Challenges of Word Sense Alignment: Portuguese Language Resources

Ana Salgado[1,2], Sina Ahmadi[3], Alberto Simões[4], John McCrae[3], Rute Costa[2]
[1]Academia das Ciências de Lisboa, Instituto de Lexicologia e Lexicografia da Língua Portuguesa, Lisbon, Portugal
[2]NOVA CLUNL, Universidade NOVA de Lisboa, Lisbon, Portugal
[3]Data Science Institute, National University of Ireland Galway
[4]2Ai – School of Technology, IPCA, Barcelos, Portugal
anasalgado@campus.fcsh.unl.pt; {sina.ahmadi,john.mccrae}@insight-centre.org; rute.costa@fcsh.unl.pt; asimoes@ipca.pt

Abstract

This paper reports on an ongoing task of monolingual word sense alignment in which a comparative study between the Portuguese Academy of Sciences Dictionary and the *Dicionário Aberto* is carried out in the context of the ELEXIS (European Lexicographic Infrastructure) project. Word sense alignment involves searching for matching senses within dictionary entries of different lexical resources and linking them, which poses significant challenges. The lexicographic criteria are not always entirely consistent within individual dictionaries and even less so across different projects where different options may have been assumed in terms of structure and especially wording techniques of lexicographic glosses. This hinders the task of matching senses. We aim to present our annotation workflow in Portuguese using the Semantic Web standards. The results obtained are useful for the discussion within the community.

Keywords: lexicography, sense alignment, linguistic linked data, Portuguese

1. Introduction

The concept of the dictionary has changed with the advent of the world wide web (WWW) and the digital age. The interoperability of linked data technologies has played an essential role in the evolution of lexicography (Shadbolt et al., 2006; Heath and Bizer, 2011; Gracia et al., 2017). It has been shown how lexicographic content can be represented and connected dynamically, thus allowing us to abandon once and for all the editorial perspective that still pervades most digital resources which continue to mirror the structure used in the paper versions.

The use of semantic standards enables the organization of vast amounts of lexical data in ontologies, Wordnets and other machine-readable lexical resources resorting to novel tools for the transformation and linking of multilingual datasets (McCrae and Declerck, 2019; Chiarcos et al., 2012). Linked Open Data (LLOD) promotes the use of the RDF data model to publish lexical data on the web for a global information system and interoperability issues.

There have been many efforts underway on behalf of numerous researchers to align different lexical resources (e.g. (Navigli, 2006; Knight and Luk, 1994) dealing with the word sense alignment (WSA) task. We define this task as linking a list of pairs of senses from two or more lexical resources using semantic relationships. To mention a few previous projects, Meyer and Gurevych (2011) align the Princeton WordNet with the English Wiktionary[1], and Henrich et al. (2012) link the GermaNet–the German Wordnet with the German Wikipedia[2].

WSA involves searching for matching senses within dictionary entries of different lexical resources and linking them, which poses significant challenges. The lexicographic criteria are not always entirely consistent within individual dictionaries and even less so across different projects where different options may have been assumed in terms of structure and especially wording techniques of lexicographic glosses. It has been demonstrated that the task of WSA is beneficial in many natural language processing (NLP) applications, particularly word sense disambiguation (Navigli and Ponzetto, 2012) and information extraction (Moro et al., 2013).

In this paper, we are focused on the monolingual word sense alignment (MWSA) task, which involves in sense alignment within two different resources in the same language. As an observer in the European Lexicographic Infrastructure–ELEXIS[3] (Krek et al., 2019; Declerck et al., 2018), the Academia das Ciências de Lisboa (ACL) contributed to the task of MWSA in which the Portuguese Academy of Sciences Dictionary is compared to and aligned with the senses in the *Dicionário Aberto*. We will report our experiences in annotating the senses with four semantic relationships, namely, narrower, broader, exact and related. Representing the final data in the Ontolex-Lemon model (McCrae et al., 2017), we believe that the outcomes of this project will pave the way for further research on automatic WSA for the Portuguese language and enhance the accessibility of the data on the Semantic Web and Linked Data.

The rest of the paper is organized as follows. In Section 2, we introduce our Portuguese lexicographic resources and provide a description of their content and structure. Section 3 summarises the methodology for annotation workflow. In Section 4, we point out the major challenges of the MWSA task for the Portuguese resources. We describe the conversion of the data into Ontolex-Lemon model in Section 5. Finally, we conclude in Section 6 with a summary of our contributions.

2. Lexicographic data

In the scope of ELEXIS, one of the main purposes is to extract, structure and link multilingual lexicographic resources. One of the tasks to achieve this goal consists

[1]`https://en.wiktionary.org`
[2]`https://de.wikipedia.org`
[3]This project aims to create a European network of lexical resources (`http://www.elex.is`).

of word sense alignment manual task in several languages (Ahmadi et al., 2020).The datasets are publicly freely available[4].The first established task is to provide semantic relations, as we will demonstrate in Section 3.

2.1. DLPC and DA

For the completion of this task, we align the following two Portuguese dictionaries:

- the *Dicionário da Língua Portuguesa Contemporânea* (DLPC) (Academia das Ciências de Lisboa, 2001), with the seal of ACL, coordinated by Malaca Casteleiro and published in 2001, with the financial support of the Calouste Gulbenkian Foundation, under the commercial responsibility of Editorial Verbo. This dictionary also represents the first complete edition of a Portuguese Academy dictionary, from A to Z (previous attempts in 1793 and 1976 did not go further than the letter A). The DLPC contains around 70,000 entries. In 2015, some preparatory work for an online Portuguese Academy of Science Dictionary (DACL) was performed through the Instituto de Lexicologia e Lexicografia da Língua Portuguesa (ILLLP) and a database was developed by a team working in Natural Language Processing at the University of Minho, which now draws on the participation of IPCA and NOVA CLUNL[5]. The present work, therefore, had the retro-digitised version of DLPC as a starting point.

- the *Dicionário Aberto* (DA) (Simões and Farinha, 2010), a Portuguese language dictionary obtained by the full transcription of *Nôvo Diccionário da Língua Portuguêsa*, authored by Cândido de Figueiredo, and published in 1913 by Livraria Clássica. Having the 1913 edition entered the public domain, it was digitised and text-converted by a team of distributed proofreaders volunteers between 2007 and 2010 and was made publicly available on the Gutenberg Project website on 8 March 2010. During the transcription process, and as entries got reviewed, and therefore, considered final, they were made freely available on the web. For three years, the dictionary has expanded by including more transcribed entries. After the complete transcription, the dictionary was subject to automatic orthography update and was used for different experiments regarding NLP tasks, as the automatic extraction of information for the creation of Wordnets or ontologies (Gonçalo Oliveira, 2018; Oliveira and Gomes, 2014). The updated version of the dictionary is available under license CC-BY-SA 2.5 PT. The DA contains 128,521 entries. Although the number of entries seems high, it is necessary to bear in mind that this resource registers orthographic variants of the same entry as we will mention later.

[4] https://github.com/elexis-eu/MWSA

[5] The team works with Alberto Simões (IPCA) and José João Almeida (Natural Language Processing of the Computer Science Department), and the consultancy of Álvaro Iriarte Sanromán. The participation of NOVA CLUNL is related to the DACL's transition into the TEI Lex-0 format.

2.2. Formats

Concerning formats, both Portuguese language resources are available in printed editions and XML versions.

The DLPC was published in a two-volume paper version, the first volume from A to F and the second from G to Z, in a total of 3880 pages. This dictionary, available in print and as a PDF document, was converted into XML using a slightly customized version of the P5 schema of the Text Encoding Initiative (TEI) (Simões et al., 2016). The XML was generated based on the dictionary PDF file, from which most of the information on the microstructure was recovered automatically. The new ongoing digital edition, DACL, is only privately available and has been edited with LeXmart (Simões et al., 2019). At the same time, the dictionary is being converted to the TEI Lex-0 format (Salgado et al., 2019b), a streamlined version of the TEI Dictionary Chapter. The present work, therefore, had this digital version as a starting point.

Regarding the DA, the paper version comprises 2133 pages. Currently, the dictionary is available online. Unlike DLPC, DA was transcribed manually by volunteers. This task required that the annotation format would be easy to learn, but also, that it would be similar to the format used in the transcription of other books for the Project Gutenberg[6]. Therefore, entries were only annotated with changes of font types, i.e., italics and bold, and not semantic tags. Although the dictionary is also available in XML, following the general guidelines of the Dictionary Chapter of TEI, the annotation granularity is bigger than DLPC. Specific portions of the microstructure were easy to annotate. Consider, for example, the grammatical information, geographic variant, or the knowledge domain. These entities are from a controlled list of vocabulary, and after creating the list it was straightforward to annotate them. For the construction of these lists we used the tables from the front-matter of the dictionary. Nevertheless, as these lists were manually generated, they were completed by performing dummy runs of the tagging algorithm, and finding out parts of the entries that were not detected. For other situations, like the annotation of usage examples, or to distinguish between two different senses, there are no clear marks to allow an algorithm to perform that automatically. While some hints could help, a good annotation would require manual validation. Under DA every line in the definition element tag can be a different sense, but can also be a usage example or even the continuation of the previous sense definition (Simões et al., 2012). To correctly detect other parts of the microstructure would require further manual revision that was not possible at that time.

Further developments on both dictionaries are programmed as soon as funding is available.

2.3. Micro-structure analysis

The DLPC's micro-structure is more complex than the DA's, with more structured and hierarchical information. Both dictionaries follow lexicographic conventions such as bold type in headwords. Nevertheless, comparing the sample of entries, we may observe certain typographic differences: ACL features initial lowercase entries while the DA

[6] https://www.gutenberg.org/ebooks/31552

Headword (POS)	DLPC sense	Semantic relation	Sense match	DA sense
banco (s. m.)				
	Assento estreito e comprido, de material variável, com ou sem encosto, para várias pessoas.	related	Assento, geralmente tosco, de ferro, madeira ou pedra, e de formas variadas.	Assento, geralmente tosco, de ferro, madeira ou pedra, e de formas variadas.
	banco dos réus. 1. Lugar destinado aos réus, no tribunal. 2. Situação em que se é objecto de acusação em tribunal.	none		Escabelo.
	Assento para uma pessoa, sem encosto, de tampo redondo ou quadrado, sustentado por três ou quatro pés. ≈ mocho.	related	Assento, geralmente tosco, de ferro, madeira ou pedra, e de formas variadas.	Mesa estreita e oblonga, sobre que trabalham certos artífices.
	Assento comprido e largo, com encosto alto, de tampo amovível, que pode servir também de tampa de uma arca. ≈ arquibanco, escabelo, escano.	exact	Escabelo.	Balcão de comércio.

Figure 1: An example spreadsheet used for the annotation task.

has capitalized entries. Furthermore, only the DLPC provides full pronunciation information. The DLPC etymological information figures after the grammatical properties of the lexical item while, in the DA, such information appears at the end of the entry. While the DLPC indicates the part-of-speech and gender, the DA displays the gender in the case of nouns[7]. One of the main features of the DLPC is the split of entries. Not only etymological homonyms are treated as independent entries, but also homonyms of the same etymological family belonging to different part-of-speech are differentiated by numeric superscripts to the right of the lemma in order to distinguish the respective entries (e.g. *perfurador* can function as an adjective, or a noun so is split into two entries).

Regarding the structure, the senses are numbered in the DLPC, providing better organised and more fine-grained information, while in the DA only a paragraph distinguishes the different senses. This was the result of the lack of metadata added to the dictionary during the transcription process. Nevertheless, the dictionary has the basic microstructure annotated, including grammatical information, definitions, quotations, usage examples and etymological information. The DLPC has, in general, more structured information such as synonyms (preceded by ≈), examples (shown in italics), cross-reference to lexical units that preferentially co-occur are represented by the symbol +, usage labelling, among other relevant features.

In the next section, we will explain in more detail how the workflow annotation took place. The data was delivered in XLM files and in an Excel format where the data was converted into spreadsheets.

3. Methodology

In the previous two sections, we have presented the resources we decided to analyze and pointed out that they have very different features. Before we move to the annotation workflow, we would like to define some of the terms used in this particular task:

- The lemma is a "lexical unit chosen according to lexicographical conventions to represent the different forms of an inflection paradigm" (ISO, 2007).
- A sense is one of the possible meanings or interpretations in a specific context.
- A gloss is a textual description of a sense's meaning meant for human interpretation.

3.1. Entries selection

The selection of entries took into account some points previously defined by the ELEXIS team (Ahmadi et al., 2020), namely: all open class words should be represented; monosemous and polysemous lemmas should appear; and, finally, the lemmas of both resources must had the same part-of-speech. Taking these points into account, we decided to select isolated lemmas randomly and also select data sets followed alphabetically. As a sample of entries, we chose:

A. random entries as long as they appeared in both dictionaries: *banco* [bank], *bandarilha* [banderilla], *café* [coffee], *computador* [computer], *coração* [heart], *dicionário* [dictionary], *futebol* [football], *lexicografia* [lexicography], *mililitro* [milliliter], *praia* [beach], *sorridente* [smiling] and *tripeiro* [tripe seller and native of Porto].

B. all the lexical items that came up between *especial* [special] and *esperanto* [Esperanto], *perfume* [perfume] and *perlimpimpim* [a lexical unit used in a fixed combination *pós de perlimpimpim* [magical powder], a sequence of units sorted alphabetically from letters E and P.

The total number of entries collected is 146 containing 786 distinct senses (8301 tokens).

After selecting the sample entries, we created dynamic spreadsheets as the means of the annotation task (Figure 1). This sheet contains the following information:

[7]This a common lexicographic practice: when it is marked as *m.* (masculine), it is understood that the lemma is a noun.

headwords (DLPC and DA lemmas identification); part-of-speech (DLPC POS); senses in DLPC (DLPC senses); semantic relation; sense match (DA equivalent sense); part-of-speech (DA POS); and, finally, senses in DA (DA senses).

3.2. Annotation workflow

The annotation task was carried out fully manually. Given a lemma, corresponding senses in both dictionaries, the DA and DLPC, were brought together in the spreadsheets. This way, all the possible combinations of the senses across the two resources were provided to the annotator. Unlike regular dictionaries, where a limited number of semantic relationships are defined, such as synonymy and antonymy, we considered a broader range of semantic relationships, namely the followings:

- `exact`: the two senses are semantically equivalent;
- `narrower`: the sense in DLPC describes a narrower concept than that in the DA;
- `broader`: the sense in DLPC describes a broader concept than that in the DA;
- `related`: there is a possible alignment, detecting a possible related relationship.

In the case where no semantic relationship is found for a sense, none is selected. Note that not all the semantic relationships are symmetric; therefore, the order of the columns determines the relationship. We matched the senses of the two dictionaries, using the label corresponding to the properties cited above. The result is a mapping between senses. In overall, 463 and 323 senses are aligned in the DLPC and DA, respectively. Among the whole number of 275 aligned senses, 207 exact, 38 narrower, 28 related and 2 broader are provided.

4. Challenges of MWSA

We now move on to the challenges of WSA. When we first chose these two lexicographic resources, we knew that we would be dealing with a significant time lag: the DLPC was published in 2001, and the DA in 1913. In 88 years, the Portuguese lexicon and language undergone many transformations: a Portuguese spelling reform, semantic changes of the lexical items (*computador* [computer], for example, in the DA, is not defined as an electronic device, new words have appeared, such as *futebol* [football], which is not included in the DA). All these factors are obstacles to the successful performance of this task.

The Portuguese spelling has also changed. In the DA, their development team decided to maintain old spelling variants, e.g. *periphrástico* and *perifrástico* (Figure 2), thus enabling the search of all the orthographic variants.

For this task, we have ignored the old orthographic variant forms of a given lexical unit, as they are present in duplicate in DA (with an updated version of the form). Since the DLPC is a contemporary dictionary, these orthographic forms would never appear in the DA and were not useful for the ongoing task[8].

Figure 2: *periphrástico* [periphrastic] and *perifrástico* [periphrastic] in DA

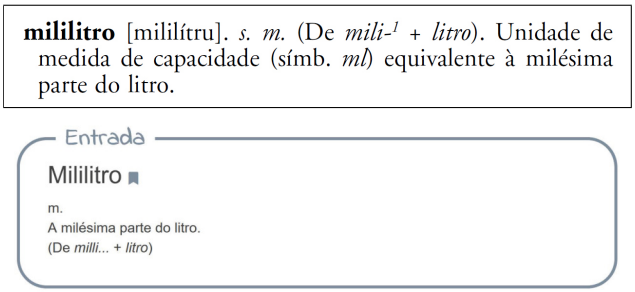


Figure 3: *mililitro* [milliliter] in DLPC (above) and DA (below)

Since we do not intend to discuss the wording techniques of the gloss, we can say that between certain lexical items senses, there is an exact correspondence of sense. There are cases where we can establish an exact relation between the senses even in structural terms (see, *mililitro* [millilitre] that has only one sense in both dictionaries, i.e., one-thousandth of a litre). However, these easily solvable cases are not what we mostly encounter when dealing with different dictionaries (Figure 3).

There are several other cases where there are exact relations, but there are other senses that appear in only one of the dictionaries. In Figure 4, DLPC sense 1 related to the bullfighting domain [banderilla] corresponds to the only sense of the DA. Sense 2 related to the bookbinding domain only appears in the DLPC.

Nevertheless, and although the first sense is identical in both resources, the disallowance is not identical in textual terms, since the meaning is described differently. The

[8]From the DA XML file, we ignored the following entries: *perhydrol, perianthado, periântheo, periânthio, periantho, periappendicite, perichécio, perichôndrio, perichondrite, perichondrio, pericoróllia, pericyclo, pericystite, perididymite, peridídymo, perídyo, perígrapho, perigynândrio, perigynadro, perigynia, perígyno, perimísio, perimorphose, perinephrite, periophthalmia, periorthógono, periosteóphyto, peripheria, periphérico, periphorantho, períphoro, períphrase, periphrástico, peripyema, peristáchio, peristéthio, peristýlico, perissýstole, perithécio, perityphlite.*

DLPC also uses a domain label, "*Taurom.*" while in the DA, there is no label.

bandarilha [bɐ̃dɐɾíʎɐ]. *s. f.* (Do cast. *banderilla*). **1.** *Taurom.* Haste munida de ponta de metal penetrante, enfeitada com uma bandeira ou com fitas de papel de cores e que se espeta no cachaço dos touros, durante a corrida. ≃ FARPA, FERRO. *A elegância com que espetou o par de bandarilhas no touro pôs a praça de pé. «Abrem-se então as portas e a manada entra, esta que será toureada hoje consoante os preceitos inteiros da arte, passada à capa, espetada de bandarilhas, castigada de varas»* (SARAMAGO, *Levantado do Chão*, p. 165). *Cravar, espetar as +s; um par de +s; tércio de +s.* **bandarilhas a quarteio,** variedade de farpas em que o toureiro faz um quarto de volta ao espetá-la no touro. **bandarilhas a recorte,** movimento que consiste em colocar os ferros no touro no momento em que o toureiro evita a marrada. **2.** *Encad.* Tira de papel que se cola na margem de um original ou prova, quando as emendas não cabem nas margens.

Entrada

Bandarilha

f.
Farpa, enfeitada com bandeiras ou fitas, e destinada a cravar-se no cachaço dos toiros, quando se correm.
(Por *bandeirilha*, cast. *banderilla*)

Figure 4: *bandarilha* [banderilla] in DLPC (above) and DA (below)

In other cases, the correspondence of senses is evident, but the lexicographic criteria adopted differ as shown in Figure 5. The structure of these lexicographic articles is different. The DLPC has two entries for *tripeiro* (*tripeiro*[1] and *tripeiro*[2]) as an adjective and a noun, part-of-speech homonyms. The first entry is an adjective, and the second is a noun; the DA has only one entry and only gender information. Between *tripeiro*[2] (DLPC) and *tripeiro* (DA), there is an exact match in the first sense, an obsolete sense, as a tripe seller although the technique of writing the gloss differs ("Pessoa que vende tripas" [Person who sells tripes] in DLPC and "Vendedor de tripas" [Tripe seller]) in DA. These two glosses point to the same concept. However, although the DA did not record sense numbers, the first two senses could be divided. We can established a match between sense two that start with "pop." [popular] in DLPC and "Deprec." [depreciative] in DA, another tricky topic is usage information. This topic is related to the various types of inconsistencies regarding usage labelling (Salgado et al., 2019a). Anyway, the only difference is that DLPC uses a cross-reference, and the DA provides the gloss.
Other times, the senses are exact correspondences, but the editorial perspective is different as shown in the example of Figure 6: for *pergaminháceo* [pergameneous] (DLPC), the DA presents a gloss and the DLPC a cross-reference. On the other hand, *pergiminháceo* (DA) has a cross-reference *pergaminháceo*.
The DA, as mentioned above, does not use numbers for senses. Thus, we have considered each paragraph as an independent sense. However, a DLPC sense may correspond to more than one DA sense. See *praia* [beach] entry in the sense of "Beira-mar" [seaside] (Figure 7).

tripeiro[1] [tripɐ́jru]. *adj. m.* e *f.* (De *tripa* + suf. *-eiro*). *Pop.* O m. que *portuense*[1].
tripeiro[2] [tripɐ́jru]. *s. m.* e *f.* (De *tripa* + suf. *-eiro*). **1.** Pessoa que vende tripas. **2.** *Pop.* O m. que *portuense*[2].

Entrada

Tripeiro

m.
Vendedor de tripas.
Aquele que se sustenta de tripas.
Deprec.
Habitante do Porto.
(De *tripa*)

Figure 5: *tripeiro* [tripe seller and native of Porto] in DLPC (above) and DA (below)

pergaminháceo, a [pɨrgɐminásju, -ɐ]. *adj.* (Do b. lat. *pergamīnum* 'pergaminho' + suf. *-áceo*). **1.** Que se assemelha ao pergaminho; que, pelo seu aspecto, faz lembrar essa pele. ≃ PERGAMINHÁCEO. **2.** Que é feito de pergaminho.

pergaminháceo, a [pɨrgɐmiɲásju, -ɐ]. *adj.* (De *pergaminho* + suf. *-áceo*). O m. que *pergaminháceo*.

Entrada

Pergaminháceo

adj.
O mesmo ou melhor que *pergaminháceo*. Cf. Arn. Gama, *Últ. Dona*, 55.

Entrada

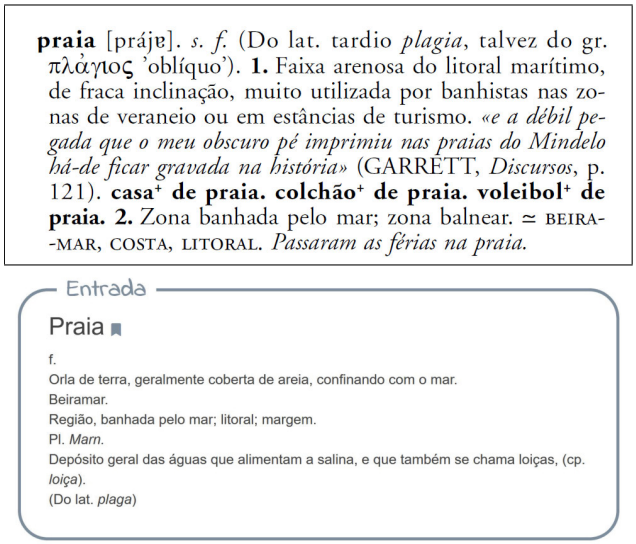
Pergaminháceo

adj.
Que tem o aspecto de pergaminho: «...*os musculos... pergaminháceos...*» Camilo, *Volcões*, 154.

Figure 6: *pergaminháceo/pergaminháceo* [pergameneous] in DLPC (above) and DA (below)

In the DA (Figure 7), the senses "Beira-mar" [seaside] and "Região, banhada pelo mar; litoral; margem" [Region, bathed by the sea; coast] correspond to sense 2 of the DLPC: "Zona banhada pelo mar; zona balnear" [Zone bathed by the sea; bathing area].
The same can be said, for example, of *especial* [special], whose DLPC gloss, "Que tem, dadas as características, uma finalidade ou um uso particular. ≈ adequado, específico, próprio. ≠ geral." [Which has, given the characteristics, a purpose or a particular use. ≈ suitable, specific, own], may correspond to three paragraphs of the DA: "Próprio. / Peculiar. / Particular." [Own. / Peculiar. / Particular.].
Looking at the three glosses of banco [stool/bench] as "assento" [seat] in the DLPC:

- "Assento estreito e comprido, de material variável, com ou sem encosto, para várias pessoas." [Narrow and long seat, of variable material, with or without backrest, for several people.]
- "Assento para uma pessoa, sem encosto, de tampo re-



praia [práje]. *s. f.* (Do lat. tardio *plagia*, talvez do gr. πλάγιος 'oblíquo'). **1.** Faixa arenosa do litoral marítimo, de fraca inclinação, muito utilizada por banhistas nas zonas de veraneio ou em estâncias de turismo. *«e a débil pegada que o meu obscuro pé imprimiu nas praias do Mindelo há-de ficar gravada na história»* (GARRETT, *Discursos*, p. 121). **casa⁺ de praia. colchão⁺ de praia. voleibol⁺ de praia. 2.** Zona banhada pelo mar; zona balnear. ≃ BEIRA-MAR, COSTA, LITORAL. *Passaram as férias na praia.*

Entrada

Praia

f.
Orla de terra, geralmente coberta de areia, confinando com o mar.
Beiramar.
Região, banhada pelo mar; litoral; margem.
Pl. *Mam.*
Depósito geral das águas que alimentam a salina, e que também se chama loiças, (cp. *loiça*).
(Do lat. *plaga*)

Figure 7: *praia* [beach] in DLPC (above) and DA (below)

dondo ou quadrado, sustentado por três ou quatro pés. ≈ mocho." [One person seat, without backrest, with round or square top, supported by three or four feet; stool]

- "Assento comprido e largo, com encosto alto, de tampo amovível, que pode servir também de tampa de uma arca. ≈ arquibanco, escabelo, escano.)" [Long and wide seat, with high back, removable top, which can also serve as a chest lid. ≈ bench cabinet; bench.]

It is tough to ascertain whether it is possible to make a correspondence with the first sense of the DA, also this one related to a seat: "Assento, geralmente tosco, de ferro, madeira ou pedra, e de formas variadas." [Seat, usually rough, of iron, wood or stone, and of different shapes.]
The last sense of the DLPC is a synonym of "escabelo" (also in the DA, so this is an "exact" correspondence), but it may also be associated with the first sense of the DLPC.
Let us now turn to the *lexicografia* [lexicography] entry in the DLPC:

- "Ling. Ramo da linguística que se ocupa dos aspectos teóricos e práticos que têm em vista a elaboração de dicionários, vocabulários, glossários." [Branch of linguistics that deals with the theoretical and practical aspects that aim to develop dictionaries, vocabularies, glossaries.]

The same entry in DA, it is described as:

- "Ciência ou estudo, que tem por objecto as palavras que devem constituir um léxico." [Science or study, whose object is the words that must constitute a lexicon.]

Although the gloss differs (we intend to explore the issue of definition in more detail in future work), in these cases, we always attribute an exact relationship since both refer to the same concept.

5. Data Conversion

In order to increase the interoperability of the annotated data with other language resources, we convert the final datasets into the Ontolex-Lemon model (McCrae et al., 2017). This model provides rich linguistic groundings for ontologies which enables various representations such as morphology and syntax. Our final output provides the headword, the part-of-speech tag along with the senses for each entry. Therefore, the following properties are respectively used: `ontolex:writtenRep`, `lexinfo:partOfSpeech` and `skos:definition`. Linking between the senses is made with the SKOS matching properties. An example of this data in Turtle is given below:

```
<#banco_noun> a ontolex:LexicalEntry ;
  rdfs:label "banco"@pt ;
  ontolex:sense <#sense0>, <#sense12>,
     <#sense13> .

<#sense0> skos:definiton
  "Assento estreito e comprido, de
  material variável, com ou sem encosto,
  para várias pessoas. "@pt .

<#sense12> skos:definition
  "banco dos réus. 1. Lugar destinado
  aos réus, no tribunal. 2. Situação
  em que se é objecto de acusação
  em tribunal."@pt .

<#sense0> skos:relatedMatch <#sense1> .
<#sense95> skos:exactMatch <#sense96> .
<#sense97> skos:narrowMatch <#sense96> .
```

The data is publicly available as part of the MWSA benchmark at `https://github.com/elexis-eu/MWSA`.

6. Conclusion

This paper focuses on the task of monolingual word sense alignment for the Portuguese language. Focusing on two lexicographic resources in Portuguese, namely, *Dicionário da Língua Portuguesa Contemporânea* and *Dicionário Aberto*, we presented the challenges and difficulties to manually align senses and annotate their semantic relationships. In addition, we also describe the conversion of our aligned data into the Ontolex-Lemon model which improves interoperability and accessibility within the Linked Data and Semantic Web technologies. We believe that our dataset is beneficial to create tools and techniques to automatically align senses within Portuguese lexicographic resources. Moreover,

7. Acknowledgements

The authors would like to thank the anonymous reviewers for their thoughtful comments towards improving our manuscript. 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/03213/2020, by the European Union's Horizon 2020 research and innovation

program under Grant Agreement No. 731015 (ELEXIS), and through the FCT/MCTES as part of the project 2Ai – School of Technology, IPCA – UIDB/05549/2020.

8. Bibliographical References

Ahmadi, S., McCrae, J. P., Nimb, S., Khan, F., Monachini, M., Pedersen, B. S., Declerck, T., Wissik, T., Bellandi, A., Pisani, I., Troelsgård, T., Olsen, S., Krek, S., Lipp, V., Váradi, T., Simon, L., Győrffy, A., Tiberius, C., Schoonheim, T., Ben Moshe, Y., Rudich, M., Abu Ahmad, R., Lonke, D., Kovalenko, K., Langemets, M., Kallas, J., Dereza, O., Fransen, T., Cillessen, D., Lindemann, D., Alonso, M., Salgado, A., Sancho, J. L., Ureña-Ruiz, R.-J., Simov, K., Osenova, P., Kancheva, Z., Radev, I., Stanković, R., Perdih, A., and Gabrovšek, D. (2020). A Multilingual Evaluation Dataset for Monolingual Word Sense Alignment. In *Proceedings of the 12th Language Resource and Evaluation Conference (LREC 2020)*, Marseille, France.

Chiarcos, C., Nordhoff, S., and Hellmann, S. (2012). *Linked Data in Linguistics: Representing and Connecting Language Data and Language Metadata*. Springer.

Declerck, T., McCrae, J., Navigli, R., Zaytseva, K., and Wissik, T. (2018). Elexis-european lexicographic infrastructure: Contributions to and from the linguistic linked open data. In *I. Kernerman & S. Krek (Arg.), Proceedings of the LREC 2018 Workshop Globalex*, pages 17–22.

Gonçalo Oliveira, H. (2018). A survey on portuguese lexical knowledge bases: Contents, comparison and combination. *Information*, 9(2):34.

Gracia, J., Kernerman, I., and Bosque-Gil, J. (2017). Toward linked data-native dictionaries. In *Electronic Lexicography in the 21st Century: Lexicography from Scratch. Proceedings of the eLex 2017 conference*, pages 19–21.

Heath, T. and Bizer, C. (2011). Linked data: Evolving the web into a global data space. *Synthesis lectures on the semantic web: theory and technology*, 1(1):1–136.

Henrich, V., Hinrichs, E. W., and Suttner, K. (2012). Automatically linking germanet to wikipedia for harvesting corpus examples for germanet senses. *JLCL*, 27(1):1–19.

ISO. (2007). Presentation/representation of entries in dictionaries — Requirements, recommendations and information. Standard, International Organization for Standardization, Geneva, CH, February.

Knight, K. and Luk, S. K. (1994). Building a large-scale knowledge base for machine translation. In *AAAI*, volume 94, pages 773–778.

Krek, S., Declerck, T., McCrae, J. P., and Wissik, T. (2019). Towards a Global Lexicographic Infrastructure. In *Proceedings of the Language Technology 4 All Conference*.

McCrae, J. P. and Declerck, T. (2019). Linguistic Linked Open Data for All. In *Proceedings of the Language Technology 4 All Conference*.

McCrae, J. P., Bosque-Gil, J., Gracia, J., Buitelaar, P., and Cimiano, P. (2017). The ontolex-lemon model: development and applications. In *Proceedings of eLex 2017 conference*, pages 19–21.

Meyer, C. M. and Gurevych, I. (2011). What psycholinguists know about chemistry: Aligning wiktionary and wordnet for increased domain coverage. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 883–892.

Moro, A., Li, H., Krause, S., Xu, F., Navigli, R., and Uszkoreit, H. (2013). Semantic rule filtering for web-scale relation extraction. In *International Semantic Web Conference*, pages 347–362. Springer.

Navigli, R. and Ponzetto, S. P. (2012). Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Navigli, R. (2006). Meaningful clustering of senses helps boost word sense disambiguation performance. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 105–112. Association for Computational Linguistics.

Oliveira, H. G. and Gomes, P. (2014). Eco and onto. pt: a flexible approach for creating a portuguese wordnet automatically. *Language resources and evaluation*, 48(2):373–393.

Salgado, A., Costa, R., and Tasovac, T. (2019a). Improving the consistency of usage labelling in dictionaries with tei lex-0. *Lexicography*, 6(2):133–156.

Salgado, A., Costa, R., Tasovac, T., and Simões, A. (2019b). Tei lex-0 in action: Improving the encoding of the dictionary of the academia das ciências de lisboa. *Electronic lexicography in the 21st century (eLex 2019): Smart lexicography*, page 93.

Shadbolt, N., Berners-Lee, T., and Hall, W. (2006). The semantic web revisited. *IEEE intelligent systems*, 21(3):96–101.

Simões, A., Sanromán, Á. I., and Almeida, J. J. (2012). Dicionário-aberto: A source of resources for the portuguese language processing. In *International Conference on Computational Processing of the Portuguese Language*, pages 121–127. Springer. `http://www.dicionario-aberto.net/`.

Simões, A., Almeida, J. J., and Salgado, A. (2016). Building a dictionary using xml technology. In *5th Symposium on Languages, Applications and Technologies (SLATE'16)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Simões, A., Salgado, A., Costa, R., and Almeida, J. J. (2019). LeXmart: A smart tool for lexicographers. In I. Kosem, et al., editors, *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference*, pages 453–466.

9. Language Resource References

Academia das Ciências de Lisboa. (2001). *Dicionário da Língua Portuguesa Contemporânea*. João Malaca Casteleiro (ed.). Lisboa. Academia das Ciências de Lisboa e Editorial Verbo.

Simões, A. and Farinha, R. (2010). Dicionário aberto: um recurso para processamento de linguagem natural. *Viceversa: revista galega de traducción*, 16:159–171.