Alberto Simões/Ana Salgado

SMART DICTIONARY EDITING WITH LeXmart

Abstract Given the relevance of interoperability, born-digital lexicographic resources as well as legacy retro-digitised dictionaries have been using structured formats to encode their data, following guidelines such as the Text Encoding Initiative or the newest TEI Lex-0. While this new standard is being defined in a stricter approach than the original TEI dictionary schema, its reuse of element names for several types of annotation as well as the highly detailed structure makes it difficult for lexicographers to efficiently edit resources and focus on the real content. In this paper, we present the approach designed within LeXmart to facilitate the editing of TEI Lex-0 encoded resources, guaranteeing consistency through all editing processes.

Keywords Dictionary encoding; Text Encoding Initiative; dictionary editing system

1. Introduction

In the last few years, many scholar projects have been encoding and placing dictionaries online, involving a wide variety of born-digital and retro-digitised lexicographic resources. Conceiving these types of lexicographic resources increasingly requires the application of adapted standards and formats capable of guaranteeing the availability of structured data and ensuring interoperability between systems, especially when the lexicographic production scenario is very heterogeneous due to its nature, form, and content. There are several types of dictionaries, in several languages, with disparate structures and different functions, purposes and users. Many of these dictionaries adopt a hierarchical data structure representation, mainly based on eXtensible Markup Language (XML).

The application of standard formats implies two different aspects: modelling and encoding. Modelling refers to the creation of an abstract model, that can account for all the lexical data and their components (Godfrey-Smith 2009). Encoding refers to the process of expressing the specific lexical data using a predefined data format. Essentially, modelling is a design task, and encoding is an implementation task. These are crucial issues for lexicography to ensure interoperability between the software components of heterogeneous lexicographic resources (Romary/Wegstein 2012).

Dictionaries are modelled and encoded in multiple diverse formats, being the information organised and stored in files of different nature. For dictionary encoding using XML, there are different structured data formats, such as the Dictionary module from the Text Encoding Initiative (TEI Consortium 2021), the XML Dictionary eXchange Format (Snegov/Soshinskiy 2019) or even the OntoLex-Lemon (McCrae et al. 2017).

Currently TEI is the dominant format for several lexicographic projects, such as BASNUM,[1] Nénufar,[2] ARTFL,[3] VICAV[4] or Berlin-Brandenburg Academy of Sciences for digitising and transcribing legacy dictionaries.[5] From the very beginning, the TEI Guidelines have a mod-

[1] https://anr.fr/Project-ANR-18-CE38-0003.

[2] http://nenufar.huma-num.fr/.

[3] https://artfl-project.uchicago.edu/.

[4] https://vicav.acdh.oeaw.ac.at/.

[5] https://gitlab.com/xlhrld/retro-dict.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

ule explicitly focused on the encoding of dictionaries. However, this module is criticised regarding its extreme flexibility, i. e., the existence of multiple possibilities to encode similar structures that affect the interoperability of the encoded formats. In some cases, TEI makes no binding requirements for the possible values since there are many possibilities across different projects. In each lexicographic project, it is likely that standardising an agreed set of values will be very helpful. In this sense, it is better to customise or change the schema by providing more restrictions. This explains why, for example, there is the need to restrict the scope of usage information (Salgado et al. 2019). Interestingly, this flexibility is also the characteristic that justifies its wide adoption.

To reduce this freedom and define a specific format for dictionaries, the TEI Lex-0 initiative (Bański/Bowers/Erjavec 2017; Romary/Tasovac 2018; Tasovac/Romary 2018), a stricter version of the TEI schema, has been promoted to reduce the encoding options. In the context of the ELEXIS project[6], TEI Lex-0 has been adopted as one of the baseline formats (McCrae et al. 2019). Other projects, such as our case-study, *Dicionário da Língua Portuguesa*[7] (ACL 2021), are also using TEI Lex-0 as an encoding format (Salgado et al. 2019).

Within the development of LeXmart[8] (Simões/Salgado/Costa 2021), this schema was also adopted. Nevertheless, while some of the verbosity of the schema helps in the automatic processing, sometimes the proposed encoding can be hard for lexicographers (examples of these structures will be discussed through this paper). To make it easier for the continuous editing of large lexical resources, LeXmart uses a layer of macros to hide this complex structure. This feature also allows the consistency of annotation, as we will discuss shortly.

While our proposal is specifically tailored for TEI Lex-0, the idea behind our approach can be replicated to other formats.

In the next section, we will provide a brief discussion of the TEI data format, as well as dictionary writing systems (DWS). We will focus on the management of the formats, and how the tools deal with them. In section 3, we will describe examples of the complex structures we identified in TEI Lex-0 and present how they were hidden using our macro approach. We will also explain how these structures are taken into consideration for the online rendering of the dictionary. In section 4, the technical details of this layer implementation will be detailed. We conclude with some insights on our approach in section 5.

2. Lexicographic Encoding and Editing

Regarding encoding, the approaches used for lexicographic content follow the same ideas that are used for encoding other types of resources: visual or semantic encoding. A dictionary can be seen as a textual artefact with its own specific publishing history and its own verbal expression and visual arrangement of the linguistic content contained within it or we can instead prioritise the linguistic content, ignoring how it is presented and the exact sequence of words used in, for example, the definitions of articles. There is also a third (poten-

[6] https://cordis.europa.eu/project/id/731015.

[7] Currently, it is being prepared under the Instituto de Lexicologia e Lexicografia da Língua Portuguesa's supervision in collaboration with researchers and invited collaborators. This project is supported by a small annual Community Support Fund Portuguese National Fund (Fundo de Apoio à Comunidade – FAC) through the Fundação para a Ciência e a Tecnologia (FCT).

[8] https://lexmart.eu/.

tially more verbose) approach, that is, to do both simultaneously and make sure both kinds of information are aligned.

These views are defined as follows: the typographical view aims to mirror the physical structure of a document using elements from the core module. It concerns the layout of individual pages and is mostly used on retro-digitisation projects, where the aspect used to present the information should be kept. These TEI elements can be used to encode the page layout, column and line breaks and highlighted words. Some elements can also be typed to provide more precision on how they are typographically presented in the original printed document. The second level of encoding deals with the semantic and logical function of text structures and is concerned with the conceptual or linguistic content of a dictionary as a whole as well as its individual entries.

While some formats are focused only on content encoding, other are more versatile, allowing these two different encoded approaches in the same document as referred above. As an example, the XDXF format is more focused on content, while TEI is versatile enough to cover both aspects in the same document.

To overcome these constraints, the TEI Lex-0 main goal is to take advantage of the TEI work but introduce stricter rules on how elements can be used. This new paradigm is quite important, as it enforces reusability and interoperability between different systems, and allow easier manipulation of the resources by computational tools. Nevertheless, it is more verbose and, while reusing some element names for different information types, it is more error-prone when the resources editing is made manually.

In this regard, it gets relevant to use a proper editor that is aware of the specific structure of a dictionary and allows the lexicographer to focus on its real work and not in the details of the annotation schema. This means that, while possible, it is better to avoid that a lexicographer edits each dictionary entry directly in a generic XML editor.

DWS are available for some time (Abel 2012), but mostly as in-house tools developed by publishing companies. Recently, free and open-source tools have been made available, such as Lexonomy[9] (Měchura 2017) and LeXmart (Simões/Salgado/Costa 2021).

Concerning DWS, the first solutions were based on forms. The use of forms allows the editing of new lexicographic articles in a clean environment. However, it is not clear how each entry maps to a specific XML element, hiding the true structure of the document. For example, in the case of Lexonomy and LeXmart, the editing is based on a specific XML schema. While for Lexonomy the user can configure their own schema, LeXmart focuses on guaranteeing interoperability and therefore enforces the use of TEI Lex-0.

Both editors suffer from the same problem: the verbosity of the used schema, as the user is presented with the complete XML structure of each entry. This paper proposes a macro layer to simplify the editing of complex XML structures. This idea was implemented on top of LeXmart.

3. TEI Lex-0 Common Patterns

TEI Lex-0 schema has a specific set of common patterns when applying its encoding rules into a specific dictionary. As a starting point, and as a motivational example, consider the

[9] https://www.lexonomy.eu/.

case of synonyms encoding. According to the TEI Lex-0 schema, these lexicographic components should be encoded as the following structure:

```
<xr type=”synonymy”>
   <ref type=”entry”>word</ref>
</xr>
```

While this structure can be a little different in specific situations (for instance, when the synonym is accompanied by a geographic label that identifies the place or region where a lexical unit is mainly used), it will be used and reused for every synonym. This solution encodes the synonym as a cross-reference inside the dictionary and reuses the `<xr>` and `<ref>` elements, making it clear the internal reference between entries and, at the same time, reusing elements that are useful in other contexts of the encoding process. Nevertheless, presenting this XML directly to the lexicographer can be confusing and overwhelming as, in some way, it distracts the lexicographers from their real work of linguistic analysis.

Bearing this in mind, we considered the possibility of using special elements, working as macros. By macro, we mean custom XML elements that are expanded to and from a more complex structure of elements. This allows the replacement of a common structure, like the one shown above for encoding synonyms, by a non-standard element that hides the original element tree. For instance, the use of an artificial element, named `<syn>`, especially used for this purpose:

```
<syn>word</syn>
```

This approach allows the user to understand the proper structure of the entry and edit all the lexicographic content faster and with less visual clutter. Thus, the lexicographer will just see this element in the DWS. The tool performs the necessary steps to guarantee that the document structure will automatically be converted to and from the original schema. LeXmart will also allow the lexicographer to choose between the original XML structure or, instead, edit the simplified version.

The next subsections discuss this approach, detecting such patterns, and propose custom and simplified custom elements, that are used only for presenting the information to lexicographers. We discuss each simplified structure, exemplifying with lexicographic articles from the new *Dicionário da Língua Portuguesa* (DLP) (ACL 2021). This lexicographic work is a retro-digitised dictionary (Simões/Almeida/Salgado 2016) whose starting point was the *Dicionário da Língua Portuguesa Contemporânea* (ACL 2001), last published in 2001.

3.1 Synonyms and antonyms

Just as mentioned above, synonyms and antonyms are encoded as cross-references to their own entries in the dictionary. While making sense, the structure can be overwhelming. We suggest the use of the `<syn>` and `<ant>` elements as a shortcut to include this kind of reference. The main issue arises when there is further information associated with these elements as, for instance, a specific geographic area where that synonym is used, or a specific domain of knowledge where that synonym is mainly used.

As an illustrative example, consider the sense of *casmurro*^{2} [pigheaded] entry with the following synonym:

```
<xr type=”synonymy”>
   <ref type=”entry”>
      <usg type=”socioCultural”>Inf.</usg>
      cabeçudo
   </ref>
</xr>
```

This example includes a usage label (*socioCultural* type),[10] which restricts the scope of the synonym in contexts of informality. When applying the <syn> macro, this part of the entry gets simplified to:

```
<syn>
   <usg type=”socioCultural”>Inf.</usg>
   cabeçudo
</syn>
```

Therefore, the usage information is not lost. This process is applied in both directions, thus generating the original markup when required.

3.2 Etymology Cross-references

When encoding etymology and data components, including foreign language words (for instance, the origin Latin word), TEI Lex-0 allows the lexicographer to include a full entry registering that word sense, directly in the etymological section. While powerful, this also requires a complex structure. Consider the following example from the *era* [era] entry:

```
<etym>
   Do latim
   <cit type=”etymon”>
      <form><orth xml:lang=”la”>aera</orth></form>
   </cit>
</etym>
```

[10] See 8.2. Types of usage: https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html#index.xml-body.1_div.8_div.2.

This structure emphasizes that the Latin word is a citation, from an etymological point of view, which form is, in Latin, *aera*. While it is possible that in some situations this structure can be useful to add further details, that will rarely be the case. In a dictionary for a Romance language such Portuguese, where most words are derived from Latin, the editing of this information gets repetitive and time-consuming. Thus, we propose the use of the `<etymon>` element, which works as a macro for etymological citations:

```
<etym>
   Do latim
   <etymon xml:lang=”la”>aera</etymon>
</etym>
```

This kind of annotation is not only smaller but can also be presented more cleanly in the online DWS.

3.3 Examples

In TEI Lex-0 the `<cit>` and `<quote>` elements are used together for different purposes, namely for the inclusions of bibliographic examples (illustrative quotations extracted from corpora obtained from known authors) and usage examples (examples made up by the lexicographer). The distinction of the two types is obtained by attributes added in the `<cit>` element. A common example in DLP (from the entry *casmurro*^{2} [pigheaded]) is codified as[11]:

```
<cit type=”example”>
   <quote type=”example”>
      Por ser tão casmurro, perdeu a oportunidade de
      fazer um bom negócio.
   </quote>
</cit>
```

For the bibliographic citations, they are encoded using these same two elements, but with different attributes, and with extra bibliographic information. Consider the following example from the *amor* [love] entry:

```
<cit type=”example”>
   <quote>
      A noite é para o amor e o amor para Guma é Lívia.
```

[11] The double use of the @type attribute with the value of “example” is a choice of the authors for codifying lexicographic examples and differentiate them from bibliographic examples. This use is not generalized, as this distinction can be done by checking the existence of the bibliographic information inside the citation. Nevertheless, this decision was taken to make formatting through Cascading Style Sheets easier.

```
      Ele não quer amor de aventuras, amor de acaso.
   </quote>
   <bibl>
      <author>J. Amado</author>
      <title>Mar</title>
   </bibl>
</cit>
```

This type of example is more complex given the introduction of bibliographic information. Thus, we decided not to simplify it. But the clear distinction between both is important and must be presented clearly for the lexicographer (and end-user). Bearing this in mind, a simpler version of the lexicographic example was developed. The example presented above for a usage example is presented in the LeXmart as:

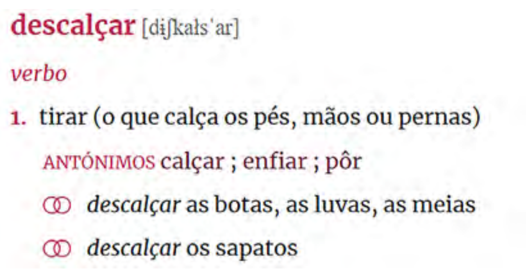
```
<example>
   Por ser tão casmurro, perdeu a oportunidade de
   fazer um bom negócio.
</example>
```

This distinction, together with the fact that the editor does not allow the inclusion of bibliographic information in an `<example>` tag, allows the lexicographer to quickly identify the type of quote.

3.4 Polylexical units

Many of the lexicographic articles from DLP include polylexical units, i. e., "a stable and recurrent sequence of lexemes that are perceived as an independent lexical unit by the speakers of a language" (Tasovac/Salgado/Costa 2020, p. 29), commonly called multiword expressions, collocations, lexical combinations, or even "*co-ocorrente privilegiado*"[12] [privileged co-occurrent]. Tasovac/Salgado/Costa (2020) argue that the encoding of polylexical units in dictionaries is a topic that has not been covered adequately and in sufficient depth by the TEI regarding the formal representation of polylexical units as they appear on the page of a single dictionary. The authors, for the case of privileged co-occurrent, recommend encoding this last type of polylexical units as `<form>` elements as they are presented to the end-user as a sequence of forms. However, and as this amendment has not yet been implemented, the DLP still maintains these sequences as a special kind of example as we explain further.

[12] 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") (ACL 2001, p. XXI).



descalçar [diʃkalsˈar]

verbo

1. tirar (o que calça os pés, mãos ou pernas)

ANTÓNIMOS calçar ; enfiar ; pôr

ↀ *descalçar* as botas, as luvas, as meias

ↀ *descalçar* os sapatos

Fig. 1: Snippet from the *descalçar* entry from DLP

Figure 1 shows the first sense of the *descalçar* [take of the shoes] lemma in the DLP. The two last lines, "*descalçar as botas, as luvas, as meias*" [to remove one's boots, one's gloves, one's socks] and "*descalçar os sapatos*" [to remove one's shoes], illustrate this type of polylexical units, that function as "*blocos semântica e sintaticamente afins*" [semantically and syntactically related blocks] (ACL 2001, p. XXI). In other words, the aim is to show that the lemma *descalçar* occurs frequently with the given nouns (boots, gloves, socks, shoes).

These polylexical units are encoded as a citation, just like the two types of examples presented before, but they need to be presented in a different way, both to the end-user and to the lexicographer. We applied a new macro for this structure. Consider the following fragment from Fig. 1 entry:

```
<cit type=”example”>
   <quote type=”collocation”>
      <hi>descalçar</hi> as botas, as luvas, as meias
   </quote>
</cit>
<cit type=”example”>
   <quote type=”collocation”>
      <hi>descalçar</hi> os sapatos
   </quote>
</cit>
```

Just like in previous situations, we decided to create a simple macro `<collocation>` to hide this structure and differentiate common bibliographic examples from this specific type:

```
<collocation>
   <hi>descalçar</hi> as botas, as luvas, as meias
</collocation>
<collocation>
   <hi>descalçar</hi> os sapatos
</collocation>
```

4. LeXmart implementation details

LeXmart is built on top of eXist-DB[13] as main backend. This document-oriented noSQL database is built with support for W3C technologies[14] like XQuery, XPath, XForms and XSLT. This allows the usage of XSL transformations to perform the conversion between the official TEI Lex-0 format and the simplified version described earlier.

The conversion process is composed of two XSLT files, one to simplify the notation, and another one to restore the correctness of the document. These stylesheets are prepared to do not perform any change when there is a structure that does not fit exactly on the pattern that was defined. This way, it is possible to ensure that there will be no loss of information during the conversion process.

To keep this pair of stylesheets working correctly, their composition needs to have the same behaviour as the mathematical identity function: return the original document.

Figure 2 shows the portion of the stylesheet that is responsible to convert a citation inside the etymology element to a `<etymon>` element.

```
  <xsl:template match=”tei:cit[@type=’etymon’]”>

    <xsl:choose>

      <xsl:when test=”./tei:form/tei:orth”>

        <etymon>

          <xsl:if test=”./tei:form/tei:orth/@xml:lang”>

            <xsl:attribute name=”xml:lang”>

              <xsl:value-of

                      select=”./tei:form/tei:orth/@xml:lang”/>

            </xsl:attribute>

          </xsl:if>

          <xsl:apply-templates

                      select=”./tei:form/tei:orth/node()”/>

        </etymon>

      </xsl:when>

      <xsl:otherwise><xsl:apply-templates/></xsl:otherwise>

    </xsl:choose>

</xsl:template>
```

Fig. 2: Extract from the XSLT template to perform the conversion of citations inside the etymology element

[13] https://exist-db.org/

[14] https://www.w3.org/

Meanwhile, the inverse process is obtained applying the stylesheet presented in Figure 3.

```
<xsl:template match=”tei:etymon”>
  <cit type=”etymon”>
    <form>
      <orth>
        <xsl:if test=”@xml:lang”>
          <xsl:attribute name=”xml:lang”>
            <xsl:value-of select=”@xml:lang”/>
          </xsl:attribute>
        </xsl:if>
        <xsl:apply-templates/>
      </orth>
    </form>
  </cit>
</xsl:template>
```

Fig. 3: Extract from XSLT template to reverse the conversion of the `<etymon>` element inside the etymology element

These transformations can be performed on server-side, applying the stylesheets when the entries are fetched or saved in the database, or performed directly on client side, using the lexicographer’s browser.

5. Final remarks

The use of XML to encode any type of content has always created discussion regarding its verbosity. The truth is that, adding elements to a digital content, and properly annotating it can duplicate the size of the document. That is a reason why XML as a serializing format is being less and less used (Fonseca/Simões 2007), in comparison with other formats such as JSON (JavaScript Object Notation) or YAML (Yet Another Markup Language).

Nevertheless, for digital humanities, the requirement of annotating texts at different levels, and not just as a structural schema, requires the use of *mixed content elements* (elements that can contain text and other elements as direct children), which are only possible with XML. But to be effective, the use of XML needs to be complemented with tools that allow content authors to quickly annotate their information.

In this paper, we proposed a solution based on macros to simplify some structures of TEI Lex-0 for editing purposes, making the user interface simpler to the lexicographer. The process is transparent to the lexicographers, in the sense that they do not need to be aware of the conversion process to properly use the tool.

The irreversible transition to the digital environment has imposed on lexicography (and the humanities and social sciences in general) the challenge of adopting new methods concerning the traditional ones. It is important to highlight that we are still in a transition phase, where lexicographers, who worked for many years on printed dictionaries, are making efforts to embrace the digital environment. The traditional lexicographer no longer exists; today, any e-lexicographer must be a digital humanist, being, for example, also an encoder. This is our real concern: to help lexicographers without encoding experience to be able to edit lexical resources with more confidence and to allow them to dedicate themselves to what is really important, the lexicographic work *per se*.

References

Abel, A. (2012): Dictionary writing systems and beyond. In: Granger, S./Paquot, M. (eds.): Electronic lexicography. Oxford, pp. 83–106.

ACL (2001): Dicionário da Língua Portuguesa Contemporânea. Malaca Casteleiro, J. (Coord.). Lisbon: Academia das Ciências de Lisboa and Editorial Verbo.

ACL (2021): Dicionário da Língua Portuguesa. Salgado, A. (Coord.) [New digital edition under revision]. Lisbon.

Bański, P./Bowers, J./Erjavec, T. (2017): TEI-Lex0 Guidelines for the Encoding of Dictionary Information on Written and Spoken Forms. In: eLex 2017: Lexicography from Scratch. Leiden, pp. 485–494.

Fonseca, R./Simões, A. (2007): Alternativas ao XML: YAML e JSON. In: XATA 2007- 5ª Conferência Nacional em XML, Aplicações e Tecnologias Associadas, pp. 33–46.

Godfrey-Smith, P. (2009): Models and fictions in science. In: Philosophical Studies, 143, pp. 101–116. http://doi.org/10.1007/s11098-008-9313-2.

McCrae, J. P. et al. (2017): The OntoLex-Lemon Model: development and applications. In: eLex 2017: Lexicography from Scratch. Leiden, pp. 587–597.

McCrae, J P. et al. (2019): The ELEXIS Interface for Interoperable Lexical Resources. In: eLex 2019 – Electronic Lexicography in the 21st Xentury, pp. 642–659.

Měchura, M. (2017): Introducing lexonomy: an open-source dictionary writing and publishing system. In: eLex 2017 – Lexicography from Scratch, pp. 662–679.

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

Romary, L./Wegstein, W. (2012): Consistent modeling of heterogeneous lexical structures. In: Journal of the Text Encoding Initiative 3. http://doi.org/https://doi.org/10.4000/jtei.540.

Salgado, A. et al. (2019): TEI Lex-0 in action: improving the encoding of the dictionary of the Academia das Ciências de Lisboa. In: eLex 2019 – Electronic Lexicography in the 21st Century, pp. 417–433.

Simões, A./Almeida, J. J./Salgado, A. (2016): Building a dictionary using XML technology. In: 5th Symposium on Languages, Applications and Technologies (SLATE'16), pp. 14:1–14:8. http://doi.org/10.4230/OASICS.SLATE.2016.14.

Simões, A./Salgado, A./Costa, R. (2021): LeXmart: a platform designed with lexicographical data in mind. In: eLex 2021 – Post Editing Lexicography, pp. 529–541.

Snegov, S./Soshinskiy, L. (2019): Why is XDXF better than other dictionary formats? GitHub. https://github.com/soshial/xdxf_makedict (last access: 22-03-2022).

Tasovac, T./Romary, L. (2018): TEI Lex-0: a baseline encoding for lexicographic data. Version 0.8.5. DARIAH Working Group on Lexical Resources. https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html# (last access: 22-03-2022).

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

TEI Consortium (2021): TEI P5: guidelines for electronic text encoding and interchange. TEI Consortium. http://www.tei-c.org/Guidelines/P5/ (last access: 22-03-2022).

Contact information

Alberto Simões
2Ai – School of Technology, IPCA, Barcelos, Portugal
asimoes@ipca.pt

Ana Salgado
NOVA CLUNL, Centro de Linguística da Universidade NOVA de Lisboa, Portugal/Academia das Ciências de Lisboa, Portugal
ana.salgado@fcsh.unl.pt

Acknowledgements

This paper was partially funded by Portuguese national funds (PIDDAC), through the FCT – Fundação para a Ciência e Tecnologia and FCT/MCTES under the scope of the projects UIDB/05549/2020 and UID/LIN/03213/2020, and by 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.