TEI Lex-0 In Action: Improving the Encoding of the Dictionary of the *Academia das Ciências de Lisboa*

Ana Salgado[1], Rute Costa[1], Toma Tasovac[2], Alberto Simões[3]

[1] NOVA CLUNL, Universidade NOVA de Lisboa

[2] Belgrade Center for Digital Humanities, Serbia

[3] 2Ai – Instituto Politécnico do Cávado e do Ave / Algoritmi, Universidade do Minho

E-mail: anasalgado@campus.fcsh.unl.pt; rute.costa@fcsh.unl.pt; ttasovac@humanistika.org; asimoes@ipca.pt

Abstract

This paper describes some experiments made while encoding the first complete dictionary of the *Academia das Ciências de Lisboa* (DACL) in the context of TEI Lex-0, a community-based interchange format for lexical data aimed at facilitating the interoperability and reusability of lexical resources. Even though the original encoding of the DACL was based on TEI, we decided to switch to TEI Lex-0 because it allowed us to streamline our encoding. Our experiments show that even though TEI Lex-0 is stricter than TEI itself (allowing fewer elements and imposing certain constraints that are not present in plain TEI), it is fully capable of representing the complexities of the entry structure of the DACL. In the paper, we discuss the TEI Lex-0 encoding of the DACL, as well as the conversion methodology and the tools used for the automatic conversion from the original encoding. We are currently focusing on the macrostructural level, more precisely on the types of lexical units and on the written and spoken forms of the lemma, providing a set of modelling principles and representation forms of every type of entry in the DACL. This paper is part of ongoing work and a contribution to the efforts of the DARIAH-ERIC Lexical Resources working group.

Keywords: dictionary encoding; lexicography; TEI; XML; TEI Lex-0

1. Introduction

The digital revolution has transformed the way we conceptualize, plan and implement lexicographic projects. While print dictionaries are slowly going out of fashion, retro-digitized and born-digital dictionaries are increasingly taking advantage of the available technologies. At the same time, however, many dictionaries continue to be designed and implemented following the typographical and editorial conventions of the print medium (Tasovac, 2010: 1). According to Trap-Jensen (2018: 34), "it is necessary that lexicographers shift their focus away from the concrete end product and towards a lexical database".

The task of updating the first complete dictionary – from A to Z – of the *Academia das Ciências de Lisboa* (DACL), published in 2001, provides the basis for this work. Its great historical value for European lexicographical heritage and the institution's willingness to update the content of the dictionary dictated the need to convert the print edition into digital format, with the ultimate goal of making this lexical resource available on the web and as a mobile app.

This dictionary – available in print and as a PDF document – was converted into XML using the P5 schema of the Text Encoding Initiative (TEI) (Simões et al., 2016). This process – as described in detail in Section 2 – was conducted with a formal format in mind, and therefore the group focused on the conceptual structure of the dictionary and not on its visual aspect. Nevertheless, the TEI format, although very complete and accompanied by comprehensive documentation, presented some challenges when encoding the DACL. Parts of the original structure diverged from the TEI proposed structure, which led to some adaptations of the official schema. This problem, coupled with the fact that TEI allows multiple solutions for encoding the same type of information, made us look into TEI Lex-0[1] (Romary & Tasovac, 2018), a streamlined version of the TEI standard for dictionaries. In Section 3, we discuss and compare these two standards.

Before we could work on the conversion between these two formats, we had to analyse the TEI Lex-0 schema and create maps from the original structure used in the DACL. Section 4 refers to this analysis, a contribution to the work developed by the DARIAH-ERIC Lexical Resources working group[2]. Section 5 discusses the technological approach used to experiment with the conversion of the DACL into TEI Lex-0, trying to accommodate every change that the TEI Lex-0 working group has published. As the standard has yet to be concluded, our technological architecture is prepared to aim at a moving target, adapting the encoding as the standard evolves. Finally, Section 6 draws some conclusions from the work that has been carried out so far.

2. Dictionary of the *Academia das Ciências de Lisboa*

As previously mentioned, the first complete edition of the DACL was only published in 2001 in a two-volume paper version (the first volume from A to F, and the second from G to Z). At the time, the computational side of the project included a Microsoft Access database and a reporting tool that could generate a Word file for the dictionary, which was then manually edited and formatted before printing. Eighteen years later, the only surviving digital data source of the published dictionary remains the final PDF file. For the Portuguese Academy to move forward and produce a new edition of the dictionary[3] using digital tools and structured data, the PDF file had to be reverse engineered in order to convert PDF strings and their typographic features into a

[1] https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html

[2] https://www.dariah.eu/activities/working-groups/lexical-resources/

[3] The original digital version of the DACL is not publicly available, but the first author of this paper is the coordinator of the new digital edition. The Natural Language Processing group of the Computer Science Department of the University of Minho has been developing the technological support of the new digital edition of DACL, counting on the participation of Alberto Simões from IPCA (Instituto Politécnico do Cávado e do Ave), responsible for the technological support, José João Almeida, and the consultancy of Álvaro Iriarte Sanromán, both from University of Minho. The participation of NOVA CLUNL (Linguistic Research Center of NEW University of Lisbon) is related to its transition into the TEI LEX-0 format.

conceptual structure. A mapping from different font typefaces and font sizes was made to specific structures (e.g., phonetic transcriptions or synonyms). Because the same font typeface and font size were used for different types of information, a heuristic procedure had to be employed, taking into account string content and string order, to infer their semantics. The TEI schema was used as the target format, since it is a well-known and documented format. Nevertheless, as already stated, some specific constructions of the standard had to be changed in order to enable the encoding of some of the dictionary entries. This process was iterative and interactive, with human interaction to fix minor issues on some entries where the default behaviour was not able to correctly determine the entry structure.

To allow the quick edition of the database, the TEI dictionary was split into thousands of small XML documents (one per dictionary entry) that were imported into a native XML database (eXist-DB). Using the eXist-DB ecosystem based on XQuery, LeXmart[4], a tool framework for lexicographic work, was developed to allow the edition, deletion and creation of new dictionary entries, as well as to validate their structure and overall dictionary coherence (Simões et al., 2016).

3. TEI Guidelines for Dictionary Encoding

The use of open formats based on standards is a crucial aspect of digital humanities initiatives. TEI is a *de facto* standard for the digital encoding of all types of written texts, ranging from novels and poetry to mathematical formulae or music notation[5]. It also defines how specific humanities resources, such as speech, morphological annotated monolingual and parallel corpora, dictionaries and other structures should be encoded.

All TEI documents must include a metadata section, named TEI header, and share a set of common annotation features, defined in the standard as the core module (Chapter 3)[6]. This set includes structural elements, such as paragraphs, lists or bibliographic references.

For dictionaries, Chapter 9[7] of the TEI Guidelines starts by defining the structure of the dictionary as a book – front matter, body or back matter. It also describes three main elements to encode dictionary entries: `entry`, `entryFree` and `superEntry`. While the document describes precisely when each should be used (`entry` forces a structure; `entryFree` provides a flat representation and allows unstructured entries that should be avoided but may be necessary for some dictionaries; and `superEntry` as a mechanism that can group other entries, such as homonyms), this freedom makes

[4] http://www.lexmart.eu/

[5] See, e.g., Music Encoding Initiative: https://music-encoding.org/

[6] Elements Available in All TEI Documents: https://tei-c.org/Vault/P5/1.3.0/doc/tei-p5-doc/es/html/CO.html

[7] Dictionaries: https://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html

it difficult for different authors to keep their dictionaries coherent in terms of structure. To these three tags we can add the `re` element, which allows the encoding of related entries (Bański et al., 2017: 485) and `hom` (homograph) which can be used for encoding homographs.

Flexibility is both a virtue and a shortcoming of TEI. For instance, to create cross-references, the preferred way is to use the `xr` tag. But it is also possible to create links using `anchor/ptr` or `link`. In order to reduce this freedom and define a specific format for dictionaries forcing dictionary encoders to follow the same structural rules, the lexicographic and dictionary-encoding communities are currently discussing a new format to encode dictionaries – TEI Lex-0[8] – a fully TEI-compliant but streamlined format for facilitating interoperability.

This new format does not intend to replace the Dictionaries Chapter in the TEI Guidelines. Instead, it is framed as a target format that can help uniformize the existing heterogeneously encoded lexical resources and is currently being tested by numerous dictionaries[9]. Given the fact that it is still a work in progress, it can be changed in order to accommodate relevant dictionary structures.

a[5] [ɐ]. *prep.* (Do lat. *ad* 'para' ou *ab* 'de'). **A** Valores semânticos: **I.** Na expressão de valores locativos, indica: **1.** Direcção para um lugar (real ou virtual). *O navio rumou a oriente. Levar a uma situação embaraçosa. Foi a casa dos sogros. Eu apenas fui a Paris; o meu irmão é que foi para Paris.* Obs. Quando introduz um complemento do verbo *ir* ou do nome *ida*, indica que a permanência no lugar de destino é breve; inversamente o uso da preposição *para* indica permanência prolongada. **2.** Termo de um movimento. *Chegou a casa.* **3.** Afastamento. ≃ DE. *Esquivar-se a trabalhos.* **4.** Distância medida em unidades de espaço ou tempo. *Há uma estação de comboio a quinhentos metros daqui. A minha casa fica a cinco minutos do mercado.* **5.** Localização, situação precisa ou aproximada. *Ela mora num palacete a São Bento. Pôs as cadeiras a todo o comprimento da sala.* **6.** Adjunção. *Amarrou o cão a um poste. A uma asneira seguiu-se outra.* **II.** Na expressão de valores temporais, indica: **1.** Tempo em que uma coisa acontece (pontual ou habitualmente); concomitância. *Tenho aulas a meio da tarde.* **2.** Distância. *O jogo está a dez minutos do intervalo. A cinco horas do desembarque.* **3.** Progressão para um tempo (em correlação com a prep. *de*). *De mês a mês. De cinco dias a esta parte. A exposição estará aberta ao público de Junho a Setembro.* **4.** Intervalo regular ou duração periódica. *Ele trabalhava a tempo inteiro. Há muitos contratos a prazo.* **III.** Na expressão de outros valores, indica: **1.** Causa. ≃ POR. *Fez isso a solicitação dos parentes.* **2.** Instrumento, meio e modo. *Pintura a óleo. Navegava a todo o vapor. O móvel apresentava entalhaduras a canivete. Há quem aguente muito tempo a pão e água.* **3.** Finalidade. ≃ PARA. *O patrão deu-lhe vinho a beber. Pôs*

Figure 1: *a* preposition (DACL).

[8] To secure interoperability, the Working Group "Retro-digitised Dictionaries", lead by Toma Tasovac and Vera Hildenbrandt, as part of the COST Action European Network of e-Lexicography (ENeL) started the establishment of TEI Lex-0. Then, TEI Lex-0 was taken up by the DARIAH Working Group "Lexical Resources" which is co-chaired by Laurent Romary and Toma Tasovac. Currently, the work on TEI Lex-0 is conducted by the DARIAH WG "Lexical Resources" and the H2020-funded European Lexicographic Infrastructure (ELEXIS).

[9] TEI is the basis for a large number of current lexicographic projects, such as Nénufar, ARTFL, or VICAV.

Although we followed TEI in the DACL encoding, we could not find solutions in the Guidelines that covered all the microstructural elements of the dictionary (e.g., the entry *a*, preposition, contains different types and levels of information – grammatical, semantic, pragmatic) – which made us adapt the standard features.

Considering the example referred to above, as can be seen in Figure 1, the sections that begin with an "A" or "I." are not actually 'definitions' of the headword. The information "Valores semânticos" [*Semantic values*] or "Na expressão de valores locativos […]" [*When expressing locative values* [...]] indicates the properties of the preposition. In the original encoding we used the `def` element to encode the description and created a grouping mechanism (named `group`) that can be used recursively to create as many levels as needed.

4. TEI Lex-0 encoding of the DACL

In order to have an interoperable lexical database and aiming at dictionary content reusability, we intend to convert the DACL into TEI Lex-0 encoding, especially if it allows us to encode the complete extension of the dictionary structure without any kind of adaptation. Therefore, we present some experiments on the encoding of specific parts of the dictionary entries.

It is important to stress that the TEI Lex-0 working group is aiming at a standard that is able to encode a dictionary taking into account its structure and semantic meaning for each specific part of the entry, and not how it looks visually. While the authors agree that there may be cases where the latter approach is useful (namely for the digital preservation of ancient documents), the development of new lexical resources should take into account their own structure. This is crucial if the goal of the lexical resource is not only to be used by humans but also by Natural Language Processing algorithms.

For our experiment, we started by identifying every element in the dictionary. A typical entry includes the following elements: headword, pronunciation, usually followed by some linguistic information (e.g., part-of-speech), the different meanings, usage information, synonyms, antonyms, collocations, etymology, and notes. Examples of usage, cross-references, etc., may also be present.

In a TEI-style encoding, each of these or even other elements of an entry must be distinguished as clearly as possible.

4.1 Macrostructural level: different types of lexical items

In order to be able to define a valid approach to annotate all the entries of the dictionary, we performed an analysis of the different types of lexical units that can be headwords, so that a sample entry for each type was chosen and encoded, enabling us to understand the versatility of the standard. Thus, we first worked on the macrostructural level of the dictionary.

At an initial stage, we listed all the entries of the DACL and identified all the types of lexical units that are summarized in Figure 2: monolexical unit, polylexical unit, affix and abbreviation.

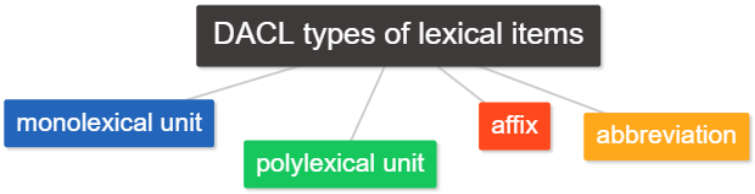

Figure 2: Formal representation of lexical entries (DACL).

In the following chapter, we will illustrate each type of lexical item found in DACL.

4.2 TEI encoding of different types of lexical items

In TEI encoding, the outermost structural level of an entry is marked with the `entry` element that begins with information about the form of the headword – `form` element – i.e., information on the written and spoken forms of one headword related to the description of its spelling and phonetics. The different types of entries are currently being marked with the attribute `type` into the `entry` element. As of this writing, there is no complete agreement within the TEI Lex-0 community on where to encode this information. Currently, as shown below, we are still adding this property to the whole entry. Nevertheless, as this is grammatical information, it should probably be encoded together with the morphologic information.

To illustrate the application of TEI Lex-0, we present the original encoding of the lemma and the conversion to TEI Lex-0 of some entries of the DACL for each of the entry types illustrated in Figure 2.

4.2.1 Monolexical units

Monolexical units can be divided into two types: lexical units, such as nouns, adjectives, verbs and grammatical units, such as conjunctions, determiners, prepositions, and pronouns.

palácio [pɐlásju]. *s. m.* (Do lat. *palatium*). **1.** Edifício sumptuoso, de grandes dimensões, geralmente construído num espaço urbano e destinado a residência da família real, de personalidades nobilitadas, de dignidades eclesiásticas ou altas individualidades. + *ducal, episcopal, presidencial, real.* **2.** Edifício sumptuoso, de dimensões significativas, onde se encontram sediados determinados organismos públicos. **palácio da justiça,** edifício, em cada localidade, onde funcionam os serviços judiciais e se realizam os julgamentos. *Um advogado seu amigo trabalha no palácio da justiça.* **3.** Casa solarenga, ampla, sumptuosa que lembra um palácio. **olhar para** *alguma coisa* **como boi para palácio.** 1. Não perceber nada de alguma coisa. 2. Não ligar importância; não dar valor, apreço a. Dim. palacete.

Original encoding

```
<entry id="palácio">
  <form>
    <orth>palácio</orth>
    <pron>pɐlˈasju</pron>
  </form>
  <gramGrp>s. m.</gramGrp>
<!--etc. -->
</entry>
```

Conversion to TEI Lex-0

```
<entry type="monolexicalUnit" xml:lang="pt"
xml:id="palácio">  <form type="lemma">
      <orth>palácio</orth>
      <pron>pɐlˈasju</pron>
   </form>
  <gramGrp>
        <gram type="pos" norm="NOUN">s.</gram>
        <gram type="gen">m.</gram>
  </gramGrp>
<!--etc. -->
</entry>
```

Example 1: DACL monolexical unit – original encoding and conversion from TEI to TEI Lex-0.

As can be seen in Example 1, in TEI Lex-0, `entry` is used to encode the basic element of the dictionary microstructure and requires the attributes `xml:id` and `xml:lang` in compliance with ISO Standard 16642 for terminological data.

Note that TEI Lex-0 schema only allows `entry` to be used to typeset entries – the `entryFree`, `superEntry` and `re` elements of the TEI Guidelines are not allowed. As for the DACL itself, only `entry` and `re` were being used, and therefore little adaptation was needed at this point.

Lexicographical articles always start with a lemma (headword), which is a non-inflected unit considered as the canonical form. The lemma is encoded using the `form` element with the attribute `type` and value "lemma". The `orth` element (orthographic form) gives the orthographic form of the headword.

Sometimes the lemma is a borrowed word. In TEI encoding, a unit borrowed from a foreign language is identified within the TEI element `etym`, where etymologic information is encoded, and labelled with the attribute `type` and the value "borrowing" (Bowers & Romary, 2017), as exemplified in Example 2.

workshop [wórkʃɔp]. *s. m.* (Ingl.). Reunião destinada à discussão ou realização de trabalho prático sobre um assunto específico, em que é feita uma aprendizagem através da troca de conhecimentos e experiências. «*Durante o 'workshop' sobre a articulação dos hospitais com os tribunais, foi visível a desconfiança de algumas pessoas*» (*DN*, 21.2.1992). Pl. workshops.

Original encoding

```
<entry id="workshop">
  <form>
    <orth>workshop</orth>
    <pron>wˈorkʃɔp</pron>
  </form>
  <gramGrp>s. m.</gramGrp>
  <etym>Ing.</etym>
<!--etc. -->
</entry>
```

Conversion to TEI Lex-0

```
<entry type="monolexicalUnit" xml:lang="en"
xml:id="workshop">
  <form type="lemma">
    <orth>workshop</orth>
    <pron>wˈorkʃɔp</pron>
  </form>
  <gramGrp>
    <gram type="pos" norm="NOUN">s.</gram>
    <gram type="gen">m.</gram>
  </gramGrp>
  <etym type="borrowing"><lang>Ing.</lang></etym>
<!--etc. -->
</entry>
```

Example 2: DACL borrowed word – original encoding and conversion from TEI to TEI Lex-0.

The lexical units formed from other units or bases – derivative lexical units (e.g. *infeliz* [unhappy]; *ensonado* [sleepy]) – are also classified as monolexical units, as shown in Example 3.

ensonado, a [ẽsunádu, -ɐ]. *adj.* (De *en-* + *sono* + suf. *-ado*). Que tem ou está com sono. ≃ SONOLENTO. «*Sertório assoma à porta do quarto: vem, ensonado, a esfregar os olhos.*» (D. MOURÃO-FERREIRA, *Gaivotas em Terra*, p. 139).

Original encoding

```
<entry id="ensonado">
  <form>
    <orth fem="a">ensonado</orth>
    <pron>ẽsunˈadu, -e</pron>
  </form>
  <gramGrp>adj.</gramGrp>
<!--etc. -->
</entry>
```

Conversion to TEI Lex-0

```
<entry type="monolexicalUnit" xml:lang="pt"
xml:id="ensonado">
  <form type="lemma">
    <orth>ensonado</orth>
  </form>
  <form type="inflected">
    <orth>ensonado</orth>
    <pron>ẽsunˈadu</pron>
  <gramGrp>
     <gram type="gen">m.</gram>
  </gramGrp>
</form>
  <form type="inflected">
    <orth>ensonada</orth>
    <pron>ẽsunˈade</pron>
  <gramGrp>
     <gram type="gen">f.</gram>
  </gramGrp>
</form>
  <gramGrp>
     <gram type="pos" norm="ADJ">adj.</gram>
  </gramGrp>
<!--etc. -->
</entry>
```

Example 3: DACL monolexical lexical units – original encoding and conversion from TEI to TEI Lex-0.

This last example also shows that, when a specific inflected form is featured in the entry, it should be clearly defined as an independent form, and have enough information about the inflected type (in this case, that the item is a feminine form).

For the grammatical information, the TEI Lex-0 standard suggests the use of the `gramGrp` tag. This element can be used in two different places: as a sibling of the `form` element, when the annotation is referring to all the forms present in the entry, or as a child of the `form` element, when the information is specific for that form.

As XML is verbose enough, for DACL annotations will appear mostly following the `form` element, and when used inside it, it will describe only the properties that differ for that form. This way, in the example above, we do not repeat the information about the part-of-speech.

4.2.2 Polylexical units

Polylexical units are present in almost every dictionary. Under this classification, we have included compounds and all kinds of lexical combinations, such as collocations or phrasemes. By compounds we mean every lexical unit formed by two or more elements with autonomy within the language that together form a new lexical unit with a new meaning. By definition, in a general-language dictionary we can only find compounds and more rarely fixed combinations in an entry.

The encoding of compounds can be seen in Example 4:

decreto-lei [dɨkrɛtulɐ̟j]. *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.

Original encoding

```
<entry id="decreto-lei">
  <form>
    <orth>decreto-lei</orth>
    <pron>dɨkrɛtulˈɐ̟j</pron>
  </form>
  <gramGrp>s. m.</gramGrp>
<!--etc. -->
</entry>
```

Conversion to TEI Lex-0

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

Example 4: DACL polylexical unit – original encoding and conversion from TEI to TEI Lex-0.

In the DACL, Latin phrases, i.e., fixed combinations, appear as headwords too (see Example 5):

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

Original encoding

```
<entry id="fiat lux">
  <form>
    <orth>fiat lux</orth>
  </form>
  <gramGrp>loc. lat. </gramGrp>
```

Conversion to TEI Lex-0

```
<entry type="polylexicalUnit" xml:lang="pt"
xml:id="fiat_lux">
  <form type="lemma">
    <orth>fiat lux</orth>
  </form>
```

```
<!--etc. -->
</entry>
```

```
<gramGrp>
  <gram type="pos">loc.</gram>
</gramGrp>
<etym type="borrowing"><lang>lat.</lang></etym>
<!--etc. -->
</entry>
```

Example 5: DACL polylexical unit – original encoding and conversion from TEI to TEI Lex-0.

In this example, even if "locução latina" [latin phrase] is not a part-of-speech, for now we decided to keep it encoded that way. While we are trying to use the Universal Dependencies Part-of-Speech Tagset[10], we needed to add our own tags for specific cases due to the lack of accurate tags for our purpose.

4.2.3 Affixes

In certain dictionaries, such as the DACL, affixes appear as headwords, as shown in Example 6[11]. The DACL uses bracketed hyphens as visual clues of the position the given affix takes in relation to the lexical unit it is attached to: the headword *(-)carpo(-)* indicates that *carpo* can be used as both a suffix and a prefix. Bracketed hyphens play the role of labels signalling the morphological property of the affix, but are not part of the affix itself. We therefore encode the affix itself as `<orth>carpo</orth>`, while using the element `<lbl>` to reflect the positional labels used in the dictionary.

(-)carpo(-) *elem. de form.* (Do gr. καρπός 'fruto'). Exprime a noção de *fruto*. *Mesocarpo, carpologia, pericarpo.*

Original encoding

```
<entry id="carpo">
  <form>
    <orth>(-)carpo(-)</orth>
  </form>
  <gramGrp>elem. de form.</gramGrp>
<!--etc. -->
</entry>
```

Conversion to TEI Lex-0

```
<entry type="affix" xml:lang="pt" xml:id="carpo">
  <form type="lemma">
    <lbl>(-)</lbl><orth>carpo</orth><lbl>(-)</lbl>
  </form>
  <gramGrp>
    <gram type="pos">elem. de form.</gram>
  </gramGrp>
<!--etc. -->
</entry>
```

Example 6: DACL affix headword – original encoding and conversion from TEI to TEI Lex-0.

[10] When labelling entries with part-of-speech appropriate linguistic terminology is crucial, mainly when we are talking about interoperability between lexical resources. This information must be one of the values from the Universal Dependencies Part-of-Speech Tagset: @norm attribute. See https://universaldependencies.org/u/pos/.

[11] Even if "*elemento de formação*" [affix] is not a part-of-speech, for now we decided to keep it encoded that way.

4.2.4 Abbreviations

Concerning abbreviations, the DACL registers different types of these: abbreviation (*Cf.*), alphabetism (*AAC*), acronym (*VIP*), symbol (*Ag*), contractions (*do* [of the]) and clipped forms (*metro* [metropolitan]).

Ag *símb.* (De *a<r>g<entum>* 'prata'). *Quím.* Símb. da *prata*.

Original encoding

```
<entry id="Ag">
  <form>
    <orth>Ag</orth>
  </form>
  <gramGrp>símb.</gramGrp>
<!--etc. -->
</entry>
```

Conversion to TEI Lex-0

```
<entry type="abbreviation" xml:lang="pt"
xml:id="Ag">
  <form type="lemma">
    <orth>Ag</orth>
  </form>
  <gramGrp>
    <gram type="pos">símb.</gram>
  </gramGrp>
<!--etc. -->
</entry>
```

Example 7: DACL abbreviation – original encoding and conversion from TEI to TEI Lex-0.

VIP [víp]. *s. m. e. f.* Sigla de *Very Important Person* (Pessoa Muito Importante).

Original encoding

```
<entry id="VIP">
  <form>
    <orth>VIP</orth>
    <pron>víp</pron>
  </form>
  <gramGrp>s. m. e. f.</gramGrp>
<!--etc. -->
</entry>
```

Conversion to TEI Lex-0

```
<entry type="abbreviation" xml:lang="pt"
xml:id="VIP">
  <form type="lemma">
    <orth>VIP</orth>
    <pron>víp</pron>
  </form>
  <gramGrp>
    <gram type="pos" norm="NOUN">s.</gram>
    <gram type="gen">m.</gram>
    <lbl>e</lbl>
<gram type="gen">f.</gram>
  </gramGrp>
<!--etc. -->
</entry>
```

Example 8: DACL abbreviation – original encoding and conversion from TEI to TEI Lex-0.

In these examples, we would like to call attention to the usage of the `pos` element to annotate this type of abbreviation. Again, these are not proper part-of-speech attributes and might change in the future.

Finally, clipped forms are usually treated as nouns, as shown in Example 8:

metro[2] [mɛ́tru]. *s. m.* (Red. de *metropolitano*). **1.** Sistema de transporte urbano efectuado por comboios de tracção eléctrica, em linhas parcial ou totalmente subterrâneas. ≃ METROPOLITANO. *Encontraram-se na estação de metro. O metro está em greve.* **boca**[+] **de metro. 2.** Comboio que assegura esse sistema de transporte. *Apanhar, perder o +.*

Original encoding

```
<entry id="metro:2">
  <form>
    <orth>metro:2</orth>
    <pron>mˈɛtru</pron>
  </form>
  <gramGrp>s. m.</gramGrp>
<!--etc. -->
</entry>
```

Conversion to TEI Lex-0

```
<entry type="abbreviation" xml:lang="pt"
xml:id="metro_2" n="2">
  <form type="lemma">
    <orth>metro</orth>
    <pron>mˈɛtru</pron>
  </form>
  <gramGrp>
    <gram type="pos" norm="NOUN">s.</gram>
    <gram type="gen">m.</gram>
  </gramGrp>
<!--etc. -->
</entry>
```

Example 9: DACL abbreviation – original encoding and conversion from TEI to TEI Lex-0.

This example shows yet another detail regarding visual information. There is more than one entry for the lexical unit *metro*. Therefore, as usual, the dictionary includes a superscript number, near the headword, to differentiate each entry. To encode this information we do it two ways: first, the entry identifier has the entry number following the headword, separated by a undescore. As this information is also important to the reader, it is encoded as the attribute `n` (number) in the entry element.

From these examples it is clear that TEI Lex-0 is going in a good direction, making the encoding more verbose but more structural, allowing machines to process this information better.

From these examples it is clear that TEI Lex-0 is going in a good direction, making the encoding more verbose but more structural, allowing machines to process this information better.

5. Automatic conversion of the original TEI schema to TEI Lex-0

Given that we are not dealing with a standard but with the process of creating it, the schema is not fixed. Therefore, our present goal is not to have the dictionary in TEI Lex-0 only, but to keep the original version in our own interpretation of TEI and have another version that can be used for tests and to promote the discussion with the TEI Lex-0 community.

Also, as our entries are stored independently in the XML database, our goal is not to produce a complete XML document for the dictionary, but a set of small XML files per dictionary entry. Therefore, details about the TEI header are deliberately being ignored at this stage, and thus we are not using the complete schema but only the entry portion, considering the entry tag as the document root element. In the future, the header can be stored in an independent record in the database, and a simple tool can be used to construct a TEI/TEI Lex-0 file with the complete dictionary, validating the complete schema.

The conversion between structured formats is not difficult as long as the information is somehow annotated in the source document. This is the case for most of the encoding changes needed in the dictionary, with a few exceptions.

If we were only dealing with structural changes, an interesting approach would be to use the eXtensible Stylesheet Language Transformations (XSLT) language. This would allow the transformation to run on top of eXist-DB, and could even be performed on demand for any desired entry. Nevertheless, to allow us more control when dealing with partially structured content, our approach was to use a generic high-level programming language (Perl).

In order to allow progressive validation, we chose to edit our schema in order to accommodate TEI Lex-0 recommendations, one at a time. For each of these changes, a new part of the script was added to perform the desired changes.

Two main changes needed human intervention: grammatical groups and etymology:

- While TEI allows the grammatical information (under the `gramGrp` element) to be unstructured (i.e., only the visual information, such as “n. m.” for masculine noun), TEI Lex-0 enforces the tagging of the part-of-speech information using specific tags. In order to guarantee the accuracy of this conversion, a list of the complete possibilities for the content of that tag was computed, and the desired annotation was manually added with part-of-speech. Taking the opportunity, we also normalized situations where the entry lexical unit had more than one grammatical analysis — e.g. *vegano* [vegan] whose morphological information is “adj., n. m.” [adjective and masculine noun].

Figure 3: *vegano* [vegan] (DACL new edition).

In these cases, the `gramGrp` element stores a list of possible `gram` entries, one for each analysis. This mapping was defined manually as a table, and the conversion script simply replaced the existing information with the new one.

- The other tricky conversion is the entry's etymology. It is challenging mainly because, when the PDF document was converted to TEI, not every detail of the etymology was properly annotated. While no information was lost, some portions were stored simply as plain content (text) without proper XML annotation. Unlike grammatical information, the creation of a list of all the possibilities is unthinkable, as the amount of entries that completely share their etymological information is close to zero. Thus, the process for etymology conversion had to be based on an approach that is similar to the one executed during the PDF to TEI conversion: a definition of a set of regular expressions to detect clear portions of the etymology (that do not include any ambiguity), which are annotated first, as anchors. Then, new rules and heuristics are applied using these anchors to detect other bits of information. This process is currently being done and it is expected that 95 % of the entries can be completely automated. The remaining ones might need direct manual intervention. This is a work in progress, and, just like most of the TEI Lex-0 encoding, further discussion on how to encode most of the information properly is still needed.

6. Conclusions and future work

In this paper, we focused on encoding information of different types of lexical units, providing examples, and thus contributing to a more consistent encoding of lexicographic data, constraining the variety of possibilities offered by the TEI Guidelines.

The results obtained are useful for the discussion and definition of the TEI Lex-0 standard. The definition of a standard is very important, as it allows resources or tools to be used interchangeably, but it is also a complex task, as the resulting standard should be able to encode different types of dictionaries, and not just for different languages, but with different purposes as well.

7. Acknowledgements

Research financed by 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/03213/2019, and by the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015 (ELEXIS).

8. References

Bański, P., Bowers, J. & Erjavec, T. (2017). TEI-Lex0 guidelines for the encoding of dictionary information on written and spoken forms. In I. Kosem et al. (eds.) *Electronic Lexicography in the 21st Century: Proceedings of eLex 2017 Conference*, pp. 485–94.

Bowers, J. & Romary, L. (2017). Deep encoding of etymological information. In *TEI. Journal of the Text Encoding Initiative*, TEI Consortium, 2017, ⟨https://jtei.revues.org/1643⟩. ⟨10.4000/jtei.1643⟩. ⟨hal-01296498v2⟩.

Gouws, R. H. (2018). Internet lexicography in the 21st century. In *Wortschatz: Theorie, Empirie, Dokumentation*, pp. 215–236.

ISO 16642:2017 *Computer applications in terminology – Terminological markup framework*

Khemakhem, M., Foppiano, L. & Romary, L. (2017). Automatic Extraction of TEI Structures in Digitized Lexical Resources using Conditional Random Fields. In Proceedings of *eLex 2017*, Leiden, Netherlands, September.

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. Available at: https://tei2018.dhii.asia/AbstractsBook_TEI_0907.pdf.

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

Tasovac, T. (2010). Reimagining the Dictionary, or Why Lexicography Needs Digital Humanities. In *Digital Humanities* 2010, pp. 254–256.

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

Websites:

TEI Consortium, eds. TEI P5: Guidelines for *Electronic Text Encoding and Interchange*. [Version 3.5.0]. [Last updated on 29th January 2019, revision 3c0c64ec4]. TEI Consortium. http://www.tei-c.org/Guidelines/P5/ ([13.07.2019]).

DARIAH WG: Lexical Resources and the H2020-funded European Lexicographic Infrastructure (ELEXIS), https://github.com/DARIAH-ERIC/lexicalresources/tree/master/Schemas/TEILex0.

Dictionaries:

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