LeXmart: A platform designed with lexicographical data in mind

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

[1]2Ai – School of Technology, IPCA, Barcelos, Portugal
[2] Academia das Ciências de Lisboa, Lisboa, Portugal
[3] NOVA CLUNL, Centro de Linguística da Universidade NOVA de Lisboa, Lisboa, Portugal
E-mail: asimoes@ipca.pt, anacastrosalgado@gmail.com, rute.costa@fcsh.unl.pt

Abstract

LeXmart is an open-source web platform used to support the lexicographer's work through editing, control, validation, management, and publication of lexical resources. This tool was specifically developed to facilitate the compilation of general monolingual dictionaries in which data is encoded according to the Text Encoding Initiative (TEI) schema (chapter 9). Here, we will describe the challenges of adapting LeXmart to deal with TEI Lex-0 and distinct types of lexical resources, namely *Dicionário da Língua Portuguesa* (DLP) and *Vocabulário Ortográfico da Língua Portuguesa*, lexicographic works from Academia das Ciências Lisboa, and *Dicionário Aberto*, the retro-digitised version of the Cândido de Figueiredo dictionary. This article describes the steps taken to update the LeXmart platform to deal with the TEI Lex-0 schema and describe the challenges on properly encoding these three projects while allowing the lexicographical team to work continuously. This work builds on automatic operations performed on top of the original resources. It also includes the changes made to the editor to make it capable of dealing with the encoding updates and the new types of resources.

Keywords: dictionary editing system; e-lexicography; online dictionary; TEI Lex-0

1. Introduction

Compiling a dictionary is both a challenging and time-consuming task. For centuries, data collection and compilation of lexicographic data had been done on pen and paper, making the lexicographic work as a Herculean task. Nowadays, there are various computerised tools that can support the writing of dictionaries.

Since the beginning of computer-aided dictionary editing, publishers and some academic institutions have developed their own software to create dictionaries for commercial purposes. The first-generation dedicated dictionary writing systems were developed in the previous century in order to make life easier on the entry-writing front (Rundell & Kilgarriff, 2011). On the other hand, as secrecy is always the lifeblood of a business, these systems were not shared with third parties, which nowadays has a significant impact on issues concerning interoperability between different lexicographic resources.

The irreversible transition to a digital environment in the past two decades has imposed new challenges on lexicography in terms of adopting new methods, mainly due to technological advances, the fall of many publishers, and the changes introduced in their business models (Rundell, 2010). Nevertheless, independent software continues to be developed to assist lexicographers from different institutions. LeXmart[1] is one of these tools, designed from scratch to support an existing dictionary in an era where there is no great commercial interest in a dictionary distributed in physical mode, i.e. a printed version. Our main concern is to develop a lexicographic tool that responds to heterogeneous lexicographic structures and ensures that the structural components of a lexicographic article, known for their extreme complexity, are well identified and represented in a well-defined hierarchical organisation and appropriate metalanguage.

[1] Available at http://lexmart.eu/.

In the next section, we briefly discuss the LeXmart tool. Then, in Section 3, we describe the lexicographic resources that are currently under development using LeXmart. These resources are then analysed in Section 4 in terms of their structure and encoding details, using the TEI Lex-0 specifications. This is followed by Section 5, where we describe how LeXmart is being designed to help produce valid TEI Lex-0 documents keeping, at the same time, the interface as simple as possible for the lexicographer. Finally, in Section 6, we conclude the paper by presenting some insights into the project's status and propose several promising future research areas.

2. LeXmart

LeXmart is an open-source web platform used to support the lexicographer's work. It aims to support the activities involved in the whole lifecycle of preparing a dictionary, including editing the lexicographic articles, controlling, validating, and managing the dictionary and its content Simões et al. (2019).

LeXmart was developed using a bottom-up approach to solve a specific problem: storing and allowing the editing and quality management processes of the *Dicionário da Língua Portuguesa Contemporânea* (DLPC) (ACL, 2001). Further details on this project will be elaborated in Section 3.1.

This bottom-up approach means that, instead of creating a dictionary editing system from scratch (thereby restricting how dictionaries are defined), a basic version is first built and then further refined according to users' data management needs. In this way, the LeXmart tool was shaped as per the lexical data, rather than requiring the data to fit the tool.

Despite the evident benefits of the bottom-up approach, LeXmart was clearly created as a biased tool used to deal with a single lexical resource. DLPC was encoded following version 5 of the TEI Guidelines for Electronic Text Encoding and Interchange[2] while including some adaptations to match the TEI standard to preserve the original structure of the dictionary. Although LeXmart responds to the editing needs of the DLPC, the flexibility of the scheme was very restricted as it was designed specifically for only this dictionary. This limitation severely limited the advantages of using LeXmart in other lexicographic resources, which, in a way, are characterised by high structural heterogeneity. Meanwhile, the LeXmart platform has been heavily used to edit DLP and make it robust enough to deal with an actual-sized dictionary. Therefore, the team associated with it has an interest in using the tool to edit and maintain other lexical resources, namely:

- The *Dicionário Aberto* (DA) (Simões & Farinha, 2011), a transcription of a 1913 dictionary in the Portuguese language that was encoded using a custom TEI schema.
- *The Vocabulário Ortográfico da Língua Portuguesa* (VOLP-1940) (ACL, 1940), published by *Academia das Ciências Lisboa* (ACL) in 1940, which is currently being encoded using the TEI Lex-0.

These resources have different structures and have been encoded using different schemas. We cannot maintain LeXmart with a specific bias for each resource it includes. Therefore,

[2] Available at https://tei-c.org/guidelines/.

the current task in progress is rewriting LeXmart to focus on a specific and strict schema that can adequately encode all projects currently under development, following TEI Lex-0 specifications. Nevertheless, this target requires that the current LeXmart database's lexical resources be properly transformed and encoded into TEI Lex-0. Therefore, this article focuses not only on the tool and its changes, but also on the original dictionary's encoding process and the newly added resources.

3. Lexicographic Resources

This section presents the three resources for which LeXmart is used. For each one, we share some insights into their origins and the goals of including each of those resources into LeXmart.

3.1 Dicionário da Língua Portuguesa

The *Dicionário da Língua Portuguesa* (DLP) (ACL, 2021) is a scholarly dictionary of the Portuguese language being developed by the ACL. DLP aims at being the first digital academic Portuguese dictionary. The main objective of this endeavour is to update the DLPC 2001 edition by presenting an entirely new lexicographic resource. The database will be available online for free, and currently there are no plans to publish a printed version of the dictionary. It is a monolingual dictionary that is descriptive in nature, but with normative indications, as can be expected from a dictionary prepared by an academy of sciences. It is based on a retro-digitised dictionary created by converting the DLPC, described in the previous section, that was last published in the year 2001. This retro-digitisation process was previously described by Simões et al. (2016).

The result of this retro-digitisation was the creation of a database with over 68,000 entries, each of them stored independently in an XML file. These entries did not follow the guidelines in the chapter 9 of the TEI, on dictionaries. It was designed in accordance with the metainformation that was possible to extract from a PDF file, which was the only information source. This resulted in well-formed XML files, which included non-standard XML elements and attributes. Some examples of newly added elements are the `group` tag for enclosing a set of senses with the same morphological information and the `syn` and `ant` elements for encoding the lists of synonyms and antonyms. Similarly, custom attributes were also added. One of them is the `@fem` attribute, added to the `orth` tag, that registers the feminine suffixes for the lemmas.

To guarantee interoperability, the DLP is being transformed to ensure its compliance with the TEI Lex-0 format (Salgado et al., 2019b), a streamlined version of the TEI dictionary chapter. This decision is also behind the adaptation of LeXmart to follow this specific schema. Section 4.1 elucidates this conversion process.

3.2 Dicionário Aberto

The *Dicionario Aberto* (DA) (Simões & Farinha, 2011) is a Portuguese-language dictionary obtained by the OCR and a fully manual validation of the *Nôvo Diccionário da Língua Portuguêsa*, authored by Cândido de Figueiredo in 1913. This retro-digitisation process was done in close cooperation with the Distributed Proofreaders project of the Project Gutenberg[3]. The transcription took nearly four years to complete, and in 2010

[3] Available at https://www.pgdp.net/c/.

its full version was made publicly available on the Project Gutenberg website. The DA contains 128,521 entries: almost twice the number of entries in the DLP. This significant difference is explained by the DA registering orthographic variants of the same entry, as its original dictionary was published in troubled times for Portuguese language orthography.

This transcription was performed by volunteers with no lexicographic background. Thus, they were asked to encode the dictionary following quite a simple set of rules, which are used across all transcriptions performed in the Distributed Proofreaders website: each line in the original document should be presented independently (only hyphenated words were glued to the end of the top line), and bold and italics should be encoded using a custom markup, surrounding words by one asterisk character to encode bold words and one underscore character to encode italic words.

This simple markup was then converted to a custom TEI schema. The details on this encoding are in Section 4.2, where we discuss the process of transforming this original encoding into TEI Lex-0.

For years, DA has been subject to different transformations. The most relevant was the automatic orthography update, which allowed the dictionary to be used for experiments in natural language processing tasks, such as the automatic extraction of information to create Wordnets and ontologies (Gonçalo Oliveira, 2018; Gonçalo Oliveira & Gomes, 2014).

In the future, DA will be included in another broader project that aims to encode different dictionaries currently in the public domain into a single, more comprehensive resource.

3.3 Vocabulário Ortográfico da Língua Portuguesa

The *Vocabulário Ortográfico da Língua Portuguesa* [Orthographic Vocabulary of the Portuguese Language] (VOLP-1940) is the first orthographic vocabulary published by the ACL, in 1940. The *Digital Edition of the VOLP-1940*[4] (Salgado & Costa, 2020) aims at the digitisation of all the vocabularies of the ACL. The goal is to analyse the vocabularies with computational methods to better assess the importance of this work for the evolution of the Portuguese language in the 20th century and to contribute to the current movement of creating innovative, data-driven computational methods for text digitisation, encoding, and analysis. VOLP-1940's digitisation aims to create a lexicographical resource encoded in TEI, with structured information in the Simple Knowledge Organisation System (SKOS), to guarantee its future connection to other systems and resources, particularly in the Portuguese-speaking world.

The digitisation of the VOLP-1940 resulted in a series of image files of the original PDF manuscript that were converted to plain text using a commercial character recognition program (OCR) — the *Omnipage Pro*. The text was later exported to an editing program — *Microsoft Word* — to correct typos and inconsistencies generated by OCR.

Identifying the VOLP-1940 lexicographic conventions (for example, the comma used after each lemma or the use of abbreviations listed on the initial pages of the paperwork)

[4] Further details of the project at https://clunl.fcsh.unl.pt/en/investigacao/projetos-curso/edicao-digital-do-vocabulario-ortografico-da-lingua-portuguesa-volp-1940/ and at https://www.volp-acl.pt/index.php/vocabulario-1940/projeto.

was carried out to experiment a possible automated annotation of the entire work. Using *Microsoft Word* styles, we identified the different VOLP lexicographic article components, such as grammatical information, geographic information, etc.

4. TEI Lex-0 Encoding

LeXmart is being adapted to support the TEI Lex-0 standard properly. Although it would be interesting to have the tool dealing with different encoding formats, we are only targeting TEI Lex-0 as its community is currently growing, and it is being applied in projects such as BASnum[5] and Nénufar[6].

This format's groundwork started in 2016, and it is currently led by the Digital Research Infrastructure for the Arts and Humanities (DARIAH) Lexical Resources Working Group[7]. TEI Lex-0 aims to define a clear and versatile, albeit not too permissive, annotation structure to facilitate heterogeneously encoded lexical resources' interoperability. TEI Lex-0 should be regarded as "a format that existing TEI dictionaries can be unequivocally transformed to, so that they can be queried, visualised or mined uniformly" (Tasovac et al., 2018). As this format's layout has not been finished yet, we have been actively contributing to its development by raising GitHub[8] issues.

4.1 Dicionário da Língua Portuguesa

The *Dicionário da Língua Portuguesa* (DLP) is being developed, both lexicographically and computationally, without any direct funds. This results in a slower pace of work. As such, its conversion from the custom TEI schema to TEI Lex-0 is being done progressively, using small steps that fix some specific aspect of the original encoding. Simultaneously, as the lexicographic work is being performed concurrently, the LeXmart tool also needs adaptations to support the new elements.

The designed approach is cyclical, consisting of the following steps:

1. A specific detail of the original encoding is chosen for conversion.
2. Then, its conversion to TEI Lex-0 is discussed and evaluated.[9]
3. This is followed by the complete rewrite of the dictionary files, considering that specific encoding structure.
4. While this process runs[10], the LeXmart code is edited to support this specific TEI Lex-0 encoding.

As soon as this cycle ends, the complete dictionary is validated accordingly with the TEI Lex-0 and RelaxNG schema (REgular LAnguage for XML Next Generation), so that we can account for the progress and choose what the next conversion step is.

[5] Available at https://anr.fr/Projet-ANR-18-CE38-0003.
[6] Available at http://nenufar.huma-num.fr/?article=3813.
[7] See https://www.dariah.eu/activities/working-groups/lexical-resources/.
[8] Available at https://github.com/DARIAH-ERIC/lexicalresources/projects/1.
[9] In some specific situations, the TEI Lex-0 team is contacted in order to understand and/or discuss how some information should be encoded.
[10] It can take from a few minutes to more than half an hour.

Before putting this approach into practice, the original TEI Lex-0 schema was included in another RelaxNG schema that allows the dictionary to be stored in different XML files, without repeating the whole TEI Header[11], and allows the inclusion of an extra element, named meta, that includes some metadata about the entry state. To keep the XML files as compliant as possible, this extension was done properly, using XML namespaces.

To give an idea of the adaptation process, a list of steps that were taken during the conversion is shown below:

1. To each entry, the required `@xml:id` attribute was added, using the entry filename as the base, thus guaranteeing uniqueness. At the same time, the attribute `@xml:lang` was also added.
2. The `@type` attributes for the `usg` element were normalised using the standard values for geographic and domain instead of the suggested names from the TEI schema: '*geo*' and '*dom*' (Salgado et al., 2019a).
3. As noted before, one of the adaptations during the bootstrap process was the addition of the `group` tag. For all entries which contain only one `group` element, it was removed, keeping its contents intact.
4. According with the TEI Lex-0 schema, every sense element should include the `@xml:id` attribute. These attributes were also added automatically, taking as the base the entry identifier, and adding a suffix with the sense number.
5. The `cit` elements need a `@type` attribute. This was easy to add as, at this specific stage, any occurrence of this element was a bibliography example. Thus, the attribute `@type` was added to all `cit` elements with the same value: '*example*.'
6. To encode the page part of a citation (under the `bibl` element), the original schema used the `pag` element. TEI Lex-0 suggests the usage of the `citedRange` element.
7. In the etymology, references to words in the dictionary, and references to words in other languages, were both encoded with the `mentioned` element. To be able to perform the replacement correctly we needed to use some context. Thus, the sequence

   De `<mentioned>`*word*`</mentioned>`

   was replaced by

   De `<ref type="entry">`*word*`</ref>`.
8. As every reference needs a `@type` attribute, as seen in the previous item, every `ref` element present in the dictionary was edited to include this attribute, with the entry value.
9. In the original dictionary the `ph` element was used in expressions that required placeholders (specific multiword expressions, where a specific token is a word from a class, and not a concrete word). As this element is not supported by TEI Lex-0, but the `hi` (from highlight) is valid, these were replaced.
10. Synonyms and antonyms have initially been encoded with the `syn` and `ant` elements. These were changed to a more complex structure of a reference with a specific type (synonymy or antonymy), as shown in the example below.

[11] We are dividing the dictionary into individual files, for easy concurrent editing. Nevertheless, while specified individually, the whole set of files constitutes the real document. Therefore, a TEI Header will be generated every time the full dictionary is exported in a single XML document. While in the database, that information would be redundant.

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

11. Non-bibliographic examples were originally encoded as quotes, directly inside the `sense` element. This is not supported by the TEI Lex-0, requiring every occurrence to be replaced by the more complex structure shown below.

    ```
    <cit type="example"><quote type="example">...</quote></cit>
    ```

    Note that the `@type` attribute in the `quote` element is not required but useful for us to distinguish between bibliographic citations.
12. While DLP is being developed with the Internet as the target media, the project keeps track of entries or senses that should not be included in a paper dictionary. For this, the attribute `@digital` was originally created. To keep it with TEI Lex-0, the `@rend` attribute was chosen to encode this information. Thus, digital-only entries include the attribute `@rend="digital"`.
13. The references to words in other languages present in the etymology were encoded as mentioned elements. These were changed to citations, as shown in the next example:

    ```
    <cit type="etymon">
        <form><orth xml:lang="la">word</orth></form>
    </cit>
    ```

Even though we already converted much of the original syntax, the mentioned changes achieved 33,093 of the 70,726 entries in the dictionary as valid with regard to TEI Lex-0 (about 46.79%). There are some details needing changes that have not yet beenadequately discussed. One example is the `@fem` attribute in the `orth` element, which currently holds the suffix to generate the feminine form. One of the possibilities to encode this in TEI Lex-0 is to replace it with a full form entry. Nevertheless, for that to be done automatically we will require a morphological analyser to derive the feminine forms automatically.

4.2 Dicionário Aberto

Although the DA is also available in XML, following the dictionary chapter of TEI's general guidelines, the annotation granularity is bigger than DLP. This simplicity is derived from the lack of detailed annotation in the original document after the volunteer transcription, which only marked bold and italic words. Thus, the conversion to TEI was based only on that information, the knowledge of the dictionary's microstructure and a set of abbreviation lists (Simões & Farinha, 2011). These hints allowed a quite interesting structure to present the dictionary online with some quality but lack detailed annotations. Thus, its conversion to TEI Lex-0 is also simpler, as only the top-level structure is required.

As can be seen in Figure 1, originally each entry was encoded with only one sense. Only words with more than one grammatical class have more than one sense element. Different definitions are currently encoded in a single `def` element, where new lines are used to distinguish between senses.

While this structure is quite poor, its conversion to TEI Lex-0 is straightforward: the sense elements are removed from their current places. As for definitions (`def` element), their content is split by a new line and, for each line, a pair of `sense`/`def` elements is added. What follows is the addition of the required attributes, the identifier (`@xml:id`)

```
<entry id="drogaria">
  <form><orth>Drogaria</orth></form>
  <sense>
    <gramGrp>f.</gramGrp>
    <def>
      Porção de drogas.
      Estabelecimento, em que se vendem drogas.
    </def>
  </sense>
</entry>
```

Figure 1: Example of an entry before the TEI Lex-0 conversion.

```
<entry xml:id="drogaria" xml:lang="pt">
  <form><orth type="lemma">Drogaria</orth></form>
  <gramGrp>f.</gramGrp>
  <sense xml:id="drogaria-1"><def>Porção de drogas.</sense>
  <sense xml:id="drogaria-2"><def>Estabelecimento, em que se vendem drogas.</def></sense>
</entry>
```

Figure 2: Entry from Figure 1 after the TEI Lex-0 conversion.

and the language (`@xml:lang`). After these changes, we obtain a simple but valid TEI Lex-0 document.

While there are entries with some more annotation than in the presented example, in their transformation into a TEI Lex-0 file it is possible to keep the same basic structure.

4.3 Vocabulário Ortográfico da Língua Portuguesa

In microstructural terms, a lexicographical article from the VOLP-1940 may, as a rule, include the following elements: lemma, orthoepy, part of speech, and a gloss.

A lexicographical article in the VOLP-1940 starts with a base structure corresponding to the entry, followed by the grammatical information. Figure 3 shows the basic and regular structure of a VOLP-1940 entry to which the TEI Lex-0 annotation was applied.

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

Figure 3: Basic and regular structure of a VOLP-1940 entry.

While the entry element encompasses all the information contained in the lexicographical article, the form element is used to note the information relating to the base, detailing its `@type` attribute as "lemma," and the orthographic form is provided in the `orth` element. It is important to note that in TEI Lex-0, the `entry` element requires the attributes

`@xml:id`, the entry identifier and `@xml:lang`, the appropriate language code. Since we are dealing with vocabulary entries, we use the form `@type="lemma"`.

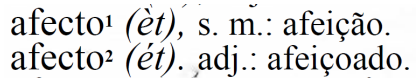

Figure 4: Example of homonymous words on VOLP-1940.

In the particular case of homonymous words, as shown in Figure 4, "afecto", the lemma is split. In TEI Lex-0, avoiding possible structural ambiguities, the `superEntry` element originally available in TEI (which groups a sequence of entries, such as a set of homographs) is no longer allowed, and therefore we use entry element systematically. To mark the numeric index, the element `lbl` preserves the digit of the original document while the attribute `@n` of the entry will, in turn, provide the information for the further processing of the entry by computational tools.

There is also information about words that are almost exclusively used in phrases. For example, when a particular word is only used in a particular phrase, this indication appears as an entry in what is considered the core word of that phrase — for instance, "cavalitas, el. nom. f. pl. na loc. adv. mod. às cavalitas" [riding piggyback, plural feminine noun element].

Another indication of a prescriptive nature concerns constructions that begin with the expression "Melhor que" [Better than]. The forms indicated as preferable are those that are considered to be closest to their origin or more correct for specific reasons, such as "canon" and "cânone" — "cânone, s. m. Melhor que canon" [cânone [canon], s. m. better than canon (Portuguese orthographic variant of the first form)][12]. So far, we have identified the essential and most relevant elements of the VOLP-1940's microstructure.

5. Simplifying TEI Lex-0 Interface

TEI Lex-0 is an interesting format, as it is much less permissive than the original guidelines in the chapter 9 of the TEI, on dictionaries. To make this process more straightforward and structured, the TEI Lex-0 team is reusing some elements for different, although near, semantics. As an example, TEI allows the use of the quote element by itself, to add an authorless quote, while quotes with bibliographic information are stored inside the `cit` element. TEI Lex-0 does not allow the direct usage of the quote element and suggests the use of a `cit` element in both situations. While this makes the automatic processing of the resource easier, as element trees are shared, it creates a large overhead of XML annotations. There are other examples of such situations, namely the inclusion of synonyms or antonyms, which have already been mentioned, that require a complex reference structure, or the encoding of foreign words, that could be encoded with the mentioned element in the original TEI schema, and that requires a more complex nested entry when properly encoded using TEI Lex-0.

As an option during its development, the LeXmart editor shows entries in a format very close to its XML structure. That is interesting for experienced users, as it clearly shows

[12] However, today the non-preferential form is the most common.

the annotation details. Nevertheless, if this editor includes the full structures for some of the situations described above, entries would be challenging to edit on a web browser.

During the development we also faced some issues regarding the versatility of Xonomy[13], the JavaScript library that implements the LeXmart web editor. While Xonomy has a very interesting application programming interface (API), and allows a high level of customisation, we faced some issues during the implementation of some functionalities, as they would require a large amount of coding.

The solution for both of these problems is the XML rewrite before the editing process, removing some complex structure and hiding it under a set of custom elements, and a post-processing pipeline that transforms this custom XML back into TEI Lex-0. This process is an excellent approach to make entry editing simpler and a straightforward way to guarantee the correct usage and respective element structure for some specific constructions.

This mapping is done automatically by the eXist Database backend that supports LeXmart, running a pair of eXtensible Stylesheet Language Transformations (XSLT) that transform the document structure.

Figure 5: LeXmart editor, showing two types of examples: bibliographic or not.

In Figure 5 we show two senses for the entry "drogaria" [drugstore] from DLP. Note that the first block, that corresponds to the second sense, shows a citation, of type example, that includes the quote and its bibliography information. The second block, which corresponds to the third sense, shows an `example` element. Although this element is not part of the TEI Lex-0 standard, it gets converted back and forth from the following structure:

```
<cit type="example"><quote type="example"> ⟹ <example>
```

[13] Available at https://github.com/michmech/xonomy.

Figure 6: LeXmart editor, showing etymological information with formant information.

Figure 6 shows a different situation for this same entry. To keep the editor as clean as possible, a `formant` element was created to hide the structure behind the inclusion of a foreign word in the etymology:

```
<etym>Do francês <cit type="etymon"><form>
  <orth xml:lang="fr">droguerie</orth></form></cit></etym>
```

These simple changes allow quicker editing for the lexicographer without jeopardising the document structure's adequacy to the TEI Lex-0 schema. In order to reduce the ambiguity, these new elements have different designations from the entries available either in TEI Lex-0 or the original TEI schema[14].

6. Conclusions

This article briefly described three different lexicographic resources, with different origins, and belonging to projects with independent goals. Nevertheless, it was shown that these resources can be encoded using the TEI Lex-0 schema, and therefore, their editing can be performed in a tool supporting this specific structure.

With this in mind, LeXmart has been modified to comply with this schema, and therefore allow their editing. To keep the tool as simple to use as possible, a set of mechanisms were developed to hide some of the XML encoding's verbosity.

For the future of LeXmart, a diverse number of features are already planned:

- The codebase of the tool requires generalisation, as much of it was developed with DLP in mind. While the code itself is easy to apply to different resources, the configuration of the system is currently hardcoded.
- LeXmart aims at allowing the lexicographer to manage labels (domain labels, geographic labels, etc.): not just to add or remove labels, but also to account for their usage. We also intend to have a taxonomy or an ontology to structure the labels. This would allow a very detailed annotation of the entries and allow interesting search scenarios for the end-user.
- With DLP going online during 2021, the system is being tested for exporting the dictionary database to a non-XML, but still document-oriented database for fast querying. Using the eXist database is quite helpful during the editing process, as the tool is aware of the XML structure, but it is relatively inefficient for simple querying. This will also allow the creation of dictionary snapshots, keeping the lexicographers' work on a non-public version of the dictionary.

[14] The designations currently in use might be changed in the future, as they were not yet a matter of discussion with all the involved parties.

- LeXmart, by itself, needs further improvements. A lot of the code is still too specific for DACL. Nevertheless, given it is available as an open-source project, we expect to have, sooner or later, new users testing the system with other languages and other kinds of resources, thus allowing for the development of new features but also the possibility of the customisation.

7. 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 the European Union's Horizon2020 research and innovation programme under grant agreement No 731015 (ELEXIS — European Lexicographic Infrastructure).

8. References

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

ACL (2001). *Dicionário da Língua Portuguesa Contemporânea*. Lisboa: Academia das Ciências de Lisboa & Editorial Verbo.

ACL (2021). *Dicionário da Língua Portuguesa*. Lisboa: Academia das Ciências de Lisboa.

Gonçalo Oliveira, H. & Gomes, P. (2014). ECO and Onto.PT: A flexible approach for creating a Portuguese wordnet automatically. *Language Resources and Evaluation*, 48(2), pp. 373–393.

Gonçalo Oliveira, H. (2018). A Survey on Portuguese Lexical Knowledge Bases: Contents, Comparison and Combination. *Information*, 9(2).

Rundell, M. (2010). What future for the learner's dictionary? In I.J. Kernerman & P. Bogaards (eds.) *English Learners' Dictionaries at the DSNA 2009*. Jerusalem: Kdictionaries, pp. 169–175.

Rundell, M. & Kilgarriff, A. (2011). Automating the creation of dictionaries: Where will it all end? *Studies in Corpus Linguistics*, 45, pp. 257–282.

Salgado, A. & Costa, R. (2020). O projeto 'Edição Digital dos Vocabulários da Academia das Ciências': o VOLP-1940. *Revista Da Associação Portuguesa De Linguística*, 7, pp. 275–294.

Salgado, A., Costa, R. & Tasovac, T. (2019a). Improving the consistency of usage labelling in dictionaries with TEI Lex-0. In *Lexicography ASIALEX 6*. pp. 133–156.

Salgado, A., Costa, R., Tasovac, T. & Simões, A. (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. Proceedings of the eLex 2019 conference*. pp. 417–433.

Simões, A., Almeida, J.J. & Salgado, A. (2016). Building a Dictionary using XML Technology. In M. Mernik, J.P. Leal & H.G. Oliveira (eds.) *5th Symposium on Languages, Applications and Technologies (SLATE)*, volume 51 of *OASIcs*. Germany: Schloss Dagstuhl, pp. 14:1–14:8.

Simões, A. & Farinha, R. (2011). Dicionário Aberto: um recurso para processamento de linguagem natural. *Viceversa: revista galega de traducción*, 16, pp. 159–171.

Simões, A., Salgado, A., Costa, R. & Almeida, J.J. (2019). LeXmart: A Smart Tool for Lexicographers. 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. Proceedings of the eLex 2019 conference*. pp. 453–466.

Tasovac, T., Romary, L., Banski, P., Bowers, J., de Does, J., Depuydt, K., Erjavec, T., Geyken, A., Herold, A., Hildenbrandt, V., Khemakhem, M., Petrović, S., Salgado, A. & Witt, A. (2018). TEI Lex-0: A baseline encoding for lexicographic data. Version 0.8.6. Technical report, DARIAH Working Group on Lexical Resources. URL https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html.