<table>
<thead>
<tr>
<th><strong>Title</strong></th>
<th>Challenges of word sense alignment: Portuguese language resources</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Author(s)</strong></td>
<td>Salgado, Ana; Ahmadi, Sina; Simões, Alberto; McCrae, John P.; Costa, Rute</td>
</tr>
<tr>
<td><strong>Publication Date</strong></td>
<td>2020-05-16</td>
</tr>
<tr>
<td><strong>Publisher</strong></td>
<td>National University of Ireland Galway</td>
</tr>
<tr>
<td><strong>Link to publisher's version</strong></td>
<td><a href="https://doi.org/10.13025/htat-gw66">https://doi.org/10.13025/htat-gw66</a></td>
</tr>
<tr>
<td><strong>Item record</strong></td>
<td><a href="http://hdl.handle.net/10379/15922">http://hdl.handle.net/10379/15922</a></td>
</tr>
<tr>
<td><strong>DOI</strong></td>
<td><a href="http://dx.doi.org/10.13025/htat-gw66">http://dx.doi.org/10.13025/htat-gw66</a></td>
</tr>
</tbody>
</table>
Challenges of Word Sense Alignment: Portuguese Language Resources

Ana Salgado1,2, Sina Ahmadi3, Alberto Simões4, John McCrae3, Rute Costa2
1Academia das Ciências de Lisboa, Instituto de Lexicologia e Lexicografia da Língua Portuguesa, Lisbon, Portugal
2NOVA CLULN, Universidade NOVA de Lisboa, Lisbon, Portugal
3Data Science Institute, National University of Ireland Galway
42Ai – 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 Wiktionary1, and Henrich et al. (2012) link the GermaNet—the German Wordnet with the German Wikipedia2.

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–ELEXIS3 (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

1https://en.wiktionary.org
2https://de.wikipedia.org
3This 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. 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 **Diccionário 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. The present work, therefore, had the retro-digitised version of DLPC as a starting point.

- the **Diccionário Aberto** (DA) (Simões and Farinha, 2010), a Portuguese language dictionary obtained by the full transcription of **Novo Diccionário da Língua Portuguesa**., 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.

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. 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

\[4\text{https://github.com/elexis-eu/MWSA}\]

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

\[6\text{https://www.gutenberg.org/ebooks/31552}\]
<table>
<thead>
<tr>
<th>Headword (POS)</th>
<th>DLPC sense</th>
<th>Semantic relation</th>
<th>Sense match</th>
<th>DA sense</th>
</tr>
</thead>
<tbody>
<tr>
<td>banco (s. m.)</td>
<td>Assento estreito e comprido, de material variável, com ou sem encostos, para várias pessoas.</td>
<td>related</td>
<td>Assento, geralmente tosco, de ferro, madeira ou pedra, e de formas variadas.</td>
<td>Assento, geralmente tosco, de ferro, madeira ou pedra, e de formas variadas.</td>
</tr>
<tr>
<td></td>
<td>banco dos réus. 1. Lugar destinado aos réus, no tribunal. 2. Situação em que se é objeto de acusação em tribunal.</td>
<td>none</td>
<td>Escabelo.</td>
<td>Escabelo.</td>
</tr>
<tr>
<td></td>
<td>Assento para uma pessoa, sem encostos, de tampo redondo ou quadrado, sustentado por três ou quatro pés. ≈ mocho.</td>
<td>related</td>
<td>Assento, geralmente tosco, de ferro, madeira ou pedra, e de formas variadas.</td>
<td>Mesa estreita e oblonga, sobre que trabalham certos artífices.</td>
</tr>
<tr>
<td></td>
<td>Assento comprido e largo, com encosto alto, de tampo amovível, que pode servir também de tampa de uma arca. ≈ arquibanco, escabelo, escano.</td>
<td>exact</td>
<td>Escabelo.</td>
<td>Balcão de comércio.</td>
</tr>
</tbody>
</table>

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

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 have 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 perlimpinpin [a lexical unit used in a fixed combination pós de perlimpinpin [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:

7This 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. perifrástico 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.

Since we do not intend to discuss the wording techniques of the gloss, we can say that between certain lexical items, 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 [milliliter] 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.

\[\text{Figure 2: perifrástico [periphraistic] and perifrástico [periphraistic] in DA}\]

\[\text{Figure 3: mililitro [milliliter] in DLPC (above) and DA (below)}\]

\[\text{From the DA XML file, we ignored the following entries: perhydrol, perianthado, periântaho, periantaho, peri-}
\[\text{antio, periappendicite, perichécio, perichôndrio, perichondrite,}
\[\text{perichondrio, pericorôllia, pericyclo, pericystite, perididlome,}
\[\text{perididymo, peridíodo, perigrapho, perignândrio, perignadrio,}
\[\text{perigniaux, perignia, perinísmo, perimorphose, perinephrite, pe-}
\[\text{riothamnia, periorióngono, periosteóphyto, peripheria, peripé-}
\[\text{rhipérico, periperhanhoro, periphoro, periphraise, periphraístico,}
\[\text{peripêma, peristâchio, peristêtvio, peristylíco, perissystole, perihéció, perisyphlite.}\]
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 tripéiro1 and tripéiro2 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 tripéiro2 (DLPC) and tripéiro (DA), there is an exact match in the first sense, an obsolete sense, as a tripé 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 establish 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 pergamináceo [pergameneous] (DLPC), the DA presents a gloss and the DLPC a cross-reference. On the other hand, pergaminháceo (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).

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

**bandarilha** [bandarilha], 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 espetá no cachaço dos touros, durante a corrida. = Farpa, Ferro. A elegância com que espetam o par de bandarilhas no touro pê a praça de pê. «Abrem-se então as portas e a matada entra, esta que será toureada hoje consoante os preceitos íntimos 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; tético de +s. bandarilhas a quarteto, variedade de farpas em que o toureiro faz um quarto de volta ao esperti-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.

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

**Figure 5:** *tripéiro* [tripère] 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 [práij]. 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 estáticas de turismo, «e a debil pe-
gada que o meu obscuro pé imprímui nas praias do Mindele há-de ficar gravada na história» (GARRETT, Discursos, p. 121). caso* de praia, colchão* de praia, voleibol* de praia. 2. Zona banhada pelo mar; zona balnear; = beira-
-mar, costa, litoral. Passaram as férias na praia.

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,
coreado 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:

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

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:

The data is publicly available as part of the MWSA bench-
mark at https://github.com/elixis-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, Di-
cionário da Língua Portuguesa Contemporânea and Di-
cionário Aberto, we presented the challenges and difficul-
ties to manually align senses and annotate their semantic re-
lationships. In addition, we also describe the conversion of
our aligned data into the Ontolex-Lemon model which im-
proves interoperability and accessibility within the Linked
Data and Semantic Web technologies. We believe that our
dataset is beneficial to create tools and techniques to au-
tomatically align senses within Portuguese lexicographic
resources. Moreover,

7. Acknowledgements

The authors would like to thank the anonymous review-
ers for their thoughtful comments towards improving our
manuscript. Research financed by Portuguese National
Funding through the FCT – Fundação para a Ciência e Tec-
nologia as part of the project Centro de Linguística da Uni-
versidade 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


9. Language Resource References
