TEI Lex-0: a good fit for the encoding of the Portuguese Academy Dictionary?

[1] Ana Salgado, [1] Rute Costa, [2] Toma Tasovac

[1] NOVA CLUNL, Faculdade de Ciências Sociais e Humanas, Universidade NOVA de Lisboa, Portugal

[2] Belgrade Center for Digital Humanities, Serbia

anasalgado@campus.fcsh.unl.pt

rute.costa@fcsh.unl.pt

ttasovac@humanistika.org

NOVAFCSH
FACULDADE DE CIÊNCIAS SOCIAIS E HUMANAS
UNIVERSIDADE NOVA DE LISBOA

CLUNL
CENTRO DE LINGUÍSTICA DA UNIVERSIDADE NOVA DE LISBOA

BELGRADE CENTER FOR DIGITAL HUMANITIES

Acknowledgements

This research has been 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

- European Union's Horizon 2020 research and innovation programme under grant agreement No 731015 (ELEXIS)

Outline

1. Motivation and goals
2. Portuguese Academy Dictionary (DLPC)
3. TEI Guidelines for Dictionary encoding
4. TEI Lex-0
5. TEI Lex-0 encoding of the DLPC
6. Conclusions and future work

1. Motivation

A new PORTUGUESE ACADEMY DICTIONARY

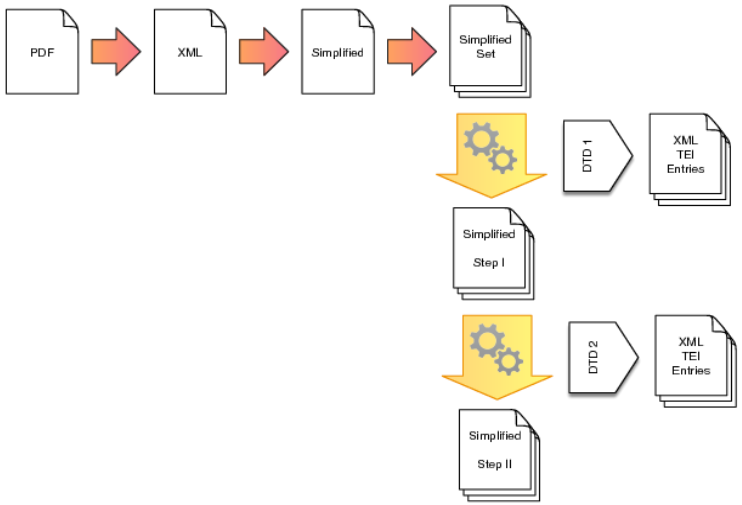
- Preservation
- Updating (of the content)
- Acessibility
- Interoperatibility
- Reusability
- Linguistic analysis

1. … and Goals

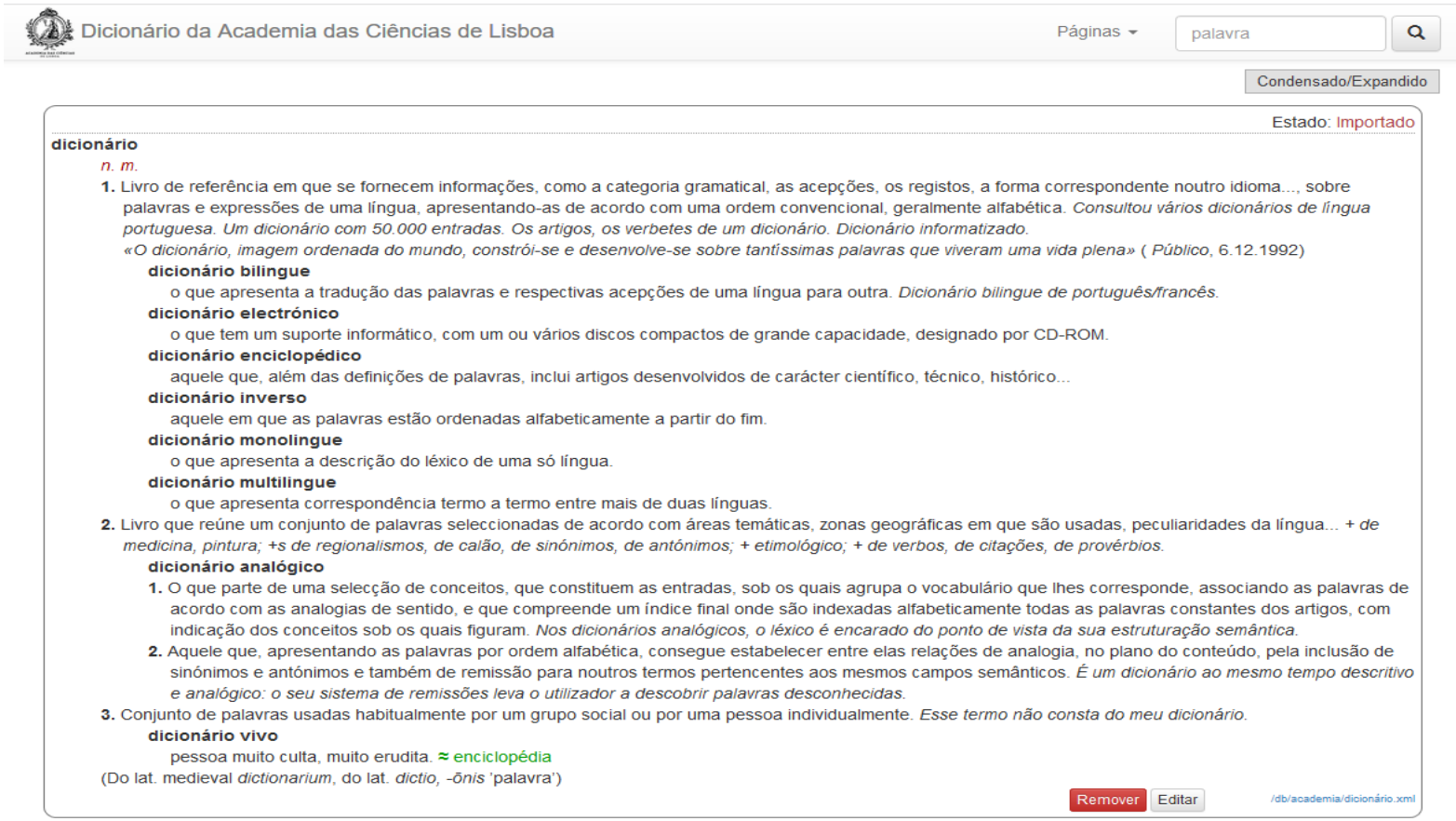
- to refer the conversion of TEI P5 schema into TEI Lex-0
- to demonstrate the application of TEI Lex-0 to mark up ‘special entries’ in DLPC
- to present arguments in favour of using TEI Lex-0 dictionary encoding
- to contribute to the efforts of the TEI Lex-0 group (DARIAH-ERIC Lexical Resources group)

2. Portuguese Academy Dictionary

DLPC = *Dicionário da Língua Portuguesa Contemporânea* (2001)
69 426 entries, 167 556 senses



Simões *et al.* (2016). Building a Dictionary using XML Technology



Salgado, Costa, Tasovac – TEI 2019

3. TEI Guidelines for Dictionary encoding

TEI

dictionary encoding

- TEI is a *de facto* standard in digital edition or text annotation projects, from novels, letters to music notation.
- It is frequently used in Digital Humanities as the basis for a large number of current lexicographic projects .
- TEI has a specific module for encoding dictionaries.

BUT...

- We could not find solutions in the Guidelines that covered all the microstructural elements of the dictionary.
- TEI Guidelines contain multiple encoding possibilities for the same dictionary features.

4. TEI Lex-0

TEI Lex-0

Romary & Tasovac (2018). TEI Lex-0: A Target Format for TEI-Encoded Dictionaries and Lexical Resources

- a streamlined version of the TEI Guidelines, simplified and enhanced for regular use to improve interoperability
- given its (still) non-standard nature, it can be changed in order to accommodate relevant dictionary structures
- we have been participating in the TEI Lex-0 discussion during the Portuguese Academy Dictionary encoding: https://github.com/DARIAH-ERIC/lexicalresources/tree/master/Schemas/TEILex0
- the advantage of applying TEI Lex-0 lies in the fact that lexicographers and terminologists are currently making efforts to align the ongoing revision of ISO LMF with TEI (Romary, 2015)

TEI Lex-0 will not replace the Dictionaries Chapter in the TEI Guidelines. It is being discussed as **a target format** that will facilitate the integration and analysis of the existing heterogeneously encoded lexical resources.

5. TEI Lex-0 encoding of the DLPC

XML ESSENTIAL CHANGINGS (some examples)	
Original encoding	Conversion into TEI Lex-0
<entry>	<entry xml:id="|*id*|" xml:lang="pt">
<term>	<form type="lemma">
<gramGrp>|*part of speech* and *gender*|</gramGrp>	<gramGrp><gram type="pos"</gram> <Gram type="gen">f.</gram></gramGrp>
<sense>	<sense xml:id="|*id*|">
<quote type="example">	<cit type="example"><quote>
<syn>|*synonym*|</syn>	<xr type="synonym"><ref type="entry|sense">|*synonym*|</ref></xr>
<cit><quote>|*example*|</quote> <bibl>|author|,<title>|title|</title>,|page|</bibl></cit>	<cit type="example"><quote>|*example*|</quote> <bibl><author>|*author*|</author> <title>|*title*|</title><citedRange>|*page*|</citedRange></bibl></cit>
<sense><def>V.<xr><ref>|*cross-reference*|</ref></xr></def></sense>	<xr><lbl>V.</lbl><ref>|*cross-reference*|</ref></xr>

5. TEI Lex-0 encoding of the DLPC

Basic entry structure

bibliofilia [bibljɔfilíɐ]. *s. f.* (De *biblio-* + gr. φιλία 'amor'). Amor aos livros.

ENTRY (headword)

PHONETIC TRANSCRIPTION

PART-OF-SPEECH

ETIMOLOGY

DEFINITION (meaning)

bibliofilia [bibliophilia], DLPC (2001)

```
<entry type="monolexicalWord" xml:lang="pt" xml:id="bibliofilia">
    <form type="lemma">
      <orth>bibliofilia</orth>
      <pron>bibljɔfilˈiɐ</pron>
    </form>
    <gramGrp>
      <gram type="pos" norm="NOUN">s.</gram>
      <gram type="gen">f.</gram>
    </gramGrp>
    <etym type="inheritance">
      <lbl>De</lbl>
      <cit type="etymon">
      <form>
        <orth>biblio-</orth>
      </form>
<etym type="grammaticalization">
          <seg type="desc">De</seg>
          <cit type="etymon">
            <form>
              <orth extent="pref">biblio-</orth>
            </form>
          </cit> <lbl>+</lbl>
<etym type="inheritance">
          <seg type="desc">Do</seg>
          <cit type="etymon" xml:lang="grc">
            <form><orth>φιλία</orth></form>
          </cit>
        </etym>
      <sense>
      <def>Amor aos livros</def><pc>.</pc>
      </sense>
</entry>
```

`entry`

- @xml:id
- @xml:lang – code (BCP 47)

`gram type element`

- part-of-speech of the entry and further specifications
- @norm attribute – values from the Universal Dependencies Part-of-Speech: https://universaldependencies.org/u/pos/

5. TEI Lex-0 encoding of the DLPC

Mark up of 'special entries'

- part-of-speech homonyms: *antepassado*[1], *antepassado*[2]
- entries that have a different meaning in the plural: *antepassados*
- etymological homonyms: *cota*[1], *cota*[2] (Bowers & Romary (2017). Deep encoding of etymological information in TEI)
- homographs: *lobo*[1] /ó/, *lobo*[2] /ô/
- spelling variants: *ouro, oiro*
- lexical variants: *missanga, miçanga*
- trademarks: *donut*®, *walkman*®

antepassado[1], a [ɐ̃tɨpɐsádu, -ɐ]. *adj.* (De *ante-* + *passado*). Que pertence ou viveu numa época anterior. ≃ ANTECESSOR, PREDECESSOR. ≠ DESCENDENTE, SUCESSOR.

antepassado[2] [ɐ̃tɨpɐsádu]. *s. m.* (De *ante-* + *passado*). **1.** Pessoa que é ascendente de outra ou outras. ≃ ASCENDENTE. ≠ DESCENDENTE. *Certos povos crêem-se descendentes de um antepassado comum. «o vaqueiro, pai do vaqueiro, o avô e outros antepassados mais antigos haviam-se acostumado a percorrer veredas, afastando o mato com as mãos.»* (G. RAMOS, *Vidas Secas*, p. 36). **2.** *pl.* Pessoas anteriormente ao momento actual. ≃ ANTECESSORES. ≠ VINDOUROS. *Herdámos estes costumes dos nossos antepassados. Culto dos antepassados.*

antepassado [ancestor, past], DLPC (2001)

```
<entry type="derivativeWord" xml:lang="pt"
xml:id="antepassado_1" n="1">
      <form type="lemma">
        <orth>antepassado</orth>
</form>
      <form type="inflected">
        <orth>antepassado</orth>
        <pron>ɐ̃tɨpɐsˈadu</pron>
      <gramGrp>
        <gram type="gen">m.</gram>
      </gramGrp>
      </form>
      <form type="inflected">
        <orth>antepassada</orth>
        <pron>ɐ̃tɨpɐsˈadɐ</pron>
      <gramGrp>
        <gram type="gen">f.</gram>
      </gramGrp>
       </form>
      <gramGrp>
      <gram type="pos" norm="ADJ">adj.</gram>
      </gramGrp>
      <!--etc. -->
</entry>
```

```
<entry type="derivativeWord" xml:lang="pt"
xml:id="antepassado_2" n="2">
      <form type="lemma">
        <orth>antepassado</orth>
        <pron>ɐ̃tɨpɐsˈadu</pron>
      </form>
      <gramGrp>
        <gram type="pos" norm="NOUN">s.</gram>
        <gram type="gen">m.</gram>
      </gramGrp>
      <!--etc. -->
      <sense xml:id="antepassado_2.2" n="2">
      <gramGrp>
        <gram type="number">pl.</gram>
      </gramGrp>
      <!--etc. -->
</entry>
```

cota[1] [kɔ́tɐ]. *s. f.* (Do fr. ant. *cotte*). **1.** *Mil.* Vestimenta ou peça da armadura usada antigamente por cavaleiros e guerreiros, sobre o busto, para os proteger dos golpes do adversário. **cota de armas** ou **cota.** 1. Vestimenta usada pelos cavaleiros sobre o arnês ou a armadura, tanto na guerra como em torneios. 2. Peça de vestuário, onde estava bordado o escudo real, usada pelos reis em actos públicos. **cota de malha(s),** armadura defensiva, com o formato de uma camisa feita de anéis ou malhas de metal entrelaçadas, usada antigamente para proteger o corpo contra golpes de armas. **fidalgo**+ **de cota de armas. 2.** *Mil.* Antiga peça de vestuário, usada para protecção por cavaleiros e guerreiros, que podia chegar ao joelho e apertar-se na cintura. **3.** Peça de vestuário parecida com uma bata ou gibão, usada antigamente. **4.** Antiga peça de vestuário feminino que consistia numa espécie de corpete. **5.** *Rel.* Veste litúrgica, de cor branca, com ou sem rendas, semelhante a uma sobrepeliz mas com mangas mais curtas, envergada sobre o roquete, e usada sobretudo em Itália, por certos dignitários religiosos e acólitos.

cota[2] [kɔ́tɐ]. *s. f.* (De origem obscura). Lado de uma ferramenta ou de um instrumento cortante, oposto ao que está afiado, ao gume. ≃ COSTAS.

cota [armour, …, back], DLPC (2001)

```
<entry type="monolexicalWord" xml:lang="pt"
xml:id="cota_1" n="1">
    <form type="lemma">
      <orth>cota</orth>
      <pron>kˈɔtɐ</pron>
    </form>
    <gramGrp>
      <gram type="pos" norm="NOUN">s.</gram>
      <gram type="gen">f.</gram>
    </gramGrp>
    <etym type="inheritance">
      <seg type="desc">Do</seg>
      <cit type="etymon" xml:lang="fro">
      <form><orth>cotte</orth></form>
      </cit>
    </etym>
    <!--etc. -->
</entry>
```

```
<entry type= "monolexicalWord" xml:lang="pt"
xml:id=“cota_2" n="2">
    <form type="lemma">
      <orth>cota</orth>
      <pron>kˈɔtɐ</pron>
    </form>
    <gramGrp>
      <gram type="pos" norm="NOUN">s.</gram>
      <gram type="gen">f.</gram>
    </gramGrp>
    <etym type=“undefinied">De origem obscura.
    </etym>
   <!--etc. -->
</entry>
```

lobo[1] [lóbu]. *s. m.* (Do gr. λοβός). **1.** *Anat.* Parte arredondada e saliente, existente em diversos órgãos. *Os lobos do cérebro. Lobo frontal. Lobos do pulmão.* **lobo de orelha,** prolongamento arredondado e carnudo do pavilhão auricular. **2.** *Bot.* Parte arredondada entre duas reentrâncias

lobo[2]**, a** [lóbu, -ɐ]. *s.* (Do lat. *lupus*). **1.** *Zool.* Animal mamífero carnívoro, da família dos canídeos (*Canis lupus*, Lin.), que vive nas regiões montanhosas da Europa, da Ásia e da América, em lugares solitários e tranquilos e é aparentemente semelhante a um cão grande. *No Verão, os lobos caçam solitários, mas, no Inverno, esfomeados, reúnem-se geralmente em alcateias, tornando-se muito perigosos.* **lobo asnal,** o que é grande, por oposição ao lobo-cerval. **2.** Pessoa cruel, de maus tratos. **3.** *Bras. Zool.* Animal mamífero carnívoro que vive no estado selvagem em certas regiões da América do Sul. **4.** *s. m.* Jogo infantil em que umas crianças fazem de lobo, outras de cordeiro e uma delas faz de pastor que as guarda. **5.** *s. m.* Máquina constituída por um tambor guarnecido de grossos dentes de ferro ou de aço e que é utilizada para abrir a lã, nas fábricas de lanifícios. **6. cova**[+] **do lobo. 7. fome**[+] **de lobo. 8. lobo do mar,** marinheiro corajoso, habituado às tempestades, ao perigo do mar. **9. lobo com pele de cordeiro,** pessoa que se faz passar por alguém calmo, dócil, sem maldade, ocultando o carácter, as intenções... **10. tripa**[+] **de lobo. entre o lobo e o cão,** *loc. adv.,* ao escurecer; ao lusco-fusco. **cair na boca do lobo,** cair numa cilada; ficar numa situação perigosa ou comprometedora que se pretendia evitar. **comer como um lobo,** comer muito e avidamente. **encomendar as ovelhas ao lobo,** confiar uma coisa a quem é passível de ser tentado por ela. **estar na goela do lobo,** estar numa situação muito grave. **meter-se na boca do lobo,** expor-se inadvertidamente a perigos ou dificuldades; cair numa cilada. **parecer que viu lobo.** 1. Apresentar-se com o cabelo hirto ou todo despenteado. 2. Ficar muito amedrontado com qualquer coisa inesperada. **quem não quer ser lobo não lhe veste a pele,** quem não quer ter trabalhos ou problemas não se envolva neles. **são lobos da mesma alcateia,** são pessoas com os mesmos interesses; entendem-se umas com as outras. **ser um lobo com pele de cordeiro,** ser velhaco ou mau com capa de santo. **ter uma fome de lobo,** estar esfomeado.

lobo [lobe, wolf], DLPC (2001)

```
<entry type="simpleWord" xml:lang="pt"
xml:id="lobo_1" n="1">
      <form type="lemma">
        <orth>lobo</orth>
        <pron>lóbu</pron>
      </form>
      <gramGrp>
        <gram type="pos"
norm="NOUN">s.</gram>
        <gram type="gen">m.</gram>
      </gramGrp>
      <!--etc. -->
</entry>
```

```
<entry type="simpleWord" xml:lang="pt"
"xml:id="lobo_2" n="2">
      <form type="lemma">
        <orth>lobo</orth>
      </form>
      <form type="inflected">
        <orth>lobo</orth>
        <pron>lóbu</pron>
      <gramGrp>
        <gram type="gen">m.</gram>
      </gramGrp>
      </form>
      <form type="inflected">
        <orth>loba</orth>
        <pron>lóbɐ</pron>
      <gramGrp>
        <gram type="gen">f.</gram>
      </gramGrp>
       </form>
      <gramGrp>
      <gram type="pos" norm="NOUN">s.</gram>
      </gramGrp>
      <!--etc. -->
</entry>
```

ouro, oiro [óru], [ójru]. *s. m.* (Do lat. *aurum*). **1.** *Quím.* Elemento sólido (símb. Au, nº at. 79, p. at. 196,997), metal precioso de cor amarela, macio e maleável, inoxidável e não atacado pelos ácidos. *O ouro é muito usado em joalharia.* **de ouro.** 1. Diz-se de alguma coisa que tem a cor, o aspecto, o brilho desse metal precioso. *Banho de ouro.* 2. De grande valor. ≃ PRECIOSO. *O silêncio é de ouro.* **ensaiador⁺ de ouro e prata. ouro branco.** 1. Liga deste metal que contém 20 a 50% de níquel, muito usada na ourivesaria. 2. Liga deste metal, níquel e paládio. 3. *Bras.* Algodão, considerado como riqueza agrícola. **ouro de lei.** 1. Aquele cujos quilates são determinados por disposições e que varia consoante o país. 2. Material de excelente qualidade. **ouro fino,** o que possui 24 quilates. **2.** Qualquer objecto feito desse metal. **medalha⁺ de ouro. 3.** Coisa muito valiosa. **a preço⁺ de ouro. bodas⁺ de ouro. coração⁺ de ouro. ouro em pó.** 1. Coisa de qualidade superior. 2. Pessoa muito virtuosa, sincera, leal. **menino⁺ de ouro. ouro negro,** petróleo, considerado como riqueza económica. **ouro coronário,** *Hist.,* coroa deste metal que davam, na Roma Antiga, aos generais vencedores. **sonho⁺ de ouro. 4.** Símbolo de riqueza. ≃ FORTUNA, OPULÊNCIA. **a galinha⁺ dos ovos de ouro. mina⁺ de ouro. 5.** Cor amarela dourada. *Tinha cabelos claros, cor de ouro.* **6.** *Heráld.* Metal heráldico, amarelado, representado por uma série de pontos. **7.** *funç. adj.* **córdoba⁺ ouro. 8. livro⁺ de ouro. 9. número⁺ de ouro. 10. ouro potável,** líquido oleoso e alcoólico. *O silêncio é de ouro.* **cobrir** *alguém* **de ouro,** tratar muito bem alguém, dando-lhe tudo de bom. **nem tudo o que brilha/luz é ouro,** as coisas nem sempre são o que parecem ser. **ouro sobre azul.** 1. Coisa muito boa. 2. Reunião de duas coisas excelentes, que combinam na perfeição. **pagar a peso de ouro,** pagar muito bem. *O patrão pagava-lhe a peso de ouro.* **valer o seu peso em ouro,** ser muito valioso.

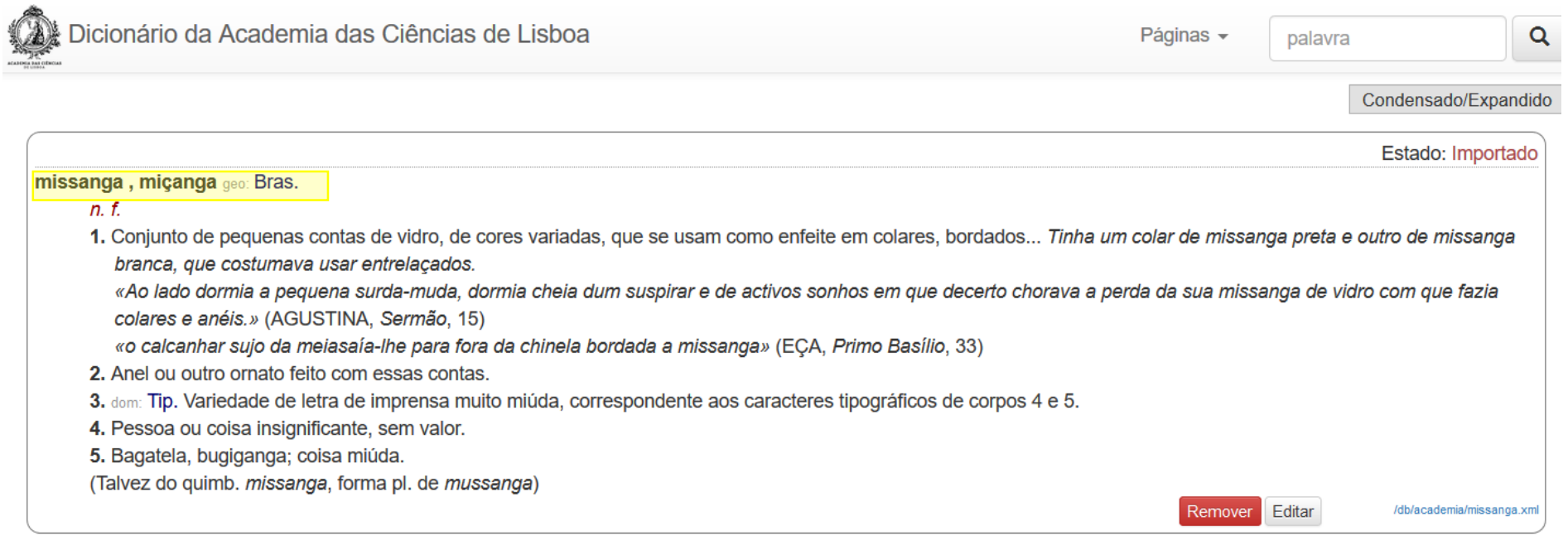
ouro [gold], DLPC (2001)

```
<entry type="simpleWord" xml:id="ouro" xml:lang="pt">
    <form type="lemma">
      <orth>ouro</orth>
      <pron>'oru</pron>
    </form>
    <form type="variant" xml:id="oiro" xml:lang="pt">
      <orth>oiro</orth>
      <pron>ˈojru</pron>
    </form>
  <!--etc. -->
</entry>
```

missanga [misɐ̃gɐ]. *s. f.* (Talvez do quimb. *missanga*, forma pl. de *mussanga*). **1.** Conjunto de pequenas contas de vidro, de cores variadas, que se usam como enfeite em colares, bordados... *Tinha um colar de missanga preta e outro de missanga branca, que costumava usar entrelaçados. «Ao lado dormia a pequena surda-muda, dormia cheia dum suspirar e de activos sonhos em que decerto chorava a perda da sua missanga de vidro com que fazia colares e anéis.»* (AGUSTINA, *Sermão*, p. 15). *«o calcanhar sujo da meia saía-lhe para fora da chinela bordada a missanga»* (EÇA, *Primo Basílio*, p. 33). **2.** Anel ou outro ornato feito com essas contas. **3.** *Tip.* Variedade de letra de imprensa muito miúda, correspondente aos caracteres tipográficos de corpos 4 e 5. **4.** Pessoa ou coisa insignificante, sem valor. **5.** Bagatela, bugiganga; coisa miúda.

missanga [glass bead], DLPC (2001)

```
<entry type="monolexicalWord" xml:id="missanga" xml:lang="pt">
    <form type="lemma">
      <orth>missanga</orth>
    </form>
    <!-- It isn´t on paper; new edition information -->
    <form type="variant" xml:id="id" xml:lang="pt">
      <orth>miçanga</orth>
      <usg type="geo">
      <placeName>Brasil</placeName>
      </usg>
    </form>
  <!--etc. -->
</entry>
```



donut [dɔ́nut]. *s. m.* (Ingl.). V. *dónute*.
dónute [dɔ́nutɨ]. *s. m.* (Do ingl. *donut*). Argola de massa frita enrolada em açúcar ou com diferentes coberturas, em forma de anel ou bola.

donut [donut], DLPC (2001)

walkman [wɔ́ɬkmɐn]. *s. m.* (Ingl., da marca registada). *Mús.* Leitor de cassetes portátil, por vezes com rádio acoplado, com auscultadores leves, para as pessoas ouvirem música quando se deslocam. *«O mesmo garoto tem um 'walkman', objecto que todos os mais novos sonham possuir»* (*DN*, 12.3.1990). Pl. walkmans.

walkman [walkman], DLPC (2001)

```
<entry type="simpleWord" xml:lang="en" xml:id="donut">
    <form type="lemma">
      <orth>donut</orth>
      <pron>dˈɔnut</pron>
    </form>
    <gramGrp>
      <gram type="pos" norm="NOUN">s.</gram>
      <gram type="gen">m.</gram>
    </gramGrp>
    <etym type="borrowing">
      <seg type="desc">Do</seg>
      <cit type="etymon" xml:lang="en">
      <form><orth>donut</orth></form>
     <!-- It isn´t on paper; new edition information -->
       <usg type="domain">marca registada</usg>
      </cit>
    </etym>
    <!--etc. -->
</entry>
```

6. Conclusions and future work

- The results are useful for the discussion and definition of the TEI Lex-0 standard.
- The need to adapt and rewrite the guidelines.
- An agreement between academies and other institutions would be desirable to systematize and optimize resources that can provide a better representation of the entire European lexicographical heritage.

Thank you for your attention.

Obrigada pela vossa atenção.

anasalgado@campus.fcsh.unl.pt
rute.costa@fcsh.unl.pt
ttasovac@humanistika.org

NOVAFCSH
FACULDADE DE CIÊNCIAS SOCIAIS E HUMANAS
UNIVERSIDADE NOVA DE LISBOA

CLUNL
CENTRO DE LINGUÍSTICA DA UNIVERSIDADE NOVA DE LISBOA

BELGRADE CENTER FOR DIGITAL HUMANITIES