ENCODING POLYLEXICAL UNITS WITH TEI LEX-0: A CASE STUDY

Toma TASOVAC
Belgrade Center for Digital Humanities, Belgrade, Serbia

Ana SALGADO
NOVA CLUNL Universidade NOVA de Lisboa, Lisbon, Portugal,
Academia das Ciências de Lisboa, Lisbon, Portugal

Rute COSTA
NOVA CLUNL Universidade NOVA de Lisboa, Lisbon, Portugal

Tasovac, T., Salgado, A., Costa, R. (2020): Encoding polylexical units with TEI Lex-0: A case study. Slovenščina 2.0, 8(2): 28–57.

DOI: https://doi.org/10.4312/slo2.0.2020.2.28-57

The modelling and encoding of polylexical units, i.e. recurrent sequences of lexemes that are perceived as independent lexical units, is a topic that has not been covered adequately and in sufficient depth by the Guidelines of the Text Encoding Initiative (TEI), a de facto standard for the digital representation of textual resources in the scholarly research community. In this paper, we use the Dictionary of the Portuguese Academy of Sciences as a case study for presenting our ongoing work on encoding polylexical units using TEI Lex-0, an initiative aimed at simplifying and streamlining the encoding of lexical data with TEI in order to improve interoperability. We introduce the notion of *macro- and microstructural relevance* to differentiate between polylexicals that serve as headwords for their own independent dictionary entries and those which appear inside entries for different headwords. We develop the notion of *lexicographic transparency* to distinguish between those units which are not accompanied by an explicit definition and those that are: the former are encoded as `<form>`–like constructs, whereas the latter becomes `<entry>`–like constructs, which can have further constraints imposed on them (sense numbers, domain labels, grammatical labels etc.). We codify the use of attributes on `<gram>` to encode different kinds of labels for polylexicals (implicit, explicit and normalised),

concluding that the interoperability of lexical resources would be significantly improved if dictionary encoders would have access to an expressive but relatively simple typology of polylexical units.

Keywords: TEI, Lexicography, Language Resources, Polylexical Units, Interoperability

1 INTRODUCTION

A polylexical unit can be defined as a stable and recurrent sequence of lexemes that are perceived as an independent lexical unit by the speakers of a language. In the specialized literature, different authors with different theoretical backgrounds (Gantar et al., 2018; Fellbaum, 2016; Baldwin and Kim, 2010; Calzolari et al., 2002; Sag et al., 2001; Moon, 1998; Cowie, 1994, 1998; Mel'čuk, 1984–1999, 1998; among others) have referred to these morphosyntactic sequences as multiword expressions, collocations, phrasemes, phraseologies, idiomatic expressions, lexical combinations, and so forth. Each of these designations is often defined inside a particular theoretical linguistic framework.

At the same time, scholars have long recognised that polylexical units are essential components of lexical resources (Svensén, 2009; Atkins and Rundell, 2008; Fontenelle, 1997; Hausmann, 1979; Mel'čuk et al., 1984–1999; Zgusta, 1971). When including a polylexical item in a dictionary, lexicographers decide on the degree of its lexical independence based on several criteria from different fields of knowledge, including statistics, semantics, morphosyntax, pragmatics and/or, broadly speaking, culture. This kind of lexicographic judgement, enacted through a particular editorial policy *and* influenced by the conventions of a given lexicographic tradition, necessarily leads to multiple ways of capturing, classifying and presenting lexicographic knowledge about polylexical units. The lack of a more general agreement within the lexicographic community makes the process of encoding dictionaries particularly challenging: how can we identify, describe and consistently represent this type of linguistic phenomena in lexical resources if we do not agree on what they are and/or what to call them?

Unlike corpus linguists who try to describe linguistic evidence as it appears in recorded instances of genuine language use, or practising lexicographers who

try to systematise their knowledge about words and their meaning by laying it out in dictionary articles, dictionary encoders work on formally representing the concrete lexicographic content of existing dictionaries. This is an important distinction to be kept in mind in the context of what we are trying to achieve in this paper. When, in the rest of this paper, we discuss polylexical units, we will do so from the point of view of lexicographic data modelling, i.e. the process of explicitly marking up the structural hierarchies and the scope of particular textual components appearing in existing dictionary entries in order to convert them to electronic format as part of lexicographic digitisation workflow (Tasovac and Petrović, 2015). In other words, our starting point will be polylexical units as a stable and recurrent sequence of lexemes that are *perceived as independent lexical units by the lexicographers of a given dictionary*. Our focus will be on how these linguistic phenomena appear on a printed dictionary page and at which level of the dictionary microstructure. Our main goal will be to explore how these phenomena can be formally described using the recommendations of the Text Encoding Initiative (TEI),[1] in general, and TEI Lex-0,[2] in particular.

The encoding of polylexical units in dictionaries is a topic that has not been covered adequately and in sufficient depth by the TEI, a *de facto* standard for the digital representation of textual resources in the scholarly research community. We will discuss the challenges and propose some solutions to this problem. We will also argue that a typology of polylexical units for dictionaries encoding – especially given both the limited resources which are usually available for this kind of work *and* data interoperability as a worthy goal to pursue – need to be relatively general so that it can be used and applied by dictionary encoders in a straight-forward fashion.

The terminology we use in this paper aims to be supra-theoretical, and consequently, as neutral as possible, hence our preference for "polylexical units". We recognize, nonetheless, that the term "multiword expression" (MWE) is already widely used, including in the LMF standard, ISO 24613-1:2019. In this paper, we will, therefore, proceed as follows: when we refer to the linguistic structure of a lexical unit composed of two or more lexemes, we will use the term polylexical unit. In our discussion of TEI Lex-0, we will allow "MWE" as

1 https://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html

2 https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html

an attribute value in order to provide better alignment with LMF and because the TEI Lex-0 community has already used this term.

This article is organised as follows: in Section 2, the lexicographic treatment of polylexical units is explored based on the Dictionary of the Portuguese Academy of Sciences (DLPC) as a case study. A TEI Lex-0 representation of polylexical units in DLPC is discussed in Section 3; and, finally, in Section 4, we offer some concluding remarks and some recommendations about the future work needed in this area.

2 LEXICOGRAPHIC TREATMENT OF POLYLEXICAL UNITS

Dictionaries by design describe systematised knowledge about words and their meanings through typographic conventions that are imbued with meaning and affected by a long tradition: the use of bold typefaces to signal the lemma or headword in a dictionary article; the use of abbreviations (especially in print dictionaries) for grammatical features or usage labels (Salgado et al., 2019a); the numbering of senses and the use of different typefaces for different elements in the hierarchy (definitions, examples, etc.). Experienced dictionary users can become quite proficient at understanding and navigating the structure of the dictionary by interpreting the dictionary's typographic features and the way these features may differ from one dictionary to another. Still, that kind of understanding, based on both knowledge and experience, is not something which can always be easily formalised.

Two main challenges are affecting the modelling of polylexical units in dictionaries, both of them related to the typographical constraints of the print-based, general-language dictionaries:

1. In most general-language dictionaries, polylexical units do *not* appear as headwords, i.e., independent lexical units in the dictionary macrostructure, but rather as sub-units within entries that have a monolexical headword; and

2. Polylexical units in dictionaries are not always *explicitly* labelled as such: they may be typographically singled out, using a particular typeface, but they are not always accompanied by the label which identifies the given unit as a "collocation", "idiom" or a "proverb".

The position of polylexical units in the dictionary and the benefits of lemmatisation have been discussed before (see Jónsson (2009) and Lorentzen (1996), for instance) but for our purposes, it is essential to note that when we suggest particular encodings of the Dictionary of the Portuguese Academy of Sciences, we will be following the structure and the conventions of that very dictionary. That means that we will not be trying to flatten the hierarchy or to encode all polylexical units using the same set of tags. We will be encoding them as they appear within the structure imposed by the dictionary itself.

As for the lack of explicit labels for particular types of polylexical units, we will, in the subsequent sections, explain the extent to which the types can be deduced from the entry structure. We will, in the process, also consult the Introduction to the Dictionary, which to some degree explains the structure from the point of view of the dictionary editors.

2.1 DLPC as a case study

The *Dicionário da Língua Portuguesa Contemporânea* (DLPC) is a monolingual Portuguese dictionary published by Academia das Ciências de Lisboa (2001). As such it is representative of the Academy tradition in European lexicography: large-scale and long-term dictionary projects, initiated and compiled by official national bodies established to record, maintain and promote authoritative accounts of language use (see Considine, 2014). It contains around 70,000 entries and was published in 2001 in two volumes, totalling 3880 pages. The PDF version of the printed dictionary was later converted into XML using a customised version of the P5 schema of the Text Encoding Initiative (TEI), while a custom-built dictionary writing system using TEI as a data model in the backend, was developed to serve as an editing environment for the new and improved online edition of the dictionary (Simões et al., 2016). Besides, the DLPC is currently being converted to the TEI Lex-0 format for data interoperability purposes (Salgado et al., 2019b).

We selected DLPC as a case study in our ongoing work on developing guidelines for encoding polylexicals in TEI Lex-0 for two reasons: (1) as a monolingual scholarly dictionary of the Portuguese language, DLPC covers a wide range of polylexical units from collocations to strongly lexicalised

expressions; and (2) because scholarly dictionaries, with their "pursuit of completeness concerning the entries relevant to subject matters" (see Kinable, 2015) typify detailed lexicographic information and elaborate microstructure, which can more often than not pose challenges in terms of consistent data modelling.

Given the lack of detail given to the encoding of polylexical units in the TEI Guidelines, the authors thought it was essential to take a single but complex dictionary as a starting point for our exploration of the topic in this paper. It goes without saying that further comparative work will be needed to validate and improve our recommendations. But it also goes without saying that the proposed mechanisms for marking up polylexical units in DLPC at different levels of the dictionary microstructure will generally be applicable to other dictionaries as well. While dictionaries may differ in terms of their "typographic view", i.e. page layout, column and line breaks, and their "editorial view", i.e. the sequential arrangement of individual tokens along with the use of specific font styles, punctuation and special symbols (the so-called "editorial" view), they are more easily comparable in terms of their "lexical view", i.e. the underlying structure and the types of information units contained in them.[3] While our focus on DLPC here is, above all, a matter of practicality, we will be using it as a springboard for illustrating broader encoding challenges.

Structurally speaking, we should distinguish two main types of polylexical items:

1. polylexical units which serve as headwords for their own independent dictionary entries;

2. polylexical units which appear inside entries for different headwords.

We will refer to the first category as the macrostructurally relevant polylexical units and the second as the microstructurally relevant polylexical units. The notion of relevance here is local – it refers only to the structure of the given dictionary.

3 On the difference between different "views"of the dictionary, see Section 9.5 "Typographic and Lexical Information in Dictionary Data" in the TEI Guidelines, https://tei-c.org/release/doc/tei-p5-doc/en/html/DI.html#DIMV.

2.1.1 Macrostructurally relevant polylexical units

In Salgado et al. (2019b), we identified four different types of headwords in DLPC: monolexical units, polylexical units, affixes and abbreviations. Polylexical headwords can be of two different types:

i) compounds ("palavras compostas" which are graphically realized as "palavras hifenizadas" ["hyphenated words"] (DLPC, 2001, p. XIV) (e.g. **decreto-lei** [decree-law], **franco-canadiano** [French Canadian], **pré-cristão** [pre-Christian); and

ii) Latin phrases ("locuções latinas") (e.g. **fiat lux** [let there be light]).

In the context of this particular dictionary and, more generally speaking in the Portuguese orthographic tradition, hyphenation is treated as a mark of lexicalisation and non-compositional meaning, which leads to lexicographic treatment at an entry-level. For instance, **lugar-comum** [commonplace] does not merely connote a common type of place [**lugar comum**]: the meaning of the hyphenated unit – an ordinary thing, a platitude or a cliché – cannot be obtained from its constituent parts. As such, it is considered, from the point of view of the lexicographer, headword material.[4] Latin phrases, which are used in the Portuguese language, are included in the DLPC macrostructure as entries of their own because they cannot be easily ascribed to particular Portuguese headwords.

2.1.2 Microstructurally relevant polylexical units

Microstructurally relevant polylexical units in DLPC fall into two distinct categories:

i) *lexicographically transparent* polylexical units, i.e., units which are *not* accompanied by an explicit definition; and

ii) *lexicographically non-transparent* polylexical units, i.e., units which *are* accompanied by an explicit definition.

4 The hyphen as a marker of semantic opaqueness, however, is, to a certain extent, a projection of lexicographic idealism. Many polylexicals which are are traditionally hyphenated in Portuguese dictionaries are written without the hyphen in common usage.

2.1.2.1 Lexicographically transparent polylexical units

Lexicographically transparent polylexical combinations in DLPC do not come with an explicit definition in addition to the general one already given for the sense of the headword under which they appear. The lexicographic assumption is that the user will be able to deduce their meaning from their individual components and their syntactic structure. These kinds of polylexical units serve as additional illustrations for the given sense. Still, they differ from typical full-sentence examples in that they stress the collocational aspects of the given headword: they function as lexicographical pointers to the user for how the given word is meaningful — and typically — used in combination with other words. The closeness of these polylexical combinations to actual examples in DLPC is signalled by their proximity next to each other in the dictionary entry, and by their common typographic features: both are set in italic typeface and grouped together inside a particular sense.

descalçar [dɨʃkałsár]. *v.* (Do lat. * *discalceāre*). **1.** Tirar aquilo que se tem calçado; despir os pés ou as mãos; tirar o calçado. ≠ CALÇAR, ENFIAR, PÔR. *Descalçou-se mal chegou a casa. Descalçou um chinelo e atirou-mo. Descalçou a criança e deitou-a.* *+ as botas, as luvas, as meias; + os sapatos.* **2.** Tirar aquilo que serve de apoio para que fique bem assente no chão; tirar o calço. *Descalçou a mesa e esta ficou a balançar.* **3.** Tirar as pedras que cobrem o pavimento. ≃ DESCALCETAR, DESEMPEDRAR. ≠ CALCETAR, EMPEDRAR. **4.** *Fam.* Tirar os recursos; deixar sem soluções, sem possibilidade de resolver alguma coisa; deixar descalço. ≃ DESAMPARAR, DESARMAR, DESPREVENIR. **descalçar a bota,** resolver uma dificuldade.

Figure 1: Descalçar [to remove] – DLCP (2001).

The monolexical lemma **descalçar** [to remove], as shown in Figure 1, has four numbered senses. The first sense consists of a definition “tirar aquilo que se tem calçado; despir os pés ou as mãos; tirar o calçado” [take off one’s shoes; undress one’s feet or hands], followed by three antonyms “calçar, enfiar, pôr” [to put on; to slip on] and three full-sentence examples. In addition, DLPC

lists two sets of typical collocates of the headword separated by a semicolon: + *as botas, as luvas, as meias* and + *os sapatos*. The plus sign is used as a label representing the headword, but the headword is stated only once in a given set: in other words, + *as botas, as luvas, as meias* is directly equivalent to *descalçar as botas, as luvas, as meias* [to remove one's shoes, one's boots, one's gloves], but indirectly equivalent to: *descalçar as botas, descalçar as luvas* and *descalçar as meias*. This is an example of lexicographic shorthand, typical of print dictionaries. In the given case, the user is expected to be able to decipher that the verb *descalçar*, in the given sense (removing something one is wearing), is typically used with objects such as shoes, boots or gloves.

This type of polylexical unit is classified as "co-ocorrente privilegiado" [privileged co-occurrent] in the Introduction to DLPC.[5] The sets separated by the semi-colon are described as "semantically and syntactically related blocks".[6] It appears, however, that this rule is not always followed consistently because the two sets we described above are semantically and syntactically indistinguishable: the difference in the gender of the collocate (*as botas* vs. *os sapatos*) is of no relevance to the construction of this particular type of polylexical unit.

2.1.2.2 Lexicographically non-transparent polylexical units

In DLPC, the treatment of lexicographically non-transparent polylexical units follows a minimal entry-like structure in which the polylexical unit itself is set in boldface (similar to a lemma) and accompanied by a definition (or a pointer to a definition under a different entry). These units can themselves be divided into two further categories, based on the position they take up in the entry microstructure:

1. those that are attached to particular senses; and
2. those that appear at the end of the entry, following the description of individual senses.

5 Privileged co-occurrent is a dependency relationship ("uma relação de dependência") which occurs between full words ("palavras plenas") such as nouns, adjectives, verbs and adverbs and other words in the construction of sentences ("na construção das frases") (DLPC, 2001, p. XXI).

6 "os co-ocorrentes são apresentados em blocos semântica e sintaticamente afins, separados por ponto e vírgula; dentro de cada bloco aparecem separados por vírgula." (DLPC, p. 2001, XXI).

Take, for instance, the following example (Figure 2):

bombeiro[1]**, a** [bõbɐ́jru, -ɐ]. *s.* (De *bomba*[2] + suf. *-eiro*). **1.** Pessoa que faz parte de um corpo organizado de combate a incêndios; o que trabalha com bombas de incêndio. *Sapadores bombeiros.* «*Bombeiro, o pai, apagava incêndios nas matas, trazia as chamas na vista, televisão nem pensar, entretinha-se para adormecer com o* Almanaque Cearense.» (M. O. BRAGA, *Lua*, p. 95). **bombeiro voluntário,** o que pertence a uma corporação com a obrigatoriedade de acudir a incêndios, acidentes, unicamente por filantropia. **corpo**[+] **de bombeiros. 2.** *Bras.* Pessoa que faz ou conserta bombas, canos. **bombeiro hidráulico,** *Bras.,* canalizador. **3.** *Bras. Fam.* Criança que, durante a noite, tem incontinência urinária.

Figure 2: Bombeiro [firefighter] – DLCP (2001).

The monolexical item **bombeiro** [firefighter], as shown in Figure 2, is a headword for an entry which has three distinct, numbered senses. The first sense has a definition written in regular typeface. Two unnumbered examples follow the definition in italic typeface; and of the two examples, the latter is a citation: it is surrounded by quotation marks and followed by a bibliographic reference inside brackets. Following the definition and the examples, the first sense of **bombeiro** has two polylexical items attached to it: **bombeiro voluntário** [volunteer firefighter] and **corpo de bombeiros** [fire brigade]. Both of these polylexical items appear in boldface, just like the lemma, but only the first of the two has a definition in regular typeface ("o que pertence a uma corporação com a obrigatoriedade de acudir a incêndios, acidentes, unicamente por filantropia") appearing after a comma, which is used as a field separator. The second polylexical item has no definition, but its other distinguishing feature is the superscript plus sign which appears after the word "corpo". In DLPC, this superscript label is used by convention to indicate that the given polylexical unit is defined under a different headword: **corpo**[+], in this case, can be thought of as a cross-reference: it tells the reader to look up the entry **corpo** in order to find the definition for **corpo de bombeiros.**

The Introduction to DLPC calls this type of polylexical units "combinatórias fixas" [fixed combinations].[7] They are attached to particular senses of the headword, and defined only once, the first time they appear in the dictionary. That is why **bombeiro voluntário** is defined under **bombeiro** and cross-referenced from **voluntário**, whereas **corpo de bombeiros** is defined under **corpo**, but cross-referenced from **bombeiro**.

Polylexical units that appear *outside* the sense structure are organised the same way as the "fixed combinations" described above: they have lemma-like headwords and can contain definitions, domain labels, etc. The difficulty, from the modelling point of view, is that DLPC does not use a delimiter or a label to separate the last sense in a given entry from the polylexical units that are not attached to a particular sense. That means that for all intents and purposes, a polylexical unit appearing at the end of an entry in DLPC is typographically indistinguishable from a polylexical entry appearing in the last sense of the given entry.

The Introduction to DLPC describes two types of polylexical units which appear outside the sense structure:

1. "locuções" [phrases]; and
2. "expressões idiomáticas ou fraseológicas" [idiomatic or phraseological expressions].

The two types of polylexical units appear in bold on the dictionary page, the only difference being in their labelling: "phrases" are labelled as such, whereas "idiomatic expressions" are not. Neither of the two terms is explicitly defined in the Introduction to the dictionary.

The entry for **dali**, a contraction of "de" (of, from) and "ali" (there), as shown in Figure 3, has two numbered senses. The definitions of the two senses, each of each describes one possible function of the compound preposition (indicating a point of origin of a movement; or indicating the origin of a person,

7 "Fixed combinations" are defined as "combinações de palavras cristalizadas ou em vias de cristalização, que funcionam frequentemente como verdadeiros compostos não hifenizados" [combinations of words crystallised or in the process of crystallisation, which often function as authentic non-hyphenated compounds] (DLPC, p. XXI) e.g. "pedra preciosa" [gemstone] or "sala de jantar" [dining room].

dali [dɐlí]. *contr.* Contr. da prep. *de* com o adv. *ali*. **1.** Indica ponto de partida de um movimento: daquele lugar, daquele ponto. *Fui dali para o médico, sem passar por casa.* **2.** Indica pessoa, entidade ou situação acabadas de referir e que constituem a fonte ou origem de alguma coisa. *Eu bem insisti junto dos serviços, mas parece que dali não virá a solução do problema. O meu avô está tão mal que os médicos já não esperavam nada dali.* **dali a nada,** *loc. adv.,* muito pouco tempo depois. *Dali a nada estava ele a chamar-me.* **dali a pouco (tempo),** *loc. adv.,* pouco tempo depois. *Dali a pouco chegou o meu irmão.* **dali em diante,** *loc. adv.,* a partir de então, desde esse momento. *Devido a um esforço maior tirou boa nota: dali em diante foi sempre aluno aplicado.* **dali para a frente,** *loc. adv.,* o m. que *dali em diante. Dali para a frente nunca mais senti dificuldades.* **dali por diante,** *loc. adv.,* o m. que *dali em diante. Dali por diante resolveu-me sempre qualquer problema.*

Figure 3: Dali [from there] – DLCP (2001).

thing or situation). From the typographic layout of the entry alone, it would be impossible to judge whether the five polylexical units **dali a nada**, **dali a pouco (tempo)**, **dali em diante**, **dali para a frente** and **dali por diante** are meant to be attached to the second sense or whether they appear outside the sense structure. Each of the polylexical units is explicitly labelled as *loc. adv.* [adverbial phrase].

The dictionary itself defines **locução** in its grammatical sense as a group of words that work, semantically and syntactically as a whole, equivalent to a single word.[8] The same sense also includes several different types of expressions: adjectival, adverbial, conjunctive, prepositional and verbal.

8 "Grupo de palavras que funcionam, semântica e sintacticamente como um todo, que equivalem a um só vocábulo. Rey and Chantreau (1993) underline the difference between lexical and grammatical phrases: "Locution [...] est exactement 'manière de dire', manière de former le discours, d'organiser les éléments disponibles de la langue pour produire une forme fonctionnelle. C'est pourquoi on peut parler de 'locutions adverbiales' ou 'prépositives', alors que ces mots grammaticaux complexes ne seraient jamais appelés des 'expression' (p. VI).

dura [dúɾɐ]. *s. f.* (Deriv. regres. de *durar*). **1.** Qualidade do que resiste ao tempo e ao uso. ≃ DURABILIDADE. **2.** Espaço de tempo entre o princípio e o fim de uma coisa. ≃ DURAÇÃO. **ser de pouca dura,** durar pouco tempo; passar depressa. *Foi amor de pouca dura.* **ser sol de pouca dura,** ser algo que, por ser bom ou agradável, dura pouco tempo. *Aquele bom-humor do chefe foi sol de pouca dura, pois começou logo a resmungar.*

Figure 4: Dura [durability; duration] – DLCP (2001).

The entry for **dura** [duration], on the other hand, as shown in Figure 4, has two numbered senses followed by two polylexical units: **ser de pouca dura** [to be short-lived] and **ser sol de pouca dura** [lit. to be a sun that does not last, i.e., to be a nine days' wonder]) without explicit labelling of the type of units that they are.

In DLPC proper, **expressão idiomática** has the domain label Linguistics and is defined as an expression that is peculiar to the language, usually because its meaning is not literal.[9] The **expressão fraseológica** [phraseological expression] is not defined in the dictionary.

3 REPRESENTING POLYLEXICAL UNITS IN TEI LEX-0

TEI is a *de facto* standard for the digital encoding of all types of written texts, ranging from standard books to poems, visiting other less straightforward documents, e.g., tables, mathematical formulae, cookery recipes or even music notation. It also defines how specific humanities resources, including morphologically annotated monolingual and parallel corpora, should be encoded. Chapter 9 of the TEI Guidelines[10] focuses specifically on the encoding of dictionaries and other types of lexical resources.

TEI Lex-0[11] (Romary and Tasovac, 2018) is a newer, stricter subset of TEI, which was launched in 2016 by the DARIAH Working Group on Lexical

9 "*Ling*. a que é peculiar a uma língua, geralmente devido ao facto de o seu significado não ser literal."

10 https://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html

11 https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html

Resources.[12] The goal of TEI Lex-0 is to establish a baseline encoding and a target format to facilitate the interoperability of heterogeneously encoded lexical resources. TEI Lex-0 should not be thought of as a replacement of the Dictionary Chapter in the TEI Guidelines but rather as a "format that existing TEI dictionaries can be unequivocally transformed to in order to be queried, visualised, or mined uniformly".[13] In the context of the ELEXIS project,[14] TEI Lex-0 has been adopted, together with OntoLex, as one of the baseline formats for the ingestion of existing dictionaries into the ELEXIS infrastructure (McCrae et al., 2019). While TEI Lex-0 is being developed, some of its best-practice recommendations are also changing the recommendations of TEI Guidelines themselves.

3.1 Polylexical units in TEI Guidelines

The Dictionary Chapter of the TEI Guidelines is very sparse when it comes to recommendations for encoding polylexical units. The only mention of the adjective "multi-word" appears in the definition of the element `<term>`: "contains a single-word, multi-word, or symbolic designation which is regarded as a technical term" but this is not relevant for the encoding of polylexical units in general-purpose dictionaries.

TEI includes an element `<colloc>` (collocate), which is defined as containing "any sequence of words that co-occur with the headword with significant frequency" but, in a different example, "colloc" is used as an attribute value for the element `<usg>` (usage). It is precisely this type of ambiguity that TEI Lex-0 is trying to resolve.

The TEI Guidelines recommend the use of `<re>` (related entry) to encode "related entries for direct derivatives or inflected forms of the entry word, or for compound words, phrases, collocations, and idioms containing the entry word" with barely any useful examples, or discussion of how to encode different types of polylexical units. TEI Lex-0, on the other hand, does not include `<re>`. In TEI Lex-0, `<entry>` was made recursive in order to account

12 https://www.dariah.eu/activities/working-groups/lexical-resources/

13 https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html#index.xml-body.1_div.1

14 https://elex.is/

for nestable entry-like structures without the need to resort to <re>, a differently named element whose content model would be indistinguishable from <entry> itself. Eventually, the new content model of <entry>, which allows nesting, was adopted by TEI itself.

3.2 Encoding macrostructurally relevant polylexical units

In terms of modelling, polylexical units as headwords do not present any particular challenges for TEI Lex-0. Because they function as lemmas in dictionary entries, they need to be encoded with the required @type attribute on <form>. DLPC does not label them explicitly as polylexical, which is why previously in Salgado et al. (2019b), the authors recommended that this information be encoded as a @type attribute on <entry>. At the time, the goal was to differentiate entries based on their headwords as monolexical, polylexical, affixes and abbreviations. Nevertheless, for lexicographic work with digital lexical resources, it is crucial not only to be able to extract all polylexical units but also to have the possibility to individualize them. That is why we need to go one step further and develop a mechanism for encoding different types of polylexical units.

decreto-lei [dɨkrɛtulɐ̃j]. *s. m. Dir.* Acto normativo proveniente do Governo da República. *Actualmente, os decretos-leis são publicados na primeira série-A do Diário da República.* Pl. decretos-leis.

Figure 5: Decreto-lei [decree-law] – DLCP (2001).

```
<entry xml:lang="pt" xml:id="decreto-lei" type="polylexicalUnit">
   <form type="lemma">
      <orth>decreto-lei</orth>
      <pron>dɨkrɛtul'ej</pron>
   </form>
   <gramGrp>
      <gram type="mwe" value="composto"/>
      <gram type="pos" norm="NOUN">s.</gram>
      <gram type="gen">m.</gram>
   </gramGrp>
   <!--etc. -->
</entry>
```

In Figure 5, the only addition to the encoding suggested in Salgado et al. (2019b) is the inclusion of `<gram type="mwe" value="composto"/>` to mark up the particular kind of polylexicality, even though this type of entry-level polylexicals is not explicitly labelled as such. For a detailed explanation of how one can encode different types of polylexical units, regardless of whether the given dictionary uses explicit labels for them or not, see Section 3.4 in this paper.

The situation with Latin expressions is slightly different because they are explicitly labelled in DLPC as such. See Figure 6:

fiat lux *loc. lat.* Exprime o desejo de que se torne clara alguma coisa importante.

Figure 6: *Fiat lux* – DLCP (2001).

DLPC labels the headword as *loc. lat.,* which stands for "locução latina" [Latin phrase]. This abbreviated label uses the same italic typeface in the same position as the label *s. m.* (substantivo masculino [masculine noun]), which we saw in the above example for **decreto-lei**. From DLPC's internal logic, one could argue that the label *loc. lat.* functions as a grammatical label. And yet, the two-partite structure of *loc. lat.* is internally different from that of *s. m.* While both part-of-speech and gender are grammatical categories, one can not say the same of *loc. lat.,* which combines grammatical and etymological information. Therefore, we recommend that this label be modelled as two different components: an `mwe` label for *loc. lat.,* which adequately represents the label of the source, and an `etym` element to explicitly mark up the language of origin.

```
<entry xml:lang="pt" xml:id="fiat_lux" type="polylexicalUnit">
   <form type="lemma">
      <orth>fiat lux</orth>
   </form>
   <gramGrp>
      <gram type="mwe" value="locução_latina">loc. lat.</gram>
   </gramGrp>
   <etym type="borrowing"><lang value="la"/></etym>
   <!--etc. -->
</entry>
```

The use of both grammatical and etymological tags is advantageous because it makes the same phrase findable in two different search contexts.

3.3 Encoding microstructurally relevant polylexical units

Microstructurally relevant polylexical units will be encoded differently in TEI Lex-0 depending on whether they are lexicographically transparent or not. Only the non-transparent ones will require full markup within an `<entry>` construct.

3.3.1 Encoding lexicographically transparent polylexical units

Following from our discussion in Section 2.2.2.1, the TEI Lex-0 encoding of lexicographically transparent polylexical units in DLPC should meet the following requirements:

1. each set of polylexical units should be grouped together to represent the microstructure of the entry adequately;
2. each polylexical unit should be identifiable as such for easy retrieval;
3. the explicit label "+" should be used only where it occurs in the dictionary text, but the implicit positioning of the headword in the given polylexical unit should be marked up as well.

Because lexicographically transparent polylexical units are *not* structured as mini-entries but are instead presented to the reader as a sequence of forms, we recommend to encode them as `<form>` elements:

```
<sense xml:id="descalçar.1">
   <!--etc.-->
   <form type="collocations">
      <form type="collocation">
         <orth>
            <ref type="oRef"><lbl>+</lbl></ref>
            <seg>as botas</seg>
         </orth>
         <gramGrp>
            <gram type="mwe" value="co-ocorrente_privilegiado"/>
         </gramGrp>
      </form>
      <pc>,</pc>
```

```
      <form type="collocation">
         <orth>
            <ref type="oRef"/>
            <seg>as luvas</seg>
         </orth>
         <gramGrp>
            <gram type="mwe" value="co-ocorrente_privilegiado"/>
         </gramGrp>
      </form>
      <pc>,</pc>
      <form type="collocation">
         <orth>
            <ref type="oRef"/>
            <seg>as meias</seg>
         </orth>
         <gramGrp>
            <gram type="mwe" value="co-ocorrente_privilegiado"/>
         </gramGrp>
      </form>
   </form>
   <pc>;</pc>
   <form type="collocations">
      <form type="collocation">
         <orth>
            <ref type="oRef"><lbl>+</lbl></ref>
            <seg>os sapatos</seg>
         </orth>
         <gramGrp>
            <gram type="mwe" value="co-ocorrente_privilegiado"/>
         </gramGrp>
      </form>
   </form>
   <pc>.</pc>
</sense>
```

The `<ref>` element of the type `oRef` (orthographic reference) is used to encode the position of the headword in the polylexical unit. Optionally, this element can contain a `<lbl>+</lbl>` to reflect the explicit headword substitution label.

3.3.2 Encoding lexicographically non-transparent polylexical units

A sense-related non-transparent polylexical unit can be encoded in TEI Lex-0 within an <entry> construct.[15] The type of the polylexical unit is indicated by the <gram> element, which is discussed in greater detail in the following section of this paper.

```
<entry type="monolexicalUnit" xml:lang="pt" xml:id="bombeiro">
   <form type="lemma">
      <orth>bombeiro</orth>
   </form>
   <!--etc. -->
   <sense xml:id="bombeiro.1">
      <!--etc. -->
      <entry xml:id="bombeiro_voluntario" xml:lang="pt" type="relatedEntry">
         <form type="lemma">
            <orth>bombeiro voluntário</orth>
         </form>
         <gramGrp>
            <gram type=”mwe” value=”combinatória_fixa”/>
         </gramGrp>
         <pc>,</pc>
         <sense xml:id=”bombeiro_voluntario.1”>
          <def>o que pertence a uma corporação com a obrigatoriedade de acudir
a incêndios, acidentes, unicamente por filantropia</def>
            <pc>.</pc>
         </sense>
      </entry>
      <entry xml:id="corpo_de_bombeiros" xml:lang="pt" type="relatedEntry">
         <form type="lemma">
            <orth>
               <ref type="entry"><seg>corpo</seg><lbl>+</lbl></ref>
               <seg>de bombeiros</seg>
            </orth>
         </form>
         <pc>.</pc>
      </entry>
   </sense>
   <!--etc.-->
</entry>
```

15 TEI and TEI Lex-0 diverge somewhat on how they allow this, but the end result is the same: in TEI Lex-0, the content model of <sense> allows elements from the class model.sensePart as its children, and <entry> is a member of this class; whereas in TEI <sense> has a broader content model which allows members of the class model. entryPart as its children.

Because sense-related polylexical units are modelled as nested entries, they can include domain labels as well. For instance (Figure 7):

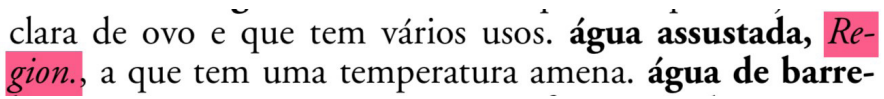
clara de ovo e que tem vários usos. **água assustada,** *Region.*, a que tem uma temperatura amena. **água de barre-**

Figure 7: Água assustada [mild water] – DLCP (2001).

```
<sense xml:id=”agua.4”>
    <!--etc.-->
    <entry xml:id="agua_assustada" xml:lang="pt" type="relatedEntry">
        <form type="lemma">
            <orth>água assustada</orth>
        </form>
        <pc>.</pc>
        <usg type="geographic" norm="regionalism">Region.</usg>
        <pc>,</pc>
        <sense xml:id=”agua_assustada.1”>
            <def>a que tem uma temperatura amena.</def>
        </sense>
    </entry>
    <!--etc.-->
</sense>
```

Sense-related polylexical units can themselves be polysemous. For instance (Figure 8):

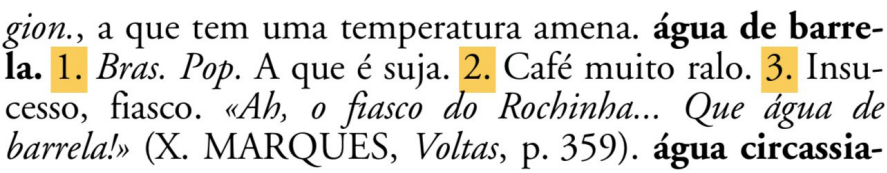
gion., a que tem uma temperatura amena. **água de barre-la.** 1. *Bras. Pop.* A que é suja. 2. Café muito ralo. 3. Insucesso, fiasco. *«Ah, o fiasco do Rochinha... Que água de barrela!»* (X. MARQUES, *Voltas*, p. 359). **água circassia-**

Figure 8: Água de barrela [dirty water; weak coffee; fiasco] – DLCP (2001).

```
<sense xml:id="agua.4">
    <!--etc.-->
    <entry xml:id="agua_de_barrela" xml:lang="pt" type="relatedEntry">
        <form type="lemma">
            <orth>água de barrela</orth>
        </form>
        <sense xml:id=”agua_de_barrela.1” n=”1”>
```

```
            <lbl>1.</lbl>
            <usg type=”geographic”>Bras.</usg>
            <usg type=”socioCultural”>Pop.</usg>
            <def>A que é suja.</def>
        </sense>
        <sense xml:id=”agua_de_barrela.2” n=”2”>
            <lbl>2.</lbl>
            <def>Café muito raro.</def>
        </sense>
        <sense xml:id=”agua_de_barrela.3” n=”3”>
            <lbl>3.</lbl>
            <def>Insucesso, fiasco.</def>
            <cit type=”example”>
                <quote>Ah, o fiasco do Rochinha... Que água de barrela!</quote>
                <bibl>
                    <author>X. MARQUES</author>
                    <pc>,</pc>
                    <title>Voltas</title>
                    <pc>,</pc>
                    <citedRange>p. 359</citedRange>
                </bibl>
            </cit>
        </sense>
    </entry>
    <!--etc.-->
</sense>
```

Entry-related polylexicals have the same structure as the sense-related ones, only they appear as children of the main entry:

```
<entry type="monolexicalUnit" xml:lang="pt" xml:id="dali">
   <form type="lemma">
      <orth>dali</orth>
   </form>
   <!--etc.-->
   <entry xml:id="dali_a_nada" xml:lang="pt" type="relatedEntry">
      <form type="lemma">
         <orth>dali a nada</orth>
      </form>
      <gramGrp>
         <gram type="mwe" value="locução_adverbial">loc. adv.</gram>
      </gramGrp>
      <pc>,</pc>
```

```
      <sense xml:id="dali_a_nada.1">
         <def>muito pouco tempo depois</def>
         <pc>.</pc>
         <cit type=”example”>
            <quote>Dali a nada estava ele a chatear-me.</quote>
         </cit>
      </sense>
   </entry>
   <!--etc.-->
</entry>
```

The same type of encoding applies to idiomatic expressions:

```
<entry type="monolexicalUnit" xml:lang="pt" xml:id="dura">
   <form type="lemma">
      <orth>dura</orth>
   </form>
   <!--etc.-->
   <entry xml:id="ser_de_pouca_dura" xml:lang="pt" type="relatedEntry">
      <form type="lemma">
         <orth>ser de pouca dura</orth>
         <gramGrp>
            <gram type=”mwe” value=”expressão_idiomática”/>
         </gramGrp>
      </form>
      <pc>,</pc>
      <sense xml:id=”ser_de_pouca_dura.1”>
         <def>durar pouco tempo; passar depressa</def>
         <pc>.</pc>
         <cit type=”example”>
            <quote>Foi amor de pouca dura.</quote>
         </cit>
      </sense>
   </entry>
   <!--etc.-->
</entry>
```

3.4 Encoding types of polylexical units

We saw above that some polylexical units in DLPC are explicitly labelled as such (for instance *loc. lat.* or *loc. adv.*, but some are not – for instance, hyphenated compounds as headwords, or idiomatic expressions. TEI Lex-0 should

provide a consistent but flexible mechanism for labelling types of polylexical units in dictionaries regardless of whether these labels exist explicitly in the dictionary source or not. We propose to encode this information using the existing TEI `gramGrp/gram` mechanism, in order to have the maximum flexibility to cover these three distinct types of labels:

1. *implicit labels*, i.e., those labels whose value can only be deduced from its typographical properties or its position in the entry structure, but are not present on the dictionary page (for instance, compounds as headwords in DLPC);
2. *explicit labels*, i.e. labels which appear on the dictionary page (for instance, *loc. adv.* in DLPC);
3. *normalised labels*, i.e. normalised versions of either implicit or explicit labels, which can be used to improve the interoperability of the labels.

The consistent labelling of polylexical units in a dictionary can be achieved by adopting the following principles:

1. Any polylexical unit should be identified by the presence of a generic element-attribute combination: `<gram type="mwe"/>`. Without any further classification, `<gram type="mwe"/>` does not tell us anything about the specific type of the polylexical unit.
2. Explicit labels should be encoded as text nodes of `gram`: `<gram type="mwe">loc. adv.</gram>`.
3. Implicit labels should be placed in the `@value` attribute.
4. Normalised values should be placed in the `@norm` attribute.

In addition to being encoded as text nodes, explicit labels should, for the sake of consistency with implicit labels, also use the `@value` attribute. This is to avoid situations in which some labels are encoded as text and some as attributes. The consistent use of the `@value` attribute for both explicit and implicit labels will make it easier to retrieve all labels of a specific type regardless of how they are labelled in the text of the dictionary. Also, it is important to emphasize that the `@value` and `@norm` attributes should be kept conceptually distinct: the former should be used as a locally non-ambiguous identifier of both

the explicit and implicit labels in a given dictionary; the latter, on the other hand, should be optionally used as a placeholder for a dictionary-independent classification of the local label.

```
<gramGrp>
    <!--NOT RECOMMENDED: explicitly labelled MWE as text node only -->
    <gram type="mwe">loc. adv.</gram>
    <!--RECOMMENDED: explicitly labelled MWE as text node + attribute -->
    <gram type="mwe" value="locução_adverbial">loc. adv.</gram>
    <!--implicitly labelled MWE-->
    <gram type="mwe" value="co-ocorrente_privilegiado"/>
    <!--more work needed: normalizing values-->
    <gram type=”mwe” value=”locução_adverbial” norm=”???”>loc. adv.</gram>
    <gram type=”mwe” value=”co-ocorrente_privilegiado” norm=”???”/>
</gramGrp>
```

A typology of labels for polylexical units that would work across multiple dictionaries and languages would be needed if we were to suggest possible values for the `@norm` attribute. Neither TEI nor TEI Lex-0 currently refers to any such typology. However, such a typology would be very helpful for any work on aligning multiple dictionaries, studying them in parallel or pooling various lexical resources together. For instance, in DLPC, the Latin phrase **habeas corpus** is a headword labelled as *loc. lat.* [Latin phrase] but the same polylexical unit in the *Grande Dicionário Houaiss da Língua Portuguesa* (Houaiss, 2015) is labelled as *loc. subst.* [locução substantiva; noun phrase] and “[lat.]”, which is an explicit label for Latin etymology. A typology of polylexical units would make it possible to normalize both explicit and implicit labels across different dictionaries.

4 CONCLUDING REMARKS

Our recommendations for encoding polylexical units using TEI Lex-0 show that TEI Lex-0 is fully capable of consistently marking up polylexical units as constituent parts of the dictionary macro- and microstructure, regardless of whether they appear as headwords in independent entries, or in nested entry-like structures inside entries for monolexical units. The use of nested `<entry>` elements to encode polylexical units inside dictionary entries is a robust mechanism which can take care of all kinds of lexicographic constraints

imposed on the description of polylexical units (polysemy, domain labels, grammatical labels etc.), whereas the combination of `<gram>` element and attributes `@type`, `@value` and `@norm` can be used consistently to encode explicit, implicit and normalised versions of the labels.

In this paper, we focused on the formal representation of polylexical units as they appear on the page of a single dictionary because we wanted to document the process of translating lexicographic and typographic conventions from linear text strings to hierarchical, tree-like structures using the vocabulary and syntactic constraints of TEI Lex-0. While further comparative work will be needed to validate our recommendations on a larger sample, the process we described in this paper and the markup solutions we proposed are sufficiently abstract to serve as a basis for marking up the lexical view of polylexical items in various dictionaries, even though we can expect to see more pronounced differences in their editorial and typographic views. When it comes to designing and applying TEI Lex-0 markup to dictionary entries, the question of whether a dictionary is a paper dictionary, a retrodigitised one or a born-digital resource is of little consequence: what matters is that one can consistently identify, represent and validate all the microstructural elements in a given dictionary entry using a standardised vocabulary.

As we could see in the penultimate section of this paper, the interoperability of encoded lexical resources would be significantly improved if dictionary encoders would have access to a typology of polylexical units that was both expressive and straightforward enough to apply when modelling lexical data. It would be safe to say that very detailed typologies, like the one proposed by Bergenholtz (2013), which includes twenty different types of MWEs, would be challenging to implement in practice. That is why more work on the classification of polylexical items *specifically for encoding purposes* will be necessary. One could argue that there is "no hope of finding a single classification or taxonomy of polylexical units that can be used for all purposes" (Sailer, 2018, p. vi), but a comparative study of multiple dictionaries in different languages would bring us one step closer to proposing, discussing and eventually agreeing on a sensible typology that could be used in the context of TEI Lex-0 as a set of attribute values for normalizing local lexicographic classifications. We hope to pursue this line of work in the future.

Acknowledgements

This paper is supported by European Union's Horizon 2020 research and innovation programme under grant agreement No 731015 (ELEXIS) (European Lexicographic Infrastructure), and 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.

REFERENCES

Dictionaries

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

Dictionnaire des Expressions et Locutions. (1993). Alain Rey and Sophie Chantreau (Eds.). Col. Les Usuels. Paris: Éd. Dictionnaires Le Robert.

Grande Dicionário Houaiss da Língua Portuguesa. (2015). Instituto António Houaiss Bloco Gráfico, Lda. Lisboa: Círculo de Leitores.

Websites

DARIAH WG = *Lexical Resources and the H2020-funded European Lexicographic Infrastructure (ELEXIS)*. Retrieved from https://github.com/DARIAHERIC/lexicalresources/tree/master/Schemas/TEILex0 (23. 2. 2020)

TEI Consortium (Ed.) = *TEI P5: Guidelines for Electronic Text Encoding and Interchange* (2019). Version 3.5.0. [Last updated on 29th January 2019, revision 3c0c64ec4.] TEI Consortium. Retrieved from http://www.tei-c.org/Guidelines/P5/ (23. 2. 2020)

Other

Atkins, B. T. S., & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

Baldwin, T., & Kim, S. (2010): Multiword Expressions. In N. Indurkhya & F. J. Damerau (Eds.), *Handbook of Natural Language Processing* (2nd ed., pp. 267–292). Boca Raton, USA, CRC Press.

Bergenholtz, H., & Gouws, R. (2013). A Lexicographical Perspective on the Classification of Multiword Combinations. *International Journal of Lexicography*, *27*(1), 1–24. doi: 10.1093/ijl/ect031

Calzolari, N., Fillmore, C. J., Grishman, R., Ide, N., Lenci, A., MacLeod, C., & Zampolli, A. (2002). Towards Best Practice for Multiword Expressions in Computational Lexicons. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2002)* (pp. 1934–1940). Spain: Las Palmas, Canary Islands.

Considine, J. (2014). Academy Dictionaries 1600-1800. Cambridge, New York: Cambridge University Press.

Cowie, A. P. (1994). Phraseology. In R. E. Asher (Ed.), *The Encyclopedia of Language and Linguistics* (pp. 3168-3171). Oxford, UK: Pergamon.

Cowie, A. P. (Ed.). (1998). *Theory, Analysis, and Applications*. Oxford: OUP.

Fellbaum, C. (2016). Treatment of Multi-Word Units. In P. Durkin (Ed.), *The Oxford Handbook of Lexicography* (pp. 411–424). Oxford: Oxford University Press.

Fontenelle, T. (1997). *Turning a Bilingual Dictionary into a Lexical-Semantic Database*. Tübingen: Niemeyer.

Gantar, P., Colman, L., Parra Escartín, C., & Martínez Alonso, H. (2018). Multiword Expressions: Between Lexicography and NLP. *International Journal of Lexicography, 32*(2), 138–162. doi: 10.1093/ijl/ecy012

Hausmann, F. J. (1979). Un Dictionnaire des Collocations Est-Il Possible? *Travaux de Linguistique et de Littérature, 17*(1), 187–195.

ISO 24613-1 (2019). Language Resource Management — Lexical Markup Framework (LMF) — Part 1: Core Model. Genève: Organisation Internationale de Normalisation.

Jónsson, J. H. (2009). Lemmatisation of Multiword Lexical Units: Motivation and Benefits. In H. Bergenholtz, S. Nielsen & S. Tarp (Eds.), *Lexicography at a Crossroads. Dictionaries and Encyclopedias Today, Lexicographical Tools Tomorrow* (pp. 165–194). Bern: Peter Lang AG.

Kinable, D. (2015). Reflections on the Concept of a Scholarly Dictionary. *Kernerman Dictionary News*, *23*, 11–2.

Lorentzen, H. (1996). Lemmatization of Multi-word Lexical Units: In Which Entry? In M. Gellerstram et al. (Eds.), *Proceedings of the 7th EURALEX*

International Congress on Lexicography: Part I (pp. 415–421). Goteborg, Sweden: Goteborg University Department of Swedish.

McCrae, J. P., Tiberius, C., Khan, F., Kernerman, A., Declerck, T., Krek, S., Monachini, M., & Ahmadi, S. (2019). The ELEXIS interface for interoperable lexical resources. In I. Kosem, T. Zingano Kuhn, M. Correia, J. P. Ferreira, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek & C. Tiberius (Eds.), *Electronic Lexicography in the 21st Century: Smart Lexicography. Proceedings of the eLex 2019 Conference* (pp. 417–433). Brno: Lexical Computing CZ, s.r.o. Retrieved from https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_37.pdf

Mel'čuk, I., Arbatchewsky-Jumarie, N., Iordanskaja, L., Mantha, S., & Polguère, A. (1984–1999). Dictionnaire Explicatif et Combinatoire du Français Contemporain. *Recherches lexico-sémantiques, IV*. Montréal: Les Presses de l'Université de Montréal.

Mel'čuk, I. (1998). Collocations and Lexical Functions. In A. P. Cowie (Ed.), *Phraseology, Theory, Analysis, and Applications* (pp. 23–54). Oxford: Oxford University Press.

Moon, R. (1998). *Fixed Expressions and Idioms in English: A Corpus-Based Approach*. Oxford: Clarendon Press.

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). Retrieved from https://tei2018.dhii.asia/AbstractsBook_TEI_0907.pdf

Sailer, M., & Markantonatou, S. (2018). *Multiword expressions: Insights from a multilingual perspective (Phraseology and Multiword Expressions): Vol. 1*. Berlin: Language Science Press. doi: 10.5281/zenodo.1182583

Salgado, A., Costa, R., Tasovac, T., & Simões, A. (2019a). Improving the Consistency of Usage Labelling in Dictionaries with TEI Lex-0. *Lexicography: Journal of ASIALEX 6*(2), 133–156. doi: 10.1007/s40607-019-00061-x

Salgado, A., Costa, R., & Tasovac, T. (2019b). TEI Lex-0 In Action: Improving the Encoding of the Dictionary of the Academia das Ciências de Lisboa. In I. Kosem, T. Zingano Kuhn, M. Correia, J. P. Ferreira, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek & C. Tiberius (Eds.), *Electronic Lexicography in the 21st Century: Smart Lexicography. Proceedings of the*

eLex 2019 Conference, 1–3 October, 2019, Sintra, Portugal (pp. 417–433). Brno: Lexical Computing CZ, s.r.o. Retrieved from https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_23.pdf

Simões, A., Almeida, J. J., & Salgado, A. (2016). Building a Dictionary using XML Technology. In *Open Access Series in Informatics (OASIcs). 5th Symposium on Languages, Applications and Technologies (SLATE'16): Vol. 51* (pp. 14:1–14:8). Germany, Dagstuhl: Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Svensén, B. (2009). *A Handbook of Lexicography: The Theory and Practice of Dictionary Making*. Cambridge: Cambridge University Press.

Tasovac, T., & Petrović, S. (2015). Multiple Access Paths for Digital Collections of Lexicographic Paper Slips. In I. Kosem, M. Jakubíček, J. Kallas & S. Krek (Eds.), *Electronic Lexicography in the 21st Century: Linking Lexical Data in the Digital Age. Proceedings of the eLex 2015 Conference* (pp. 384–396). Ljubljana/Brighton: Institute for Applied Slovene Studies and Lexical Computing Ltd. Retrieved from https://elex.link/elex2015/proceedings/eLex_2015_25_Tasovac+Petrovic.pdf

Zgusta, L. (1971). *Manual of Lexicography*. Prague: Academia; The Hague/Paris: Mouton.

KODIRANJE VEČBESEDNIH LEKSIKALNIH ENOT S TEI LEX-O: ŠTUDIJA PRIMERA

Modeliranje in kodiranje večbesednih leksikalnih enot oz. pogostih nizov leksemov, ki jih obravnavamo kot samostojne leksikalne enote, je tematika, ki v smernicah Text Encoding Initiative (TEI) ni ustrezno in dovolj poglobljeno predstavljena, čeprav je TEI v raziskovalni skupnosti de facto standard pri delu z elektronskimi besedili. V prispevku na primeru Slovarja Portugalske akademije znanosti predstavimo nekatere rešitve pri kodiranju večbesednih leksikalnih enot v formatu TEI Lex-0, iniciative, katere namen je poenostaviti in racionalizirati kodiranje leksikalnih podatkov s TEI in posledično izboljšati interoperabilnost. Vpeljemo pojem makro- in mikrostrukturne relevantnosti z namenom razločevati med večbesednimi leksikalnimi enotami, ki so samostojne slovarske iztočnice, in tistimi, ki se nahajajo v geslih enobesednih iztočnic. Vpeljemo tudi pojem leksikografske transparentnosti za razlikovanje med enotami, ki nimajo razlage, in tistimi, ki jo imajo; prve so kodirane v okviru elementa <form>, slednje pa v okviru elementa <entry> in lahko vsebujejo nadaljnje omejitve (številke pomenov, področne oznake, slovnične oznake ipd.). V elementu <gram> vpeljemo uporabo atributov za kodiranje različnih tipov oznak za večbesedne leksikalne enote (implicitne, eksplicitne in normirane). Prispevek zaključimo s sklepom, da bi se interoperabilnost leksikalnih virov močno izboljšala, če bi avtorji slovarskih shem imeli dostop do bogate, a relativno enostavne tipologije večbesednih leksikalnih enot.

Ključne besede: TEI, leksikografija, jezikovni viri, večbesedne leksikalne enote, interoperabilnost