ongoing transition to digital, the downfall of many renowned publishers and the changes introduced to editorial business models (Rundell 2010, 170). Terms such as "online dictionaries" and "e-lexicography" started to appear but were soon replaced by "digital dictionaries" and "digital lexicography". This change in terminology has led to a paradigm shift directly related to the advancement of the field of digital humanities, which has quickly become a catalyst for academic research in the interface between humanities and computation. While early definitions of digital humanities were limited to the humanities computing (Terras et al. 2013), today, its definition is far from reaching a consensus (Gold and Klein 2019). This is because it covers a wide variety and assortment of works from different branches of knowledge that are characterised by the use of tools, digital methods and standards to ensure the long-term growth of the Web, primarily implying a new look at the humanities in general.

Within digital humanities, we can find lexicography and its products, that is, lexicographical reference works. Dictionaries must be converted into digital resources to enable information retrieval on the Web. This transformation must be adequately addressed to optimise access to linguistic and lexicographical information until the dictionaries become actual digital resources. On the other hand, dictionaries are also cultural objects whose heritage must be preserved and made available to the entire community. Our research focus is to undertake a precise linguistic analysis and description of the object-language, that is, a language that is the object of study in various fields, and to organise linguistic data (e.g., linguistic variants, grammatical information and domain labels, among others) according to the microstructure of the lexicographical articles, namely the dictionary entry (the part of a dictionary that contains information related to one lemma and its variants (ISO 1951 2007)) specific to each dictionary model.

Another field that stands out within digital humanities, and is of interest to our research, is information science, an interdisciplinary field concerned with "the origination, collection, organisation, storage, retrieval,

interpretation, transmission, transformation and use of information" (Borko 1968, 3). Information includes all encoded representations (in natural language or other modalities) that can be transmitted, stored and organised for subsequent retrieval. As Saracevic (1999) noted, the study of information involves not only the encoded messages and their interpretation and processing but also the wider social context in which information is used. Born out of the so-called "information explosion" of the post-WW2 period, information science became a necessity in the newly formed information and knowledge societies, wherein information retrieval methods and technologies are paramount.

The ties between terminology science and information science were noted from the beginnings of terminology as a contemporary subject of inquiry. As one of the early proponents of terminology as a discipline in its own right noted, terminologies are fundamental for "the storage and retrieval of scientific and technical information" (Felber 1984, 1), including applications such as thesauri and the classification schemes. The ties between information science and lexicography remain less obvious. Lexicography has traditionally been understood as the art and craft of compiling general language dictionaries (Landau 2001) and is often seen as a branch of applied linguistics. However, there is a more holistic approach that embraces lexicography's relationships with lexicology, terminology, encyclopaedias and information science. According to this broader view, metalexicography "should be regarded as part of information science" (Wiegand 2013, 14). More than describing the lexicon of languages, the purpose of lexicography is to "resolve specific types of information needs detected in society" (Tarp 2018, 22). Indeed, it can be argued that lexicography is aimed "in a more general way at the production of information tools" (Bergenholtz and Gouws 2012, 40), that is, reference works currently focused on "enhanced information retrieval" (ibid.). The ties between lexicography and information science have also been noted in the latter community, especially in the context of digital lexicographical research based on end-user information needs and access to lexicographical data (Bothma 2018). Knowledge organisation (KO), a subfield of information science, is especially relevant for drawing relationships between information science and lexicography. KO is concerned with the activities of document description, indexing and classification (usually referred to as KO processes) carried out in information services, such as libraries and archives, as well as with the knowledge organisation systems (KOS) employed to carry out such activities (Hjørland 2008). The latter include widely different resources, ranging from flat term lists to structured resources, such as thesauri and ontologies (Hodge 2000; Zeng 2008). Contrary to KOS, the traditional products of lexicography and terminology are not aimed at facilitating information retrieval through the KO processes mentioned above. Instead, the structuring of knowledge present in lexicographical products aims to facilitate the retrieval of information

about the words and senses of one or more languages, e.g., through the use of lists of abbreviations representing lexicographical categories (as will be shown in the section titled Case Study with VOLP-1940). Terminological products, on the other hand, structure knowledge through concept systems based on generic, partitive and associative relations between concepts in specialised domains (ISO 1087, 2019). Therefore, terminologies are very similar to thesauri for information retrieval (ISO 25964-1, 2011; ISO 25964-2, 2013), although the former aim at improving specialised communication, while the latter are focussed on retrieving indexed information resources. Despite these differences, dictionaries, glossaries and other terminological products may also play a role in information retrieval, e.g., for extending thesauri (as a source for concepts, terms and scope notes) or complementing them in full-search applications (ISO 25964-2, 2013 §22.3).

Standards

Conceiving digital lexicographical resources increasingly requires the application of adapted standards and tools capable of guaranteeing the availability of structured data and ensuring interoperability between systems. To change a raw document into a structured one, it is necessary to define the different types of data that make up the document for modelling it according to a standardised data model, which makes interoperability feasible. Interoperability is (from manuscripts to poems, dictionaries, culinary recipes, corpora annotation and many others) despite not having the legal status of a standard (Stührenberg 2012). Interoperability is the "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" (ISO/IEC 2382, 2015). While the conversion of printed dictionaries signalled a paradigm shift, the dissemination of the Web has forced us to rethink the concept of lexicographical work. More than ever, we must learn how to take advantage of and explore the possibilities of the digital environment (Trap-Jensen 2018) by creating dynamic and robust lexicons augmented with semantic, conceptual and statistical information, wherein data from different resources can be interconnected (Linguistic Linked Open Data Cloud 2021). Although a reasonable number of Portuguese lexicographical works can currently be consulted online, these resources end up being static; hence, there is a need for some sort of icebreaker.

As Tasovac (2010, 1) stated, "we cannot think of dictionaries any more without thinking about digital libraries and the status which electronic texts have in them". Keeping in mind this new reality, we propose to apply new principles, that is, computational methods, interoperable standards and semantic technologies that facilitate the organisation of large amounts of lexical data. These methods, standards and technologies will be further described below.

De facto standard: Text encoding initiative (TEI) and TEI Lex-0

For lexical data annotation, TEI has become a *de facto* international standard for the encoding of different types of documents (manuscripts, poems, dictionaries, culinary recipes, annotated corpora and many others). TEI was created in 1987 by a consortium of several institutions, the TEI Consortium, to develop a standardised format for the electronic edition of textual content in multiple formats. It presents a metalanguage comprising a vocabulary (a set of elements and attributes) and a grammar (a schema) to annotate, structure and validate documents, whose specific syntax and semantics in Extensible Markup Language (XML) make it a textual analysis method for digital processing.

The current version of the *TEI Guidelines* (TEI Consortium) continues to be the subject of constant updates. In our case study, we chose to follow this standard format because it is commonly used to share lexicographical data and ensures the digital preservation of the dictionaries and their interoperability. The complexity of lexicographical resources has been recognised by the scientific community (Salgado et al. 2019), both because of the diversity of its structural components and as different resources follow different criteria for the representation and processing of lexicographical information.

The most recent version of the *TEI Guidelines* is known as P5. These guidelines have a specific module for dictionaries: Chapter 9. Here too, the word "dictionaries" is taken in its most general sense, that is, encompassing not only dictionaries but also, as previously mentioned, vocabularies, encyclopaedias and glossaries. Since the *TEI guidelines* are characterised by their highly flexible annotation potential – several encoding possibilities for the same elements, which poses an obstacle for interoperability – TEI Lex-0, a new, simplified TEI sub-format for dictionaries (in the broad sense of the term) is being developed specifically to encode lexical resources, the application of which will be detailed later in this chapter.

The groundwork for this format started in 2016 and is currently led by the Digital Research Infrastructure for the Arts and Humanities (DARIAH n.d.) Lexical Resources Working Group. TEI Lex-0 aims to define a clear and versatile annotation structure, albeit not too permissive, to facilitate the interoperability of heterogeneously encoded lexical resources. TEI Lex-0 should be regarded as "a format that existing TEI dictionaries can be unequivocally transformed to, in order to be queried, visualised or mined uniformly" (Tasovac et al. 2018). As the layout of this format has not been finished yet, we have been actively contributing to its development by raising issues on GitHub.

W3C recommendation for the semantic web: SKOS

SKOS is a model for sharing and linking KOS, such as thesauri, taxonomies, classification schemes and other structured and controlled vocabularies available on the Web (Baker et al. 2013). The model is expressed as

an ontology in Web Ontology Language (OWL), which enables the modelling of controlled vocabularies as Resource Description Framework graphs (RDF), as well as their mapping to external resources and integration in the Linguistic Linked Open Data Cloud (Linguistic Linked Open Data Cloud n.d.). The early developments that have led to SKOS started in the late 1990s and early 2000s in the context of several European projects focused on improving the browsing and discoverability of Web resources. SKOS answered the need for a common RDF schema for modelling thesauri, a type of knowledge organisation system and defining inter vocabulary mappings. The model became a World Wide Web Consortium (W3C) recommendation in 2009 (Miles and Bechhofer 2009). SKOS is widely used by the information science community for publishing KOS in the Semantic Web though its mostly suited for thesauri. A few notable examples include the EU Vocabularies (EU Vocabularies n.d.) and the Getty Art and Architecture Thesaurus (Art & Architecture Thesaurus Online).

The central units of SKOS are concepts, which are informally defined as ideas or notions, typically represented in thesauri, taxonomies and other KOS for information retrieval. Among other possibilities, the model allows for concepts to be identified with URIs, lexicalised with multilingual labels (preferred, alternative, and hidden), documented with notes, linked to other concepts through conceptual relations (broader, narrower or associative) and mapped to concepts in external resources. While the core SKOS model only allows for relations between concepts, the SKOS-XL extension has brought support for modelling relations between concept labels. The latter include the relations between abbreviations and their full forms (e.g., between "EU" and "European Union"), which will be exemplified later in this chapter concerning the modelling of lexicographical information.

Both standards, TEI and SKOS, have been applied to the VOLP-1940 following a precise methodology described below based on the relationship between linguistic and lexicographical knowledge and information science.

Case study: Vocabulário ortográfico da língua portuguesa (VOLP-1940)

This section is structured around research issues related to VOLP-1940. After presenting our lexicographical case study, we describe the structure of the vocabulary, focusing on the macrostructural and microstructural main components. The next subsection is devoted to front matter analysis. The two subsequent subsections are dedicated to modelling in SKOS(-XL) and encoding in TEI Lex-0.

General considerations on the VOLP-1940

The case study presented in this chapter is the digital conversion of the paper edition of the first Portuguese Academy vocabulary of a series of subsequent vocabularies – 1947, 1970 and 2012 – published in 1940. The

document is named *Vocabulário Ortográfico da Língua Portuguesa* (VOLP-1940), published by Imprensa Nacional de Lisboa with the seal of the ACL in a volume of 821 pages.

The spelling proposed in this work was governed by the 1911 spelling reform, backed by two other elements: the 1920 spelling reform, which changed some provisions of the 1911 reform, and the 1931 Portuguese-Brazilian Orthographic Agreement, signed by the Portuguese and Brazilian academies. This lexicographical work has immense historical and linguistic value since it served as the basis for the ACL and the Brazilian Academy of Letters to discuss a new orthographic measure that came to result in the Portuguese-Brazilian orthographic convention of 1945, commonly known as the Orthographic Agreement of 1945, which was in force until 2011.

We aim to do the following: (i) create a new online lexicographical resource, accessible to the entire scientific community and the general public; (ii) work on the metadata providing consistency, following an exact linguistic annotation strategy in line with TEI recommendations, while ensuring the data are accessible and reusable; (iii) organise metadata information according to SKOS; (iv) describe the linguistic annotation for further semantic enrichment of the database and (v) add new metadata information, namely, domain names and information that will be recovered from other lexicographical works that contain this annotation, and make the connection between several synonymous units that are included in the work's word list. The tasks described above are necessary for improving information retrieval within VOLP-1940 by scholars in linguistics and digital humanities, as well as for ensuring the interoperability of our dataset with third-party systems.

With the publication of the VOLP-1940, the ACL intended to establish the official spellings of Portuguese words in their national variety, having become a "referência normalizadora para a fixação da nomenclatura em quase todos os dicionários escolares e práticos publicados após a sua divulgação" (standardising reference to establish the [Portuguese] vocabulary in almost every academic and practical dictionary published after its dissemination; Verdelho, 2007).

Since the VOLP-1940 is organised around two structures, its macrostructure and microstructure, our research also focuses on these two parts separately. Rey-Debove (1971) envisioned the macrostructure as the list of every word that is described in a dictionary, while the microstructure refers to the information provided about each lexical unit, that is an "unit of language, belonging to the lexicon of a given language and which is described or mentioned in a dictionary" (ISO 1951, 2007).

The VOLP-1940 macrostructure

In macrostructural terms, the list of entries "covers only the modern Portuguese language, i.e., the linguistic period that runs from the 16th

century to the present time [i.e., 1940]" (Academia das Ciências de Lisboa, 1940, p. XII), registering lexical units that entered the language after 1500 and leaving out units "pertencentes ao período arcaico do idioma" (that belong to the archaic period of the language; Academia das Ciências de Lisboa, 1940, 12).

The preliminary pages present a dedication and an "Introdução" (Introduction; 9–86), prefaced by Francisco Rebelo Gonçalves (1907–1982), one of the great Portuguese philologists of the 20th century in Portugal. The introduction consists of three chapters: "Preliminares" (Preliminaries), "Normas da escrita portuguesa" (Standards of Portuguese spelling) and "Comentários ortográficos" (Spelling comments).

The VOLP-1940 is further divided into three main parts, namely 1) common vocabulary, 2) onomastic vocabulary and 3) registration of abbreviations.

1 COMMON VOCABULARY (3–713) of the "léxico geral da língua descontados os nomes próprios" [general lexicon of the language excluding proper names], including elements of composition (9);
2 ONOMASTIC VOCABULARY (717–809), "nomes próprios de várias categorias" [proper names of various categories] (9), such as anthroponyms, toponyms and patronyms, as well as ethnonyms, hieronyms (sacred names), mythonyms, chrononyms (calendar names) and biblionyms;
3 REGISTRATION OF ABBREVIATIONS (appendix), commonly used at the end of the 1930s (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" [Portuguese and other non-Portuguese abbreviations that are used in our writing […] the abbreviations of greatest importance for current uses and of greatest general interest for the two Portuguese-speaking countries] (9).

The lexical units that comprise the entry words of the VOLP-1940 are organised into three columns per page, listed alphabetically, and are followed by various classifications, such as grammatical information and pronunciation information, among others, as we will demonstrate in the next subsection.

The microstructure of the VOLP-1940

In microstructural terms, a lexicographical article from the VOLP-1940 may, as a rule, include the following elements:

1 Lemma: It is a "lexical unit, chosen according to lexicographical conventions to represent the different forms of an inflection paradigm" (ISO 1951, 2007). In this vocabulary, it corresponds to the singular form of the noun or adjective and the masculine form when there is gender inflection in variable words. In the case of verbs, it

corresponds to the form of the impersonal infinitive. It should be noted that the elements of composition, that is, "todo o elemento que se baseie etimologicamente num tema nominal, pronominal, ou verbal, qualquer que seja o seu lugar no composto" (any element that is etymologically based on a nominal, pronominal, or verbal base, whatever its place in the compound) (21), for instance, "mono-" and "-grafia", also appear in the word list. In this case, the base is followed by a hyphen (geo-) or preceded by a hyphen (-mente), followed by the indication "el. comp." (composition element) and a descriptive text of the employment of this element, providing examples at the end to illustrate the application of the spelling rule that is usually stated. There are also notes on spelling variants, for instance, "cenoura" and "cenoira" (carrot). Variants of the canonical form do not normally feature in the word list, e.g., "cenoira" does not appear in the list of entries but can only be found in the lexicographical article "cenoura". There are some exceptions to this criterion that are explained in the Introduction (18), such as "cousa" and "coisa" (thing). In such cases, whenever the variant is more usual than the basic form, it also features in the word list.

2 Orthoepy: The standard indication of the pronunciation of a lexical unit, which appears in parentheses after the base, and only in words of doubtful pronunciation. When it is not marked graphically, the pitch of the closed stressed vowels "e" and "o" can also be provided. Additionally, particular stressed vowels that are often pronounced incorrectly will also be marked. On the matter of orthoepy in Portuguese, see section 4 of the paper "Orthography and Orthoepy" (Gonçalves 2020, 651–677).

3 Part of speech: "A category assigned to a lexical unit based on its grammatical and semantic properties" (ISO 1951, 2007), which appears after the base or orthoepy when marked and is indicated in abbreviated lowercase. In the part corresponding to proper names, the classifications are onomastic, for instance, anthroponym (antr.) and toponym (top.). Further, although they are not parts of speech, this information is provisionally encoded in this field for practical reasons; this issue is being debated by the "Lexical Resources" DARIAH Working Group.

4 Gloss: Understood as "a textual description of a sense's meaning" (Salgado et al. 2020), it appears only to disambiguate cases of homonymy, to which a number is added (1, 2 etc.), superscripted on the right-hand side of the base as a way of distinguishing them. Consider, e.g., "afecto[1] (*ét*) s. m.: afeição" [affection] and "afecto[2] (*ét*) adj.: afeiçoado" [attached].

There is also information about words that are almost exclusively used in phrases. For example, when a particular word is only used in a particular

phrase, this indication appears as an entry in what is considered the core word of that phrase – for instance, "cavalitas, el. nom. f. pl. na loc. adv. mod. às cavalitas" (riding piggyback, plural feminine noun element).

Another indication of a prescriptive nature concerns constructions that begin with the expression "Melhor que" (Better than). The forms indicated as preferable are those that are considered to be closest to their origin or more correct for certain reasons, such as "canon" and "cânone" – "cânone, s. m. Melhor que canon" (cânone [canon], s. m. better than canon [Portuguese orthographic variant of the first form]). So far, we have identified the essential and most relevant elements of the VOLP-1940's microstructure. This analysis is crucial for the linguistic annotation phase discussed below.

The list of abbreviations and conventional signs

Now, we move on to the analysis of the front matter, specifically, the list of abbreviations. First, we describe the content and then, we focus on the modelling of the lexicographical data using SKOS. To conclude, we exemplify the encoding of a lexicographical article with TEI Lex-0.

On the initial pages of the VOLP-1940, in the front matter materials, a "Lista de abreviaturas e sinais convencionais" (List of abbreviations and conventional signs; 89–92) can be found. In this study, we focus on organising this list for computational processing using SKOS and TEI Lex-0 to ensure the interoperability that will be necessary, in the future. In the paper version, this list is sorted alphabetically and divided into two parts: (i) List of abbreviations and (ii) List of conventional signs. The list shows the abbreviations or conventional signs followed by their full form. Our analysis is anchored on the first part, from which we draw up a classification of the 220 abbreviations that comprise the list. Although this list is well organised into two columns, it is static and has some limitations inherent to the paper format. From this simple alphabetical list, whose original page is retained on the website of this project, we proceeded to its organisation and representation for the digital environment as well as its linguistic annotation.

After a thorough analysis of the abbreviations that make up the list of abbreviations in the VOLP-1940, the following types have been identified: part of speech; onomastic classification; grammatical gender; grammatical number; language; register; tense; etymology; word-formation and others (see Appendix). Thus, these categories constitute what we call the typological organisation of the list of abbreviations. In the transition from paper to digital, we had to reorganise the content of this list to be able to process it and ensure its future interoperability. Therefore, from the total list of abbreviations, we isolated those related to word classes. Based on this list, and for interoperability with other lexicographical resources,

Table 9.1 Sample matches between the VOLP-1940 word classes and Universal Dependencies Part-of-Speech values

VOLP-1940	*Universal dependencies Part-of-speech*	*Universal POS tags*
	OPEN CLASS WORDS	
adjetivo (adj.)	*adjective*	ADJ
advérbio (adv.)	*adverb*	ADV
interjeição (inter.)	*interjection*	INTJ
substantivo (s.)	*noun*	NOUN
verbo (v.)	*verb*	VERB
	CLOSED CLASS WORDS	
artigo (art.)	*determiner*	DET
conjunção (conj.)	*coordinating conjunction*	CCONJ
	subordinating conjunction	SCONJ
numeral (num.)	*numeral*	NUM
preposição (prep.)	*adposition*	ADP
pronome (pron.)	*pronoun*	PRON

we made the correspondence between word classes and the values of the Universal Dependencies Part-of-Speech (Universal Dependencies n.d.), a framework for consistent annotation of grammar, which will be exemplified in Table 9.1. The indication of *part of speech* (morphological categories and subcategories) is used in the "Common vocabulary" part. This indication provides information not only concerning the category of a lexical unit (e.g., pronoun, numeral, adverb or conjunction) but also its subcategory; for instance, there are specific labels for adverbs, such as "adv. af." (assertion adverb) or "adv. conf." (confirmation adverb). Sometimes, there is also some classifying information in this part, such as "phrase", that does not belong to a part of speech. On the other hand, in "Onomastic vocabulary", to differentiate the onomastic forms that make up the word list of this part according to the type of entities they apply to, traditional labels are used, which constitute what we call onomastic classification.

Abbreviations are also used, which are related to the indication of grammatical gender, namely, "m." (masculine), "f." (feminine) or "2 gen." (both genders); the indication of grammatical number, namely, "pl." (plural), "sing." (singular) or "2 núm." (both numbers); tense indications; etymology; word formation and abbreviations related to word-formation processes; and others. The last element is a set of abbreviations that we have not classified because they are not particularly interesting for the present research.

In addition to the abbreviations used to mark word classes, we also found abbreviations that refer to the language. This information is used in the

VOLP-1940 to identify the source language of a particular word; therefore, we also mapped these abbreviations to Tags for Identifying Languages (IETF BCP 47 n.d.), which is a set of codes to identify human languages. Tags are generally used to indicate the language of the content in a standardised way; e.g., "croché" is identified as the Portuguese version of the French "crochet", and the code used for the abbreviation "fr." (of French) in this case matches the abbreviation used in the VOLP-1940. The *register* label, defined by the standard (ISO/TR 20694, 2018, the ISO standard that gives the general principles for language registers in both descriptive and prescriptive environments) as "language register, language variety used for a particular purpose or in an event of language use, depending on the type of situation, especially its degree of formality", is also used; e.g., "ant." (old), "arc." (archaic) and "pop." (popular).

Modelling in SKOS(-XL)

After a careful analysis of the structure of the VOLP-1940, we will now move on to the first stage of modelling the list of abbreviations in SKOS. Figure 9.1 below shows the overall model of the lexicographical categories used for organising the list of abbreviations. In the examples shown

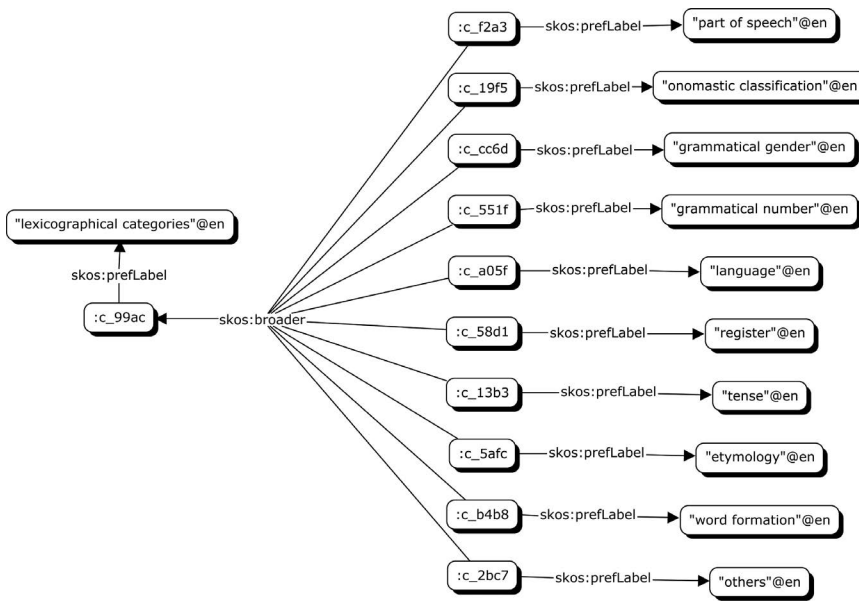

Figure 9.1 Lexicographical categories for modelling the list of abbreviations in SKOS.

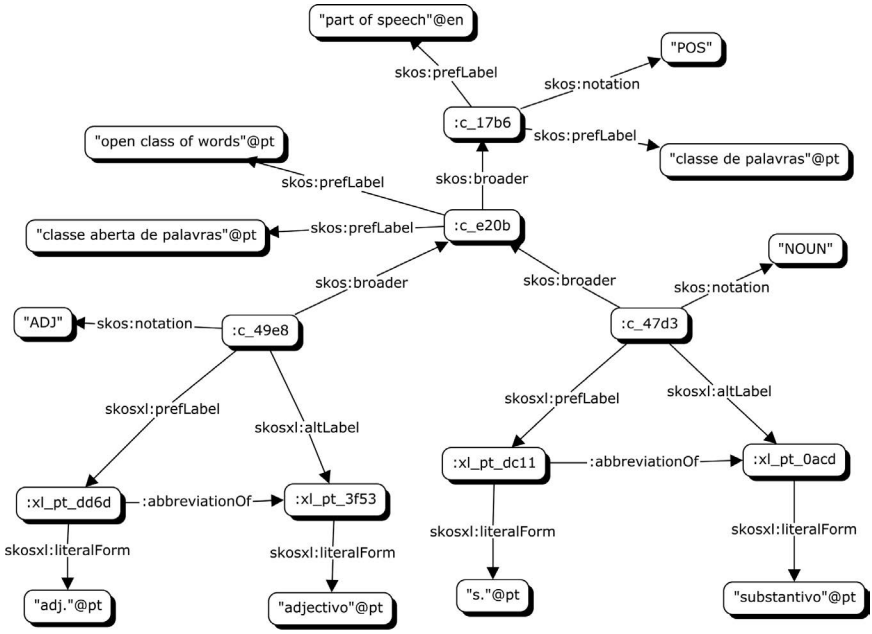

Figure 9.2 Part of speech abbreviations in SKOS (noun and adjective).

in this subsection, SKOS concepts and SKOS-XL labels are identified with URI placeholders (e.g., ":c_17b6", ":xl_pt_3f53"). We have specified the above-mentioned categories based on examples of part of speech and language concepts.

Figure 9.2 below shows the modelling of the noun and adjective concepts ("substantive" and "adjective" in Portuguese), both open classes of words. SKOS-XL is used for modelling lexical units as classes with their own URIs. This allows for the use of an abbreviation relation (abbreviationOf), which holds between the abbreviations and the full forms. For example, the label "s." (URI:xl_pt_dc11) is modelled as an abbreviation of the full form "substantivo" in Portuguese (URI:xl_pt_0acd). In this model, abbreviations are preferred labels, while the full forms are alternative labels for the concepts. The Universal Dependencies Part-of-Speech tags are modelled via the skos:notation property, which allows for the identification and retrieval of each concept regardless of language. For example, the tag for nouns (NOUN) is represented as a notation of the noun concept in our model (URI:c_47d3).

Figure 9.3 below shows the modelling of the Portuguese and French language labels. Here, abbreviations are also declared as preferred labels ("port." and "fr."), while the full forms are alternative labels ("português"

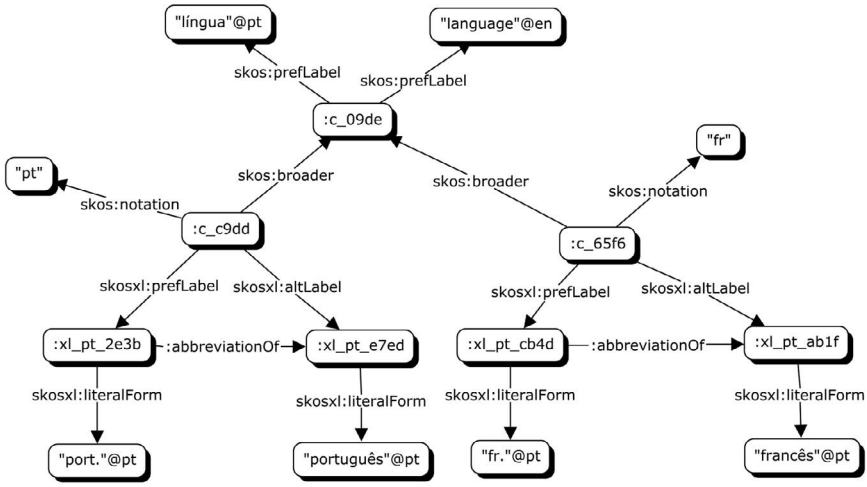

Figure 9.3 Language abbreviations in SKOS (Portuguese and French).

and "francês", respectively). For example, the label "port." (URI:xl_pt_2e3b) is declared as an abbreviation of the full form "português" in Portuguese (URI:xl_pt_e7ed). Language codes are also added through the `skos:notation` property, corresponding to IETF BCP 47 codes. For example, the language code for Portuguese (pt) is represented as a notation of the noun concept in our model (URI:c_c9dd).

These examples show how a model originating in the information community can be applied in the modelling of lexicographical resources More specifically, this approach will be used to annotate the TEI-encoded entries of the VOLP-1940 with URIs corresponding to elements of our SKOS model of the list of abbreviations. For example, the URI of the "s." element can be associated with all noun entries in the TEI encoding of the VOLP-1940. Furthermore, an information system will be able to interpret that all nouns in the VOLP-1940 correspond to an open class of words.

The approach outlined facilitates the retrieval of structured lexicographical information from VOLP-1940 and its interoperability with external systems. This approach also facilitates the use of VOLP-1940 for NLP and information retrieval applications, e.g., for word-sense disambiguation and analysis of semantic change.

Encoding in TEI Lex-0

As already mentioned, a lexicographical article in the VOLP-1940 starts with a base corresponding to the entry, followed by the grammatical information about that unit. This is the basic and regular structure of a VOLP-1940 entry to which the TEI Lex-0 annotation was applied (see Example 9.1):

Example 9.1

Basic and regular structure of a VOLP-1940 entry.

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

While the `entry` element encompasses all the information contained in the lexicographical article, the `form` element is used to note the information relating to the base, detailing its `type` attribute as `"lemma"`, and the orthographic form is provided in the `orth` element. It is important to note that in TEI Lex-0, the `entry` element requires the attributes `@xml:id`, the entry identifier and `@xml:lang`, the appropriate language code according to IETF BCP 47, which, in turn, is based on ISO 639 standards. Since we are dealing with vocabulary entries, we use the `form type=lemma`.

In the particular case of homonymous words, as in Example 9.2, "afecto", the lemma is split. In TEI Lex 0, avoiding possible structural ambiguities,

Example 9.2

Encoding of the entry "afecto[1]" of the VOLP-1940 in TEI Lex-0.

```
<entry xml:lang="pt" xml:id="afecto_1" n="1"
"type="monolexicalUnit"> <form type="lemma">
 <orth>afecto</orth>
 <lbl>1</lbl>
 <pc>(</pc>
 <pron extend="part">ét</pron>
 <lbl>)</lbl>
 </form>
 <pc>,</pc>
 <gramGrp>
 <gram type="pos" norm="NOUN">s.</gram>
 <gram type="gen">m.</gram>
 </gramGrp>
 <pc>:</pc>
 <sense>
 <def>afeição</def>
<pc>.</pc>
 </sense>
</entry>
```

the `superEntry` element (which groups a sequence of entries, such as a set of homographs) is no longer allowed, and we use `entry` systematically. To mark the numeric index, the element `lbl` preserves the digit of the original. The attribute `n` of the entry will, in turn, prove important for the further processing of the entry by computational tools.

We now focus on the TEI Lex-0 encoding word classes, that is, the "designação geral dos conjuntos distintos nos quais se agrupam as palavras do léxico, diferenciados pelas suas propriedades gramaticais e semânticas" (general designation of the different sets in which the words in the lexicon are grouped, differentiated by their grammatical and semantic properties; Raposo 2013, 326–327). We also look at how to present information about the language of origin of a lemma using language codes.

The grammatical properties of a lemma are specified in `entry/gramGrp/gram`. This `gram` element typically specifies the part-of-speech of the entry. In TEI Lex-0, specific elements of the *TEI Guidelines* for grammatical properties are dispensed with. We annotated the word classes using `@type="pos"`, e.g., `<gram type="pos">s.</gram>`, also marking the gender as `@type "gen"`, e.g., `<gram type="gender">f.</gram>`. We also considered using the `@norm` attribute for the Universal Dependencies Part-of-Speech values, as mentioned above. To ensure the accuracy of this correspondence, a complete list of possibilities for the contents of this label was calculated, and the annotation was added manually. In Table 9.1, we present a sample of the survey performed.

Considering the goals of TEI Lex-0 to serve as a common baseline and target format for transforming and comparing different lexical resources, the authors of the new guidelines decided to do away with the specific elements for grammatical properties, recommending the use of typed elements. The attribute values for `gram/@type` are a semi-closed list and the possibility of adding a new value, `"pos-sub"`, to annotate subcategories is currently being discussed. For instance, adverbs are grouped according to their function and value (subclasses), following the traditional Portuguese grammatical classification, which is obsolete. In this case, we decided to encode the part of speech with the `"pos"` value and a subcategory in the new value, `<gram type="pos" norm="ADV">adv.</gram>`, followed by `<gram type="pos-sub" expand="de afirmação">af.</gram>`.

Information about the language of origin of a lemma was encoded through the `etym` element (etymology) as a `"borrowing"`. Language information was provided in two different places. In the `lang` tag, it is presented as shown to the user, while the `@xml:lang` attribute encodes the language information as an IETF BCP 47 value. This is shown in Example 9.3, where the lemma "croché" is the Portuguese form of the French lemma "crochet".

Upon illustrating the encoding of some lexicographical articles in TEI, the examples show that this process is more detailed in TEI Lex-0 and more structured and accurate, allowing systems to better process the annotated data. TEI Lex-0 should be seen primarily as a format in which the existing

Example 9.3

Encoding of the entry "croché" of the VOLP-1940 in TEI Lex-0.

```
<entry xml:lang="pt" xml:id="croché" n="1"
"type="monolexicalUnit">
<form type="lemma">
 <orth>croché</orth>
 <pc>(</pc>
 <pron extend="part">è</pron>
 <lbl>)</lbl>
 </form>
 <pc>,</pc>
 <gramGrp>
 <gram
 type="pos" norm="NOUN">s.</gram>
 <gram
 type="gen">m.</gram>
 </gramGrp>
 <pc>:</pc>
 <etym
type="borrowing">
 <lbl>aportg. do</lbl>
 <lang>fr</lang>
 <mentioned xml:lang="fr">crochet</mentioned>
 </etym>
</entry>
```

TEI dictionaries can be annotated and exploited more uniformly, with features that will include, among others, basic and advanced search capabilities. Alongside this, SKOS will play an important role in the organisation of lexicographical data as well as in ensuring its interoperability.

Conclusion: Breaking the ice – the benefits of an interdisciplinary action

In the course of our work, we invested in an effective trans-disciplinary approach that combines theories and methods of lexicography and information science, placing the TEI and SKOS standards at the very core of our research. We therefore contributed to the creation of the linguistic digital heritage that is at the heart of digital humanities. We implemented two standards with different but complementary goals, given that TEI specifies "encoding methods for machine-readable texts, chiefly in the humanities, social sciences and linguistics" (TEI Consortium) and SKOS, in turn, "is a common data model for knowledge organization systems such as thesauri, classification schemes, subject heading systems and taxonomies" (W3C 2009).

In the context of VOLP-1940, TEI encodes contents acting primarily on the microstructure of the dictionary, whereas SKOS allows the modelling of KOS, acting on macrostructural information and enabling the connection to other existing systems and resources. The modelling of lexicographical categories and their linguistic realisations (i.e., abbreviations and full forms) in SKOS facilitates the future exploration of VOLP-1940 as Linked Data. For example, through the language category, it opens the possibility for a system to extract all entries that are adopted from other languages (e.g., "croché" in Portuguese, borrowed from the French "crochet"), which would be an important application for linguistics scholars interested in borrowing and word-formation processes. For interoperability purposes, the lexicographical categories modelled in SKOS should be aligned to external vocabularies and ontologies, such as the widely used LexInfo ontology of lexical categories (LexInfo n.d.). For example, our class for nouns should be mapped to LexInfo's noun class, which would facilitate the reuse of VOLP-1940's subset of nouns as Linked Data. We aim to foster open access to resources that have a recognised heritage value, conceived from the start as dynamic searchable resources. This is a task of linguistic, heritage and historical relevance that will certainly contribute to the establishment of the Portuguese lexicon at the time – until 1940 – around which the identity of a linguistic and cultural community has been built and preserved.

With the work we have done so far, we believe we have highlighted the need to change traditional lexicographical practices. Many of the principles now defined and adopted will be used as a guide for the annotation of the remaining entries and application to subsequent bodies of work since they share several typographic conventions that have now been identified. With this process of retro-digitisation of lexicographical reference works and the application of this methodology, we intend to represent the ever-increasing synergy between lexicographers, terminologists, computational linguists, information experts and digital humanists that we so keenly advocate.

This methodology has already proved fruitful, as the Portuguese Foundation for Science and Technology (FCT) has financed the project MORDigital – Digitisation of *Diccionario da Lingua Portugueza* by António de Morais Silva. The main goal of MORDigital is to encode the selected editions of *Diccionario de Lingua Portugueza* by António de Morais Silva (MOR), first published in 1789. MORDigital aims to promote accessibility to cultural heritage while fostering reusability and contributing towards a greater presence of digital lexicographical content in Portuguese through open access tools and standards. The methodology applied to MOR will have an enormous impact in Portuguese-speaking countries. MOR represents a great legacy, since it marks the beginning of Portuguese dictionaries, having served as a model for all subsequent lexicographical productions throughout the 19th and 20th centuries.

The strength of the methodology applied to the VOLP-19 lies in the fact that it is reproducible and reusable. In the near future, we will expand our method, link different monolingual legacy dictionaries (Portuguese, French, Spanish) and interconnect them through the "skosification" of the macrostructural elements. TEI Lex-0 will be used to encode the microstructural information of the monolingual dictionaries in the three languages, thus increasing multilingual lexicographical repositories.

One of the main challenges raised by the methodology proposed in this chapter is to combine the skills of the various scientific disciplines that make up the humanities in connection with information science. This is because the standards are cross-disciplinary tools that help build a joint methodology that benefits everyone. At the end of the project, we expect to have codified a vocabulary with a significant heritage value, compatible with the most advanced standards for academic and open-access digital editions.

We believe that this project will contribute significantly to the analysis and annotation of Portuguese lexical resources using computer-assisted processes. It will allow us to rethink how to design new lexicographical products that are truly digital and not merely a simple reproduction of paper editions, which will respond more effectively to the needs of the end-users.

In the next few years, the challenge lies in creating new profiles for the humanities. Universities must create multidimensional profiles that associate the skills of linguistics, computing, and information science. That is what defines digital humanities.

Acknowledgements

This paper is supported by the Portuguese National Funding through the FCT – Fundação para a Ciência e Tecnologia as part of the project Centro de Linguística da Universidade NOVA de Lisboa – UID/LIN/03213/2020 and the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015 (ELEXIS; European Lexicographic Infrastructure).

Bibliography

Academia das Ciências de Lisboa. 1940. *Vocabulário Ortográfico da Língua Portuguesa* (VOLP-1940). Lisboa: Imprensa Nacional. Accessed December 1, 2020. https://volp-acl.pt/index.php/vocabulario-1940/projeto.

Art & Architecture Thesaurus Online. 2017. https://www.getty.edu/research/tools/vocabularies/aat/

Baker, Thomas, Sean Bechhofer, Antoine Isaac, Alistair Miles, Guus Schreiber and Ed Summers. 2013. "Key Choices in the Design of Simple Knowledge Organization System (SKOS)." *Journal of Web Semantics*, 20: 35–49. https://doi.org/10.1016/j.websem.2013.05.001

Bergenholtz, Henning and Rufus H. Gouws 2012. "What is lexicography?" *Lexikos* 22: 31–42. https://doi.org/10.5788/22-1-996

Borko, Harold. 1968. "Information Science: What Is It?" *American Documentation* 19 (1): 3–5. https://doi.org/10.1002/asi.5090190103

Bothma, Theo J. D. 2018. "Lexicography and Information Science." In *The Routledge Handbook of Lexicography*, edited by Pedro A. Fuertes-Olivera, 197–216. London: Routledge.

DARIAH. n.d. "Lexical Resources." Accessed March 12, 2021. https://www.dariah.eu/activities/working-groups/lexical-resources/

EU Vocabularies. n.d. Accessed March 12, 2021. https://op.europa.eu/en/web/eu-vocabularies/

FAIR PRINCIPLES: Findable, Accessible, Interoperable, Reusable. 2016. Accessed March 12, 2021. https://www.go-fair.org/fair-principles/.

Felber, Helmut. 1984. *Terminology Manual*. Paris: UNESCO.

Gold, Matthew K. and Lauren F. Klein, eds. 2019. *Debates in the Digital Humanities*. Mineápolis: University of Minnesota Press.

Gonçalves, Maria Filomena. 2020. "Orthography and Orthoepy." In *Manual of Standardization in the Romance Languages, Manuals of Romance Languages (MRL)*, edited by Franz Lebsanft and Felix Tacke, 24, 649–677. Berlin/Boston: De Gruyter.

Hjørland, Birger. 2008. "What is Knowledge Organization (KO)?" *Knowledge Organization* 35 (2/3): 86–101. https://doi.org/10.5771/0943-7444-2008-2-3-86

Hockey, Susan. 2004. "The History of Humanities Computing." In *A Companion to Digital Humanities*, edited by Susan Schreibman, Ray Siemens, and John Unsworth. Oxford: Blackwell.

Hodge, Gail. 2000. *Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files*. Washington, DC: The Digital Library Federation. Council on Library and Information Resources.

IETF BCP 47. n.d. Tags for Identifying Languages. Accessed March 12, 2021. https://tools.ietf.org/html/bcp47.

ISO 1087. 2019. *Terminology Work and Terminology Science – Vocabulary*. Geneva: ISO.

ISO 1951. 2007. *Presentation/representation of Entries in Dictionaries – Requirements, Recommendations and information*. Geneva: ISO.

ISO 25964-1. 2011. *Information and Documentation – Thesauri and Interoperability with other Vocabularies—Part 1: Thesauri for Information Retrieval*. Geneva: ISO.

ISO 25964-2. 2013. *Information and Documentation – Thesauri and Interoperability with other Vocabularies—Part 2: Interoperability with Other Vocabularies*. Geneva: ISO.

ISO/IEC 2382. 2015. *Information Technology – Vocabulary*. Geneva: ISO.

ISO/TR 20694. 2018. *A Typology of Language Registers*. Geneva: ISO

Landau, Sidney I. 2001. *Dictionaries: The Art and Craft of Lexicography*. 2nd ed. Cambridge: Cambridge University Press.

LexInfo. n.d. Accessed March 12, 2021. https://www.lexinfo.net/.

Linguistic Linked Open Data Cloud (LLOD). n.d. Accessed March 12, 2021. https://linguistic-lod.org/llod-cloud.

Miles, Alistair, and Sean Bechhofer. 2009. *SKOS Simple Knowledge Organization System Reference*. 18 August 2009. http://www.w3.org/TR/skos-reference.

Raposo, Eduardo Buzaglo Paiva. 2013. Estrutura da frase. In *Gramática do Português*, edited by Eduardo Buzaglo Paiva Raposo, Maria Fernanda Bacelar do Nascimento, Maria Antónia Coelho da Mota, Luísa Segura, and Amália Mendes, 303–398. Vol. I. Lisboa: Fundação Calouste Gulbenkian.

Rey-Debove, Josette. 1971. *Étude Linguistique et Sémiotique des Dictionnaires Français Contemporains*. Paris: The Hague.

Robinson, Lyn, Ernesto Priego, and David Bawden. 2015. "Library and Information Science and Digital Humanities: Two Disciplines, Joint Future?" In *Re:Inventing Information Science in the Networked Society*, edited by Franjo Pehar, Christian Schlögl, and Christian Wolff, 44–54. Glückstadt: Verlag Werner Hülsbusch.

Rundell, Michael. 2010. "What Future for the Learner's Dictionary?" In *English Learners' Dictionaries at the DSNA 2009*, edited by Ilan J. Kernerman and Paul Bogaards, 169–175. Jerusalem: Kdictionaries.

Salgado, Ana, Rute Costa, Toma Tasovac, and Alberto Simões, A. 2019. "TEI Lex-0 In Action: Improving the Encoding of the Dictionary of the Academia das Ciências de Lisboa." In *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference*, edited by. Iztok Kosem, Tanara Zingano Kuhn, Margarita Correia, José Pedro Ferreira, Maarten Jansen, Isabel Pereira, Jelena Kallas, Miloš Jakubíček, Simon Krek, and Carole Tiberius, 417–433. Brno: Lexical Computing CZ, s.r.o. https://elex.link/elex2019/wpcontent/uploads/2019/09/eLex_2019_23.pdf

Salgado, Ana, Sina Ahmadi, Alberto Simões, John Philip McCrae, and Rute Costa. 2020. "Challenges of Word Sense Alignment: Portuguese Language Resources." In *Proceedings of 7th Workshop on Linked Data in Linguistics (LDL 2020) Building Tools and Infrastructure*, edited by Maxim Ionov, John P. McCrae, Christian Chiarcos, Thierry Declerck, Julia Bosque-Gil, and Jorge Gracia, 45–51. France: Marseille. https://www.aclweb.org/anthology/2020.ldl-1.0.pdf

Saracevic, Tefko. 1999. "Information Science." *American Documentation* 50(12): 1051–1063. https://doi.org/10.1002/(SICI)1097-4571(1999)50:12%3C1051::AID-ASI2%3E3.0.CO;2-Z.

Stührenberg, Maik. 2012. "The TEI and Current Standards for Structuring Linguistic Data." *Journal of the Text Encoding Initiative* [Online] issue 3, (November). https://doi.org/10.4000/jtei.523

Tarp, Sven. 2018. "Lexicography as an Independent Science." In *The Routledge Handbook of Lexicography*, edited by Pedo A. Fuertes-Olivera, 19–33. London: Routledge.

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

Tasovac, Toma, et al. 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, eds. 2021. "TEI P5: Guidelines for Electronic Text Encoding and Interchange. 4.2.1." Last modified March 1, 2021. TEI Consortium. http://www.tei-c.org/Guidelines/P5/

Terras, Melissa, Julianne Nyhan, Edward Vanhoutte, eds. 2013. *Defining Digital Humanities: A Reader*. Londres: Ashgate.

Trap-Jensen, Lars. 2018. "Lexicography between NLP and Linguistics: Aspects of Theory and Practice." In *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*, edited by Jaka Čibej, Vojko Gorjanc, Iztok Kosem, and Simon Krek, 25–37. Ljubljana: Ljubljana University Press, Faculty of Arts. https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/118/211/3000-1.pdf

Universal Dependencies. n.d. Accessed March 12, 2021. https://universaldependencies.org/

Verdelho, Telmo. 2007. "Dicionários Portugueses, Breve História." In *Dicionarística Portuguesa: Inventariação e Estudo do Património Lexicográfico*, edited by Telmo Verdelho, and João Paulo Silvestre. Aveiro: Universidade de Aveiro.

W3C. 2009. "SKOS Simple Knowledge Organization System Reference." https://www.w3.org/TR/skos-reference/

Wiegand, Herbert Ernst. 2013. "Lexikographie und Angewandte Linguistik." *Zeitschrift Für Angewandte Linguistik* 58 (1): 13–39. https://doi.org/10.1515/zfal-2013-0002

Wiegand, Herbert Ernst, Rufus H. Gows, Matthias Kammerer, Michael Mann and WernerWolski. 2020. *Dictionary of Lexicography and Dictionary Research*. Vol. 3 I-U. Berlin/Boston: Walter de Gruyter.

Williams, Geoffrey. 2019. "The Problem of Interlanguage Diachronic and Synchronic Markup." In *The Landscape of Lexicography*, edited by Alina Villalva and Goeffrey Williams. Lisboa–Aveiro: Centro de Linguística da Universidade de Lisboa–Universidade de Aveiro.

Zeng, Marcia. 2008. "Knowledge Organization Systems (KOS)." *Knowledge Organization* 35 (2/3): 160–182. https://doi.org/10.5771/0943-7444-2008-2-3-160

Appendix

VOLP-1940 list of abbreviations (typological organization)

Part of speech

OPEN CLASS WORDS

adj. (adjectivo) [adjective]
adv. (advérbio) [adverb]adv. af. (advérbio de afirmação) [affirmation adverb]
adv. conf. (advérbio de confirmação) [confirmation adverb]
adv. design. (advérbio de designação) [designation adverb]
adv. dúv. (advérbio de dúvida) [adverb of doubt]
adv. excl. (advérbio de exclusão) [exclusion adverb]
adv. interr. (advérbio interrogativo) [interrogative adverb]
adv. lug. (advérbio de lugar) [adverb of place]
adv. mod. (advérbio de modo) [mode adverb]
adv. neg. (advérbio de negação) [negation adverb]
adv. num. (advérbio numeral) [numeral adverb]
adv. rel. (advérbio relativo) [relative adverb]

adv. temp. (advérbio de tempo) [adverb of time]
interj. (interjeição) [interjection]interj. excl. (interjeição exclamativa) [exclamatory interjection]
interj. voc. (interjeição vocativa) [vocative interjection]
s. (substantivo) [noun]
v. (verbo) [verb]

CLOSED CLASS WORDS

art. (artigo) [determiner]
conj. (conjunção) [conjunction]conj. adv. (conjunção adversativa) [adversative conjunction]
conj. caus. (conjunção causal) [causal conjunction]
conj. comp. (conjunção comparativa) [comparative conjunction]
conj. conc. (conjunção concessiva) [concessive conjunction]
conj. concl. (conjunção conclusiva) [conclusive conjunction]
conj. cond. (conjunção condicional) [conditional conjunction]
conj. cons. (conjunção consecutiva) [consecutive conjunction]
conj. cop. (conjunção copulativa) [copulative conjunction]
conj. disj. (conjunção disjuntiva) [disjunctive conjunction]
conj. fin. (conjunção final) [final conjunction]
conj. int. (conjunção integrante) [integral conjunction]
conj. temp. (conjunção temporal) [temporal conjunction]
num. (numeral) [numeral]num. card. (numeral cardinal) [cardinal numeral]
num. distr. (numeral distributivo) [distributive numeral]
num. fracc. (numeral fraccionário) [fractional numeral]
num. mult. (numeral multiplicativo) [multiplicative numeral]
num. ord. (numeral ordinal) [ordinal numeral]
pron. (pronome) [pronoun]pron. dem. (pronome demonstrativo) [demonstrative pronoun]
pron. ind. (pronome indefinido) [indefinite pronoun]
pron. interr. (pronome interrogativo) [interrogative pronoun]
pron. pess. (pronome pessoal) [personal pronoun]
pron. pess. compl. (pronome pessoal complemento) [personal pronoun complement]
pron. pess. suj. (pronome pessoal sujeito) [subject personal pronoun]
pron. poss. (pronome possessivo) [possessive pronoun]
pron. refl. (pronome reflexo) [reflex pronoun]
pron. rel. (pronome relativo) [relative pronoun]
prep. (preposição) [adposition]

Grammatical gender

f. (feminino) [feminine]
m. (masculino) [masculine]
2 gén. (2 géneros) [dual gender]

Grammatical number

sing. (singular) [singular]
2 núm. (2 números) [dual number]
pl. (plural) [plural]

Language

al. (alemão) [Deutsch]
ár. (árabe; arábico) [Arabic]
din. (dinamarquês) [Danish]
esp. (espanhol) [Spanish]
finl. (finlandês) [Finnish]
fr. (francês) [French]
gr. (grego) [Greek]
hebr. (hebraico) [Hebrew]
hol. (holandês) [Dutch]
ingl. (inglês) [English]
it. (italiano) [Italian]
jap. (japonês) [Japanese]
lat. (latim) [Latin]
lat. vulg. (latim vulgar) [Vulgar Latin]
lit. (lituano) [Lithuanian]
nor. (norueguês) [Norwegian]
pol. (polaco) [Polish]
port. (português) [Portuguese]
rom. (romano) [Roman]
scr. (sânscrito) [Sanskrit]

Register

ant. (antigo) [old]
arc. (arcaico) [archaic]
pop. (popular) [popular]

Onomastic classification

antr. (antropónimo; antroponímico) [anthroponym; person name]
astr. (astrónimo) [astronomical name]
bibl. (bibliónimo) [renowned book name]
cogn. (cognome) [cognomen]
cron. (cronónimo) [chrononym; calendar name]
etn. (etnónimo) [ethnonym]
heort. (heortónimo) [holiday name]
hier. (hierónimo) [sacred name]
mit. (mitónimo) [mythonym; mythological name]
patr. (patronímico) [patronymic]
pros. (prosónimo) [nickname]
top. (topónimo) [toponym; place name]

Tense

fut. conj. (futuro do conjuntivo) [future subjunctive]
fut. ind. (futuro do indicativo) [future indicative]
ger. (gerúndio) [gerund]
imper. (imperativo) [imperative]
imperf. conj. (imperfeito do conjuntivo) [imperfect subjunctive]
imperf. ind. (imperfeito do indicativo) [imperfect indicative]
inf. (infinitivo) [infinitive]
inf. pess. (infinitivo pessoal) [personal infinitive]
m. q. perf. ind. (mais-que-perfeito do indicativo) [pluperfect indicative]
part. pass. (particípio passado) [past participle]
part. pres. (particípio presente) [present participle]
perf. ind. (perfeito do indicativo) [perfect indicative]
pres. cond. (presente do condicional) [conditional present]
pres. conj. (presente do conjuntivo) [present subjunctive]
pres. ind. (presente do indicativo) [present indicative]

Etimology

lat. (latino) [latin]
or. gr. (origem grega) [greek origin]
or. lat. (origem latina) [latin origin]

Word formation

adapt. (adaptação) [adaptation]
agl. (aglutinação) [agglutination]
aportg. (aportuguesamento) [adapted Portuguese form]
contr. (contracção) [contraction]
el. comp. (elemento de composição) [composition elemento]
inf. (infixo) [infix]
pref. (prefixo) [prefix]
red. (redução) [reduction]
red. pop. (redução popular) [popular reduction]
rad. (radical) [radical]
suf. (sufixo) [suffix]

Others

alf. (alfabeto) [alphabet]
át. (átono) [unstressed]
ax. (axiónimo) [honorific]
cat. morf. (categoria morfológica) [morphological category]
cf. (confira) [compare; consult]
cons. (consoante) [consonant]
constr. (construção) [construction]
dif. (diferente) [different]
diss. (dissilábico) [disyllabic]

dit. (ditongo) [diphthong]
el. (elemento) [element]
el. art. (elemento articular) [joint element]
el. nom. (elemento nominal) [nominal element]
el. part. (elemento participial) [participial element]
el. prot. (elemento protético)
[prothetic element] el. top. (elemento toponímico) [toponymic element]
equiv. (equivalente) [equivalent]
f. (forma) [form]
flex. (flexão) [inflection]
form. port. (formação portuguesa) [portuguese formation]
f. paral. (forma paralela) [parallel form]
f. verb. (forma verbal) [verbal form]
hipoc. (hipocorístico) [hypocoristic]
lig. (ligação) [connection]loc. (locução) [phrase]
loc. adj. (locução adjectiva) [adjetive phrase]
loc. adv. mod. (locução adverbial de modo) [adverbial phrase]
loc. adv. temp. (locução adverbial de tempo) [temporal phrase]
loc. prep. (locução prepositiva) [adposition phrase]
loc. pron. pess. (locução pronominal pessoal) [personal pronominal phrase]
loc. s. (locução substantiva) [noun phrase]
loc. s. f. (locução substantiva feminina) [feminine noun phrase]
loc. s. m. (locução substantiva masculina) [masculine noun phrase]
n. (nome) [name]
pal. (palavra) [word]
part. apass. (partícula apassivante) [passive particle]
part. aux. (partícula auxiliar) [auxiliary particle]
part. expl. (partícula expletiva) [expletive particle]
pess. (pessoa) [person]
p. ex. (por exemplo) [for example]
p. ext. ou abrev. (por extenso ou abreviadamente) [in full or abbreviated]
sent. (sentido) [sense]
sup. (superlativo) [superlative]
term. (terminação) [ending]
tón. (tónico) [stressed]
v. (veja) [see]
var. (variação) [variation]
vog. (vogal) [vowel]