01010
01010
01010 information

MDPI

*Article*

# A Survey on Portuguese Lexical Knowledge Bases: Contents, Comparison and Combination †

**Hugo Gonçalo Oliveira**

Department of Informatics Engineering, Centre for Informatics and Systems of the University of Coimbra (CISUC), University of Coimbra, 3030-290 Coimbra, Portugal; hroliv@dei.uc.pt

† This paper is an extended version of our paper published in Comparing and Combining Portuguese Lexical-Semantic Knowledge Bases, Proceedings of the 6th Symposium on Languages, Applications and Technologies (SLATE 2017), 26–27 June 2017, Vila do Conde, Portugal.

Received: 28 December 2017; Accepted: 31 January 2018; Published: 2 February 2018

**Abstract:** In the last decade, several lexical-semantic knowledge bases (LKBs) were developed for Portuguese, by different teams and following different approaches. Most of them are open and freely available for the community. Those LKBs are briefly analysed here, with a focus on size, structure, and overlapping contents. However, we go further and exploit all of the analysed LKBs in the creation of new LKBs, based on the redundant contents. Both original and redundancy-based LKBs are then compared, indirectly, based on the performance of automatic procedures that exploit them for solving four different semantic analysis tasks. In addition to conclusions on the performance of the original LKBs, results show that, instead of selecting a single LKB to use, it is generally worth combining the contents of all the open Portuguese LKBs, towards better results.

**Keywords:** lexical knowledge bases; Portuguese; WordNet; redundancy; semantic similarity

## 1. Introduction

Lexical-semantic knowledge bases (LKBs) are computational resources that organise words according to their meaning. In addition other features, they should have a significant coverage of the words of a language, which, according to their possible senses, should be connected by means of semantic relations. Princeton WordNet [1] is the paradigmatic resource of this kind, for English, used in many natural language processing (NLP) tasks, and with a model also adapted to many other languages, including Portuguese. However, the first Portuguese WordNet [2] is not available to be used by the research community and the first open alternatives were only developed in the last decade.

In order to cope with the lack of such a resource, several open Portuguese LKBs were since then created and most became available for download, either upon a paid license (e.g., MWN.PT (http://mwnpt.di.fc.ul.pt/)) or for free. However, those LKBs were developed by different teams, following different approaches, which resulted in LKBs with variable coverage and with slightly different features, regarding their organisation. Due to the difficulties inherent in crafting such a broad resource manually, most Portuguese LKBs have some degree of automation in their creation process, which increases the chance of noise. Furthermore, not all follow the full WordNet model. For instance, even though all of them cover one or more types of semantic relations, not all handle word senses. In fact, none of them is as consensual as Princeton WordNet, which was created manually and has a large community of users, is for English. Finally, some Portuguese LKBs are not large enough, while others have an interesting size but include several incorrect, unfrequent or unuseful relations or lexical items.

In this paper, ten open Portuguese LKBs are characterised in terms of covered lexical items and semantic relations. The redundancy across them is then analysed, towards the creation of (potentially) more useful LKBs. All the LKBs, including the new ones, are finally compared indirectly, when exploited in semantic similarity tasks with available benchmark datasets for Portuguese, namely: (i) given a word, selecting the most similar word from a predefined set; (ii) quantifying the semantic similarity of two words; (iii) filling a blank in a sentence with the correct word from a set; and (iv) quantifying the semantic textual similarity between two sentences. In addition, confirming our intuition that there are advantages in combining different LKBs, this can be seen as the first systematic comparison of the open Portuguese LKBs.

This is an extended version of a previously published paper [3], where a more detailed comparison is made, including a conversion table that maps semantic relation names in different LKBs; where two new resources are considered (ConceptNet and CARTÃO) as well as the most recent version of another (PULO). This resulted in different redundancy-based LKBs and, consequently, new experimentation results.

## 2. Related Work

The current scenario for Portuguese LKBs can be seen as atypical. There are currently many open LKBs for this language, but none is as consensual as Princeton WordNet [1] is for English. The latter started to be used by the NLP community in a time when there was nothing similar in terms of representation of the mental lexicon, with its coverage, granularity, reliability and, of course, the key factor of being freely available. On the other hand, the first Portuguese WordNet [2] was only released about a decade later and was not available to be used by the research community. Therefore, several related projects started, concurrently, for Portuguese. Those include several wordnets [4] and other simpler LKBs that, in some cases, may replace a wordnet. Looking at the *Wordnets in the World* list available at the site of the Global WordNet Association (http://globalwordnet.org/wordnets-in-the-world/ (January 2018)), one can see that previous problem is probably not specific to Portuguese. There are other languages with more than one wordnet, available or not under different licenses. For other languages, there is one "main" LKB used by the NLP community, possibly further enriched or aligned with different knowledge bases in specific domains or kinds of knowledge. For instance, there are several extensions for Princeton WordNet (e.g., subject field codes [5]), as well as alignments with other lexical resources (e.g., FrameNet and VerbNet [6], or Wikipedia and Wiktionary [7]). WordNet is also the "core" of most multilingual wordnets (e.g., EuroWordNet [8], MultiWordNet [9], Open Multilingual WordNet [10], MCR [11]) and of multilingual knowledge bases that cover linguistic and encyclopaedic knowledge (e.g., Universal WordNet [12], BabelNet [13]).

This is probably why there is not much work similar to what is presented here, where LKBs that aim at covering more or less the same kind of knowledge are combined. On the other hand, redundancy models have been proposed for assessing the confidence of relations automatically extracted from corpora [14]. The main intuition is that relation instances extracted more often, from different sources, are more plausible to be correct or useful.

## 3. Open Portuguese LKBs

Ten open Portuguese knowledge bases with lexical-semantic information were identified and explored in this work, namely:

- Three wordnets: WordNet.Br [15], OpenWordNet-PT (OWN.PT) [16] and PULO [17];
- Two synset-based thesauri: TeP [18] and OpenThesaurus.PT (http://paginas.fe.up.pt/~arocha/AED1/0607/trabalhos/thesaurus.txt (January 2018)) (OT.PT);
- Three lexical-semantic networks extracted from Portuguese dictionaries: PAPEL [19], relations extracted from Dicionário Aberto (DA) [20], and relations extracted from Wiktionary.PT (http://pt.wiktionary.org (2015 dump));
- Semantic relations available in Port4Nooj [21], a set of linguistic resources.
- Semantic relations between Portuguese words in the ConceptNet [22] semantic network, which includes common-sense knowledge, lexical knowledge and others.

As these resources do not share exactly the same structure, to enable their comparison and integration, they were all reduced to a set of relation instances of the kind "*x related-to y*", where *x* and *y* are lexical items and *related-to* is the name of a semantic relation. For synset-based LKBs, wordnets and thesauri, synsets had to be deconstructed. For example, the instance {*porta, portão*} partOf {*automóvel, carro, viatura*} resulted in: (*porta* synonymOf *portão*), (*automóvel* synonymOf *carro*), (*automóvel* synonymOf *viatura*), (*carro* synonymOf *viatura*), (*porta* partOf *automóvel*), (*porta* partOf *carro*), (*porta* partOf *viatura*), (*portão* partOf *automóvel*), (*portão* partOf *carro*), (*portão* partOf *viatura*)—In English, {*door, gate*} partOf {*automobile, car*} resulted in: (*door* synonymOf *gate*), (*automobile* synonymOf *car*), (*door* partOf *automobile*), (*door* partOf *car*), (*gate* partOf *automobile*), (*gate* partOf *car*). Adopted relation names were those defined in the project PAPEL [19], a rich set that covered most relation types in all the LKBs. However, some relation names, in other LKBs, had to be converted to a common name, always considering their semantics. Table 1 presents the performed conversions. Inverse relation names are omitted from this table, but they were also considered in the conversion process.

The size and type of contents of the LKBs obtained after conversion is summarised in Tables 2 and 3. Table 2 is focused on the number of covered lexical items, organised according to their part-of-speech (POS). Given that, without a context, the same lexical item may have different POS, the table also provides the number of distinct lexical items, when the POS is not considered. Table 3 targets the number of relations covered by each LKBs, grouped according to their broader types. The total number of relations is already provided, together with the average degree of each word, which measures the average number of relations involving each word in the network. A remark should be given on ConceptNet. In addition, being a slightly different knowledge base, not exclusively focused on lexical-semantic knowledge, it was also the last one to be included in this work. After analysing the set of available relations, several were not covered by our set of relation types. From this set, we discarded lexical relations, such as those related to word forms (e.g., FormOf, DerivedFrom, EtymologicallyRelatedTo), not so useful for semantic analysis, but we kept other interesting and potentially useful relation types (e.g., Desires, MotivatedByGoal). In the previous tables, the numbers of the latter types are only considered in the total, which is why the given number is followed by an asterisk (*). It should also be added that, for the converted relations, we only kept those for which we could identify the POS of both arguments. For this purpose, we used the POS provided by ConcepNet. However, as this information is only provided for some items, when it was not available, the possible POS of each word was automatically checked in the corpora of the AC/DC service [23]. More precisely, we considered that a word could have every POS with which its lemma occurred in AC/DC at least five times. It should also be mentioned that, although relation instances in the current version of ConcepNet have an attached confidence weight, the majority of the instances between two Portuguese words (≈95%) have this parameter set to 1.0, so it was not used.

Although the LKB with more lexical items is the one obtained from DA (≈95,000 distinct items), it contains substantially less relation instances than TeP, which covers ≈490,000 synonymy and antonymy instances but no other relation type. PAPEL, DA, OWN-PT and WN.Br all contain more than 100,000 relation instances. This is also noticeable from the average degree of each of those LKBs, which is the lowest in DA. On the other hand, WN.Br only covers verbs and is the smaller LKB in terms of lexical items, but the average degree of its words is substantially higher than others (36.9, followed by 11.9 in TeP). In fact, though lower than WN.Br, the average degrees of the synset-based LKBs are higher than for the others, which is, to some extent, a consequence of the synset deconstruction process.

On the relation types, all LKBs cover synonymy; antonymy is not covered by OT.PT, WN.Br and Port4Nooj; and hypernymy is not covered by TeP and OT.PT because the latter are originally synset-based thesauri. Other types are present in several LKBs (e.g., part, cause, property), but some types are only found in the LKBs extracted from dictionaries. ConcepNet also has an interesting range of covered types, where we highlight the quantity of purpose-of and place-of relations.

**Table 1.** Conversion of relations in different LKBs.

| POS | PAPEL, DA, Wikt.PT | TeP | OT.PT | OWN.PT | PULO | WN.Br | Port4Nooj | ConceptNet |
|---|---|---|---|---|---|---|---|---|
| Synonymy | SINONIMO_[N \| V \| ADJ \| ADV]_DE | *same synset* | *same synset* | *same synset* | *same synset* | *same synset* | É SINÓNIMO DE | Synonym |
| Antonymy | ANTONIMO_[N \| V \| ADJ \| ADV]_DE | *synset connections* | – | antonymOf | near_antonym | – | – | Antonym<br>DistinctFrom |
| Hypernymy | HIPERONIMO_DE | – | – | hypernymOf | has_hyponym | hypernymOf | É_HIPÓNIMO_DE | IsA<br>DefinedAs |
| Part | PARTE_DE<br>PARTE_DE_ALGO_COM_PROPRIEDADE<br>PROPRIEDADE_DE_ALGO_PARTE_DE | –<br>– | –<br>– | partHolonymOf<br>entails | has_holo_part<br>– | –<br>– | –<br>– | PartOf |
| Member | MEMBRO_DE<br>MEMBRO_DE_ALGO_COM_PROPRIEDADE<br>PROPRIEDADE_DE_ALGOMEMBRO_DE | – | – | memberHolonymOf | has_holo_member | | | – |
| Material | MATERIAL_DE | – | – | substanceHolonymOf | has_holo_madeof | – | – | – |
| Contains | CONTIDO_EM<br>CONTIDO_EM_ALGO_COM_PROPRIEDADE | | – | –<br>– | –<br>– | –<br>– | –<br>– | –<br>– |
| Cause | CAUSADOR_DE<br>ACCAO_QUE_CAUSA<br>CAUSADOR_DA_ACCAO<br>CAUSADOR_DE_ALGO_COM_PROPRIEDADE<br>PROPRIEDADE_DE_ALGO_QUE_CAUSA | – | – | causes | causes | – | É RESULTADO DE<br>É ACÇÃO DE | Causes |
| Producer | PRODUTOR_DE<br>PRODUTOR_DE_ALGO_COM_PROPRIEDADE<br>PROPRIEDADE_DE_ALGO_PRODUTOR_DE | – | – | – | – | – | – | – |
| Purpose | FINALIDADE_DE<br>FAZ_SE_COM<br>FINALIDADE_DA_ACCAO<br>FAZ_SE_COM_ALGO_COM_PROPRIEDADE<br>FINALIDADE_DE_ALGO_COM_PROPRIEDADE | – | – | – | – | – | – | UsedFor |
| Property | DIZ_SE_SOBRE<br>DIZ_SE_DO_QUE | – | – | similarTo<br>attributeOf | related_to | – | – | RelatedTo |
| State | TEM_ESTADO<br>DEVIDO_A_ESTADO | | – | – | be_in_state | – | | – |
| Quality | TEM_QUALIDADE<br>DEVIDO_A_QUALIDADE | – | – | – | – | – | – | – |
| Manner | MANEIRA_POR_MEIO_DE<br>MANEIRA_COM_PROPRIEDADE | – | – | – | – | – | – | – |
| Place | LOCAL_ORIGEM_DE | – | – | – | – | – | – | AtLocation |

**Table 2.** Number of lexical items extracted from each LKB.

| | | | | Lexical Items | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **POS** | **PAPEL** | **DA** | **Wikt.PT** | **TeP** | **OT.PT** | **OWN.PT** | **PULO** | **WN.Br** | **Port4Nooj** | **ConceptNet** |
| Nouns | 56,660 | 61,334 | 30,170 | 17,244 | 6110 | 32,509 | 7372 | 0 | 8109 | 9225 |
| Verbs | 21,585 | 16,429 | 8918 | 8343 | 2856 | 3626 | 2721 | 5857 | 3161 | 12,718 |
| Adjectives | 22,561 | 18,892 | 9536 | 14,979 | 3747 | 4401 | 2742 | 0 | 1055 | 214 |
| Adverbs | 1376 | 3160 | 610 | 1138 | 143 | 1120 | 312 | 0 | 475 | 295 |
| **Distinct** | 94,165 | 95,188 | 45,345 | 40,499 | 12,782 | 40,940 | 12,135 | 5857 | 12,641 | 40,778 * |

* means that additional relation types were considered for computing the total.

**Table 3.** Number of triples extracted from each LKB.

| | | | | | Relations | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Type** | **PAPEL** | **DA** | **Wikt.PT** | **TeP** | **OT.PT** | **OWN.PT** | **PULO** | **WN.Br** | **Port4Nooj** | **ConceptNet** |
| Synonymy | 83,432 | 52,278 | 35,330 | 388,698 | 51,410 | 35,597 | 69,618 | 88,488 | 559 | 30,834 |
| Antonymy | 388 | 440 | 1263 | 92,234 | – | 5774 | 8816 | – | – | 1651 |
| Hypernymy | 49,210 | 46,079 | 22,931 | – | – | 78,854 | 55,053 | 73,302 | 15,303 | 11,627 |
| Part | 5491 | 4367 | 1574 | – | – | 14,275 | 2025 | – | – | 169 |
| Member | 6585 | 1057 | 1578 | – | – | 5153 | 357 | – | – | – |
| Material | 336 | 518 | 192 | – | – | 958 | 88 | – | – | – |
| Contains | 391 | 263 | 120 | – | – | – | – | – | – | – |
| Cause | 7700 | 7211 | 3278 | – | – | 295 | 847 | – | 3325 | 281 |
| Producer | 1336 | 913 | 500 | – | – | – | – | – | – | – |
| Purpose | 9144 | 5220 | 4227 | – | – | – | – | – | 303 | 16,021 |
| Property | 23,354 | 15,732 | 7020 | – | – | 10,825 | 17,213 | – | – | 2672 |
| State | 394 | 237 | 79 | – | – | – | 889 | – | – | – |
| Quality | 1636 | 1221 | 381 | – | – | – | – | – | – | – |
| Manner | 1268 | 3381 | 439 | – | – | – | – | – | 850 | – |
| Place | 832 | 487 | 1159 | – | – | – | – | – | – | 17,246 |
| **Total** | 191,497 | 139,404 | 80,071 | 480,932 | 51,410 | 151,731 | 154,906 | 161,790 | 20,340 | 132,862 * |
| Avg. degree | 3.9 | 2.9 | 3.3 | 11.9 | 4.0 | 6.4 | 21.7 | 36.9 | 3.2 | 3.0 |

* means that additional relation types were considered for computing the total.

## 4. Redundancy in Portuguese LKBs

Open Portuguese LKBs are not only organised in slightly different models. They were also created with different approaches, most of which involve automatic or semi-automatic steps for exploiting available resources, such as dictionaries or encyclopaedias, not only in Portuguese, but also in other languages. Therefore, although they try to cover the whole language, they end up having different granularities and contents, not only in terms of covered relation types, but also of lexical items and relation instances, some of which are less useful for some tasks, or even incorrect. Table 4 shows the number of relation instances grouped by relation type and number of LKBs they were found in.

Table 5 complements Table 4 and gives an idea on the typical knowledge covered by each LKB. More precisely, for each LKB, the included relation instances are grouped into those that are exclusive from the target LKB, those that are in only one more LKB (+1), and those that are in only two more (+2). This table shows, for instance, that ConceptNet is the network with more non-overlapping knowledge. The LKBs extracted from dictionaries contain the lowest proportion of knowledge that is not found in another LKB, but this proportion is still high—≈57% for DA and ≈64% for PAPEL and Wiktionary.

The majority of relation instances found (≈82%) is in only one LKB, ≈13% is in two, ≈3% in three and just ≈1% in four. Only synonymy, and a residual number of antonymy and hypernymy instances, are in six or more LKBs, expectable because those also happened to be the types covered by more LKBs. Our intuition is that the more resources an instance is in, the more likely it is to transmit a consensual, frequent and useful relation. This does not mean, however, that most of the relations found in only one LKB are incorrect or not useful. It only means that the latter set should contain a higher proportion of relations that are either incorrect, very specific or useful only in a more limited domain of application, when compared to the set of relations in more than one LKB. This is confirmed by observed examples, including those in Table 6, which contains relation instances that are in nine to three LKBs. Each redundancy level includes only instances of relation types that were not present in the previous level, or were but with arguments with a different POS.

**Table 4.** Occurrences of the same triples in different resources, per type.

| Relation | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | Total |
|---|---|---|---|---|---|---|---|---|---|---|
| Synonymy | 276,113 | 68,983 | 20,068 | 8773 | 4194 | 2079 | 955 | 361 | 88 | 381,614 |
| Antonymy | 51,179 | 1763 | 534 | 164 | 54 | 9 | 4 | – | – | 53,707 |
| Hypernymy | 281,125 | 27,712 | 4339 | 584 | 89 | 2 | – | – | – | 313,851 |
| Part | 23,431 | 1994 | 151 | 6 | 1 | – | – | – | – | 25,583 |
| Member | 13,294 | 640 | 48 | 3 | – | – | – | – | – | 13,985 |
| Material | 1756 | 159 | 6 | – | – | – | – | – | – | 1921 |
| Contains | 635 | 65 | 3 | – | – | – | – | – | – | 703 |
| Cause | 11,481 | 3127 | 1158 | 432 | – | – | – | – | – | 16,198 |
| Producer | 2216 | 217 | 33 | – | – | – | – | – | – | 2466 |
| Purpose | 31,771 | 1333 | 142 | 13 | – | – | – | – | – | 33,259 |
| Property | 58,374 | 7569 | 870 | 146 | 22 | – | – | – | – | 66,981 |
| State | 1424 | 77 | 7 | – | – | – | – | – | – | 1508 |
| Quality | 1760 | 631 | 72 | – | – | – | – | – | – | 2463 |
| Manner | 4274 | 683 | 98 | 1 | – | – | – | – | – | 5056 |
| Place | 18,848 | 286 | 100 | 1 | – | – | – | – | – | 19,235 |
| **Total** | 777,681<br>(82.9%) | 115,239<br>(12.3%) | 27,629<br>(2.9%) | 10,123<br>(1.1%) | 4360<br>(0.5%) | 2090<br>(0.2%) | 959<br>(0.1%) | 361<br>(0.0%) | 88<br>(0.0%) | 938,530 |

**Table 5.** Proportion of relation instances in each LKB that occur only in this LKB, this and another, and this and two other LKBs.

| | Exclusive | +1 | +2 |
|---|---|---|---|
| **PAPEL** | 121,673 (63.5%) | 69,824 (36.5%) | 26,749 (14.0%) |
| **DA** | 79,010 (56.7%) | 60,394 (43.3%) | 23,792 (17.1%) |
| **Wikt.PT** | 50,881 (63.5%) | 29,190 (36.5%) | 15,418 (19.3%) |
| **TeP** | 400,334 (83.0%) | 80,598 (16.7%) | 28,676 (6.0%) |
| **OT.PT** | 36,019 (70.0%) | 15,391 (30.0%) | 10,718 (20.8%) |
| **OWN.PT** | 129,377 (85.3%) | 22,354 (14.7%) | 7577 (5.0%) |
| **PULO** | 136,223 (87.9%) | 18,683 (12.1%) | 6731 (4.3%) |
| **WN.Br** | 114,616 (70.8%) | 47,174 (29.2%) | 12,320 (7.6%) |
| **Port4Nooj** | 17,581 (86.4%) | 2759 (13.6%) | 1573 (7.7%) |
| **ConceptNet** | 123,037 (92.6%) | 9826 (7.4%) | 6042 (4.5%) |

**Table 6.** Examples of redundant relation instances.

| # | Examples of Relation Instances |
|---|---|
| 9 | *agarrar* synonymOf *pegar* (grab, catch), *apressar* synonymOf *acelerar* (rush, hasten), *punir* synonymOf *castigar* (punish, discipline) |
| 8 | *pedinte* synonymOf *mendigo* (beggar, mendicant), *vulgar* synonymOf *ordinário* (vulgar, ordinary), *porventura* synonym *talvez* (perhaps, possibly) |
| 7 | *fácil* antonymOf *difícil* (easy, hard), *legal* antonymOf *ilegal* (legal, ilegal) |
| 6 | *árvore* hypernymOf *carvalho* (tree, oak), *árvore* hypernymOf *faia* (tree, beech) |
| 5 | *degrau* partOf *escada* (step, stairs), *mítico* propertyOf *mito* (mythical, myth), *tristeza* antonymOf *alegria* (sadness, joy), *somar* antonymOf *subtrair* (add up, subtract) |
| 4 | alterar hypernymOf modificar (change, modify), *investir* causes *investimento* (invest, investment), *feliz* stateOf *felicidade* (happy, happiness), *carta* memberOf *baralho* (card, deck), *fumar* purposeOf *charuto* (smoke, cigar), *habilmente* mannerOf *habilidade* (ably, ability), *dependente* propertyOf *depender* (dependable, depend), *Equador* placeOf *equatoriano* (Ecuador, Ecuadorian) |
| 3 | *impertinente* qualityOf *impertinência* (impertinent, impertinence), *vinho* containedIn *galheta* (wine, cruet), *coqueiro* producerOf *coco* (coconut tree, coconut), *fio* materialOf *meada* (thread, hank), *condução* purposeOf *cano* (conduction, pipe), *força* partOf *robusto* (strength, robust) |

On the other hand, instances that only occur in one LKB are more likely to either be incorrect, due to noise on the automatic process, or to involve very specific meanings, which makes them less useful. Observed examples also confirm this. Some of them are presented in Table 7, which shows a list of relation instances that are in a single LKB, selected randomly for different relation types.

Following the aforementioned intuition—relation instances in more LKBs are more likely to transmit a consensual, frequent and useful relation—, new LKBs were created, based on the redundancy level: one with all the relation instances in all LKBs (*All*) and eight more with the relation instances in at least two to nine LKBs (*Redun2–9*). The resulting LKBs are characterised in Table 8. From those, the largest three (*All*, *Redun2*, *Redun3*) were used to perform the same tasks as the original LKBs, which is reported in the following section. Due to historical reasons, CARTÃO [24], an LKB completely extracted from dictionaries, that combines PAPEL, DA and Wiktionary.PT, was also used in the following experiments. Table 8 also contains information on the size of CARTÃO.

**Table 7.** Examples of relation instances in only one LKB.

*olorado* synonymOf *aromal* (smelt, aromal?), *economicamente* synonymOf *regradamente* (economically, ordely), *saltão* synonymOf *salta-paredes* (locust, wall-jumper?), *despropositado* antonymOf *razoável* (inopportune, reasonable), *em_definitivo* antonymOf *temporariamente* (definitively, temporarily), *crueza* antonymOf *clemência* (crudeness, mercy), *desgarrar* antonymOf *aprochegar* (tear apart, approach?), *despigmentado* propertyOf *perder_cor* (depigmented?, lose_color), *diluviano* propertyOf *aluvião* (diluvial, alluvium), *alfitomancia* purposeOf *farinha* (alphitomancy, flour), *cuidar_dos_pacientes* purposeOf *médico* (take_care_of_the_patients, doctor), *transformar* hypernymOf *descolorir* (transform, decolor), *atitude* hypernymOf *anticomunismo* (attitude, anticomunism), *coisa* hasState *clima* (thing, climate), *lugar-tenente* hasQuality *lugar-tenência* (lieutenant, lieutenancy?), *satanizar* causes *satanização* (demonize, demonization), *causar* causes *causa* (to cause, cause), *pressão* causes *depressão* (pressure, depression), *cobre* containedIn *hemocianina* (copper, hemocyanin), *Abissínia* placeOf *abissínio* (Abyssinia, Abyssinian), *parabolicamente* mannerOf *parábola* (paraborically?, parable), *imunoglobina* materialOf *plasma* (immunoglobulin, plasma), *pessoa* memberOf *lóbi* (person, lobby), *kibibyte* partOf *megabyte*, *caju* producerOf *castanha* (cashew, chestnut)

**Table 8.** Size of the redundancy-based LKBs.

| Redundancy | 1 (All) | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | CARTÃO |
|---|---|---|---|---|---|---|---|---|---|---|
| Lexical items | 202,000 | 58,412 | 24,959 | 13,213 | 7495 | 4196 | 2042 | 761 | 168 | 149,818 |
| Relation instances | 938,846 | 160,749 | 45,510 | 17,981 | 7858 | 3498 | 1408 | 449 | 88 | 327,405 |

## 5. Comparing Portuguese LKBs Indirectly

Due to the time-consuming work required for evaluating the contents of each LKB manually, plus the subjectivity of such a task, the Portuguese LKBs were compared indirectly, when exploited to solve semantic similarity-related tasks, for which datasets, here used as benchmarks, are available. Experiments performed in this comparison cover four different tasks, namely: selecting the most similar word from a small set ($B^2SG$, Section 5.1); computing the semantic similarity between pairs of words (SimLex-999, Section 5.2); selecting the most suitable word, in a set, for a blank in a sentence (cloze questions, Section 5.3); and computing the semantic similarity between pairs of sentences (ASSIN, Section 5.4). Table 9 organises those benchmark tests according to their type.

**Table 9.** Characterization of the benchmark tests.

| | Word Level | Sentence Level |
|---|---|---|
| Multiple choice | $B^2SG$ | Cloze questions |
| Similarity score | SimLex-999 | ASSIN |

### *5.1. Selecting the Most Similar Word from a Small Set*

The $B^2SG$ [25] test is similar to the WordNet-Based Synonymy Test [26], but based on the Portuguese part of BabelNet [13] and partially evaluated by humans. It contains frequent Portuguese nouns and verbs (target), each followed by four candidates, from which only one is related, and is organised in six files: two for synonymy, two for hypernymy, and two for antonymy, respectively, between nouns and for verbs. Table 10 illustrates the $B^2SG$ test with the first line of each file. The correct answer is always the first candidate, followed by three distractors.

**Table 10.** First entries of each file of the $B^2SG$ test.

| Relation | Target | Candidates | | | |
|---|---|---|---|---|---|
| Synonym (noun) | concorrente | **competidor** * | cortina | amurada | carmesim |
| Synonym (verb) | trancar | **barrar** | aviar | alienar | progredir |
| Hypernym (noun) | matemática | **ciência** | célula | pulseira | libertação |
| Hypernym (verb) | segar | **ceifar** | anexar | concentrar | desembrulhar |
| Antonym (noun) | esquerda | **direita** | repressão | sétimo | diácono |
| Antonym (verb) | trancar | **abrir** | praticar | dragar | empenhar |

* Correct answers in bold.

Although created for evaluating less structured resources, such as distributional thesauri, we analysed how many correct relations of this test are covered by the Portuguese LKBs. Furthermore, for the uncovered instances, the correct alternative was guessed from the top-ranked candidate, after running the Personalized PageRank [27] algorithm in each LKB, for 30 iterations, using the target word as context.

Table 11 presents the number of covered (In) and guessed (Guess) relation instances for each LKB. Coverage numbers highlight known limitations of some LKBs. For instance, antonymy relations extracted from dictionaries are mostly between adjectives; synset-based thesauri do not cover hypernymy; only the wordnet-based LKBs cover hypernymy between verbs and WN.Br covers only verbs. However, for this specific test, some limitations could be minimized by exploiting the structure of the LKB. As expected, the highest coverage and proportion of guessed relations is obtained for the *All* LKB, for which 97.4% of the instances are guessed. It is followed by OWN-PT on both coverage and guesses, except for the guesses of hypernymy and antonymy between nouns. In the former, CARTÃO gets the second highest number, followed really close by *Redun2*, which gets the second highest number of guesses of antonymy relations between nouns. However, we suspect that these numbers are positively biased towards OWN-PT because it is currently integrated in BabelNet.

### *5.2. Computing the Similarity between Word Pairs*

SimLex-999 [28] is a recent benchmark for assessing methods for computing semantic similarity. It contains 999 pairs of words, with the same POS, and their similarity score, given by human subjects who followed strict guidelines to differentiate between similarity and relatedness. No multiword expressions nor named entities are included. This dataset was originally made available for English but has been translated to other languages. The Portuguese adaptation was originally made to assess the distributional models of Portuguese words [29] and is available online (http://metashare.metanet4u.eu/ or https://github.com/nlx-group/lx-dsemvectors/ (October 2017)). Table 12 shows two adjectives, two nouns and two verbs of the Portuguese SimLex-999.

In order to exploit the LKBs in this task, two different algorithms were applied to compute the similarity between the words of each pair, namely:

- Similarity of the adjacencies of each word in the LKB, using measures such as the Jaccard coefficient (Adj-Jac, Equation (1)) or the cosine similarity (Adj-Cos, Equation (2)):

$$Adj\text{-}Jac(w_1, w_2) = \frac{|adjacencies(w_1) \cap adjacencies(w_2)|}{|adjacencies(w_1) \cup adjacencies(w_2)|}, \quad (1)$$

$$Adj\text{-}Cos(w_1, w_2) = \frac{|adjacencies(w_1) \cap adjacencies(w_2)|}{\sqrt{|adjacencies(w_1)|} + \sqrt{|adjacencies(w_2)|}}. \quad (2)$$

- PageRank vectors, inspired by Pilehvar et al. [30]. For each word of a pair, Personalized PageRank was first run in the target LKB, for 30 iterations, using the word as context; a vector was then created with the resulting rank of each other word of the LKB in each position. Finally, the similarity between the vectors for each word was computed, using: the Jaccard coefficient between the sets of words in these vectors (PR-Jac) or the cosine of the vectors (PR-CosV). Given the large vector sizes, vectors were trimmed to the top$-N$ ranked words. Different sizes $N$ were tested, from 50 to 3200.

**Table 11.** Relation instances in and guessed from the B$^2$SG test. Highest and second highest numbers are in bold.

| | LKB | Synon (1171) | | Hypern (758) | | Anton (145) | |
|---|---|---|---|---|---|---|---|
| | | In | Guess | In | Guess | In | Guess |
| Nouns | **PAPEL** | 28.9% | 84.0% | 5.0% | 78.2% | 0.0% | 63.4% |
| | **DA** | 16.5% | 71.7% | 4.6% | 66.1% | 0.0% | 59.3% |
| | **Wikt.PT** | 16.6% | 66.2% | 5.0% | 67.9% | 8.3% | 74.5% |
| | **OWN-PT** | **62.8%** | 80.1% | **59.0%** | 82.5% | **60.0%** | 82.8% |
| | **PULO** | 13.2% | 30.2% | 18.3% | 38.8% | 27.6% | 49.7% |
| | **TeP** | 33.2% | 63.9% | 0.0% | 52.9% | 32.4% | 69.7% |
| | **OT.PT** | 17.7% | 35.0% | 0.0% | 30.2% | 0.0% | 31.7% |
| | **Port4Nooj** | 0.1% | 17.1% | 0.3% | 20.4% | 0.0% | 26.2% |
| | **WN.Br** | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% |
| | **ConceptNet** | 24.3% | 60.2% | 0.1% | 54.2% | 11.7% | 65.5% |
| | **CARTÃO** | 36.8% | 89.0% | 10.4% | **86.0%** | 8.3% | 79.3% |
| | **Redun3** | 33.2% | 70.2% | 5.3% | 61.6% | 20.0% | 75.2% |
| | **Redun2** | 50.4% | **89.3%** | 20.2% | 85.5% | 41.4% | **86.9%** |
| | **All** | **81.5%** | **99.0%** | **64.9%** | **95.6%** | **71.0%** | **97.2%** |
| | **LKB** | **Synon (435)** | | **Hypern (198)** | | **Anton (167)** | |
| | | **In** | **Guess** | **In** | **Guess** | **In** | **Guess** |
| Verbs | **PAPEL** | 37.0% | 82.8% | 0.0% | 78.8% | 0.0% | 46.7% |
| | **DA** | 24.8% | 74.0% | 0.0% | 71.7% | 0.0% | 37.7% |
| | **Wikt.PT** | 18.9% | 60.9% | 0.0% | 55.1% | 3.6% | 52.7% |
| | **OWN-PT** | **84.8%** | **95.4%** | **88.4%** | **97.5%** | **86.8%** | **97.6%** |
| | **PULO** | 24.4% | 41.6% | 24.7% | 46.0% | 40.1% | 59.9% |
| | **TeP** | 53.1% | 76.8% | 0.0% | 69.7% | 47.9% | 79.0% |
| | **OT.PT** | 25.1% | 43.0% | 0.0% | 35.4% | 0.0% | 24.6% |
| | **Port4Nooj** | 0.0% | 17.7% | 0.0% | 19.2% | 0.0% | 22.8% |
| | **WN.Br** | 47.6% | 73.1% | 32.3% | 74.2% | 0.0% | 44.9% |
| | **ConceptNet** | 32.0% | 62.6% | 5.1% | 54.0% | 18.6% | 70.1% |
| | **CARTÃO** | 43.7% | 86.4% | 0.0% | 82.3% | 3.6% | 51.5% |
| | **Redun3** | 55.2% | 84.4% | 12.6% | 79.3% | 29.9% | 68.9% |
| | **Redun2** | 66.2% | 89.0% | 44.4% | 88.9% | 59.3% | 85.6% |
| | **All** | **93.1%** | **98.2%** | **91.9%** | **99.0%** | **94.6%** | **97.6%** |

**Table 12.** First two adjectives, nouns and verbs of the Portuguese SimLex-999.

| Word 1 | Word 2 | POS | Similarity |
|---|---|---|---|
| *esperto* (smart) | *inteligente* (intelligent) | A | 8.33 |
| *sujo* (dirty) | *estreito* (narrow) | A | 0.00 |
| *esposa* (wife) | *marido* (husband) | N | 5.00 |
| *livro* (book) | *texto* (text) | N | 5.00 |
| *ir* (go) | *vir* (come) | V | 3.33 |
| *levar* (take) | *roubar* (steal) | V | 6.67 |

In addition, since SimLex-999 is a similarity test, the previous methods were tested using all the relations of each LKB, or only synonymy and hypernymy relations, which are more connected with this phenomena.

The obtained results were evaluated with the Spearman correlation ($\rho$) between the similarities in SimLex-999 and the similarities computed from each of the previous methods in each LKB. Table 13 shows the best results for each combination of method, relations used, and LKB, as well as different methods for the LKB with the best results (*All*).

Results show that LKBs extracted from dictionaries have better results with PageRank-based algorithms, using all relations. This also includes CARTÃO, which we recall combines relations extracted from three dictionaries. On the other hand, LKBs extracted from wordnets have better results with adjacency-based algorithms, using only synonymy and hypernymy relations. It should be noted that there are clear advantages on using the adjacency-based algorithms, which, because of their lower time complexity, take much less to compute the similarity scores, especially in larger LKBs. The best results are clearly obtained with the combination of all LKBs, using different configurations (0.56–0.61). The original LKB with the best performance is PAPEL (0.49), which performed slightly better than *Redun2* (0.48), but lower than CARTÃO (0.53), which got second place overall. PAPEL was followed by OWN-PT (0.44) and Wiktionary.PT (0.42), both better than *Redun3* (0.44).

**Table 13.** Selection of results for the SimLex-999 test.

| LKB | Relations | Algorithm | $\rho$ |
|---|---|---|---|
| **PAPEL** | All | PR-Jac$_{800}$ | 0.49 |
| **DA** | All | PR-Jac$_{400}$ | 0.38 |
| **Wikt.PT** | All | PR-Jac$_{1600}$ | 0.42 |
| **OWN-PT** | Syn + Hyp | Adj-Cos | 0.44 |
| **PULO** | Syn + Hyp | Adj-Cos | 0.29 |
| **TeP** | Syn + Hyp | Adj-Jac | 0.36 |
| **OT.PT** | Syn + Hyp | Adj-Cos | 0.34 |
| **Port4Nooj** | All | Adj-Jac | 0.19 |
| **WN.Br** | Syn + Hyper | Adj-Jac | 0.04 |
| **ConceptNet** | Syn + Hyp | Adj-Jac | 0.43 |
| **CARTÃO** | All | PR-CosV$_{1600}$ | 0.53 |
| **Redun3** | Syn + Hyper | Adj-Jac | 0.44 |
| **Redun2** | Syn + Hyper | PR-Jac$_{50}$ | 0.49 |
| **All** | Syn + Hyper | PR-CosV$_{50}$ | 0.57 |
| **All** | Syn + Hyper | PR-CosV$_{100}$ | 0.59 |
| **All** | Syn + Hyper | PR-CosV$_{200}$ | **0.61** |
| **All** | Syn + Hyper | PR-CosV$_{400}$ | **0.61** |
| **All** | Syn + Hyper | PR-CosV$_{800}$ | **0.61** |
| **All** | Syn + Hyper | PR-CosV$_{1600}$ | 0.60 |
| **All** | Syn + Hyper | PR-CosV$_{3200}$ | 0.60 |
| **All** | Syn + Hyper | Adj-Cos | 0.58 |
| **All** | Syn + Hyper | Adj-Jac | 0.57 |
| **All** | All | PR-CosV$_{400}$ | 0.56 |

Although the top result is obtained with a PageRank-based algorithm, adjacency-based similarity is close, and even higher for some LKBs. It should thus be seen as a valuable alternative, especially because PageRank-based algorithms are either time (complexity of running PageRank) or memory-expensive (ranks can be pre-computed, but large matrices are required). As for the size of the vectors, there is no clear trend, except that the best result is never obtained with the largest size tested (3200). Further discussion of the best methods is out of the scope of this paper.

Although languages are different and so are the available resources, a final word should be given on the comparison of these results with the top state-of-the-art results for English, as reported in the ACL Wiki (https://aclweb.org/aclwiki/SimLex-999_(State_of_the_art) (October 2017)). By combining distributional vectors with knowledge from Princeton WordNet, a Spearman coefficient of 0.642

was obtained for the English SimLex-999 [31], which is not very far from the results of our best configuration (0.61). In the future, we will study the impact of combining the LKB-based approach with distributional vectors.

### *5.3. Answering Cloze Questions*

Open domain cloze questions have been generated in the scope of REAP.PT [32], an assisted language learning tutoring system for European Portuguese. Those consist of sentences with a blank, to be filled with a word from a shuffled list of candidates, of which only one is correct and the other are distractors. Some of the Portuguese LKBs have previously been exploited [33] to answer a set of 3890 of those questions, provided by the researchers involved in the REAP.PT project. Table 14 illustrates the contents of this dataset with the first two questions and the respective set of candidate words, with the correct answer in bold.

**Table 14.** First two cloze questions of the dataset used.

| # | Sentence | Candidates | |
|---|---|---|---|
| 1 | *A instalação de «superpostos» nas entradas e saídas dos grandes __________ urbanos levanta, por outro lado, algumas dúvidas à Anarec.*<br>(The installation of «overlays» at the entrances and exits of the major urban __________ raises some doubts to Anarec.) | ***centros***<br>*mecanismos*<br>*inquéritos*<br>*indivíduos* | **centers**<br>mechanisms<br>surveys<br>individuals |
| 2 | *O artista __________ uma verdadeira obra de arte.*<br>(The artist __________ a real work of art.) | ***criou***<br>*emigrou*<br>*requereu*<br>*atribuiu* | **created**<br>emigrated<br>required<br>attributed |

The experiment reported here used the same dataset, this time answered with each of the LKBs explored in this work. The selection method was similar to the one used for the B$^2$SG test (Section 5.1): for each sentence, answers were guessed from the top-ranked candidate, after running Personalized PageRank, this time using the lemmas of all the open-class words as context. For instance, for sentence #2, the words *artista, verdadeiro, obra and arte* were used.

Table 15 shows the accuracy in the selection of the correct answer, using each LKB, and with a baseline that selects the most frequent alternative, based on the frequency lists of the AC/DC corpora [23]. Results are shown as a total, and also organised according to the POS of the correct word to fill the blank. When no alternative was covered by the LKB, the answer would contain all the alternatives (25% correct).

Although all LKBs performed better than random chance (25%), this revealed to be a challenging task. WN.Br was just slightly higher than this number, possibly because it only covers verbs. Other LKBs were not much higher than the frequency baseline, which improved the random chance for nouns and verbs, but apparently did not make much difference for adjectives and adverbs. The highest rate of correct answers ($\approx$40%) was obtained with CARTÃO, with no significant differences when compared to the result obtained with the *All* LKB. On the one hand, CARTÃO got the highest proportion of correct answers when the blank was to be filled with a verb ($\approx$37%) or an adjective ($\approx$36%), while the *All* LKB got the highest proportion for nouns ($\approx$50%). For adverbs, this proportion is not significantly different than the random chance. Curiously, the highest result is for Port4Nooj ($\approx$30%). If using a smaller LKB is desired, PAPEL ($\approx$191,000 relation instances) or *Redun2* ($\approx$145,000) answer $\approx$38% of the questions correctly.

**Table 15.** Accuracy for answering cloze questions.

| | **Noun (1769)** | **Verb (1077)** | **Adj (809)** | **Adv (235)** | **Total (3890)** |
|---|---|---|---|---|---|
| Baseline | 34.43% | 32.82% | 25.28% | 25.11% | 31.52% |
| **PAPEL** | **44.19%** | **36.63%** | **33.47%** | 22.13% | **38.53%** |
| **DA** | 39.49% | 32.87% | 30.01% | 24.36% | 34.77% |
| **Wikt.PT** | 39.85% | 35.65% | 31.15% | 27.45% | 36.13% |
| **OpenWN-PT** | 38.72% | 31.78% | 25.28% | 26.17% | 33.25% |
| **PULO** | 40.77% | 31.43% | 22.16% | 23.19% | 33.25% |
| **TeP** | 41.72% | 30.71% | 31.49% | 25.00% | 35.53% |
| **OpenThes.PT** | 35.01% | 26.51% | 26.21% | 25.43% | 30.24% |
| **Port4Nooj** | 37.11% | 26.86% | 27.97% | **29.89%** | 31.93% |
| **WN.Br** | 24.82% | 29.55% | 24.44% | 25.11% | 26.07% |
| **ConceptNet** | 37.00% | 34.42% | 32.55% | 27.73% | 34.79% |
| **CARTÃO** | 46.78% | **36.86%** | **36.46%** | 27.77% | **40.74%** |
| **Redun3** | 40.54% | 32.61% | 28.83% | 27.70% | 35.13% |
| **Redun2** | 45.00% | 34.03% | 30.44% | **28.09%** | 37.90% |
| **All** | **49.90%** | 33.05% | 34.98% | 26.81% | **40.72%** |

### 5.4. Textual Similarity and Entailment

The ASSIN shared task targeted semantic similarity and textual entailment in Portuguese [34]. Its training data comprises 6000 sentence pairs ($t$, $h$), half of which in Brazilian Portuguese (PTBR) and the other half in European Portuguese (PTPT). Test data comprises 4000 pairs, 2000 in each variant. Data is available in the task's website (http://nilc.icmc.usp.br/assin/ (April 2017)), together with the gold annotations of the test data and evaluation scripts. Similarity values range from 1 (completely different sentences, on different subjects) to 5 ($t$ and $h$ mean essentially the same). Entailment can have one of the following values: *Paraphrase*, *Entailment* or *None*. Table 16 shows a selection of sentence pairs in the ASSIN training collection

**Table 16.** Selected examples from the ASSIN training collection, for EurOpean Portuguese (PTPT) and for Brazlian Portuguese (PTBR).

| **Variant** | **Id** | | **Pair** | **Sim** | **Entailment** |
|---|---|---|---|---|---|
| PTPT | 2675 | **t** | *O Chelsea só conseguiu reagir no final da primeira parte.*<br>(Chelsea were only able to react at the end of the first half) | 1.25 | None |
| | | **h** | *Não podemos aceitar outra primeira parte como essa.*<br>(We can not accept another first half like this.) | | |
| PTBR | 319 | **t** | *Cerca de 10% da Grande Muralha da China já desapareceu.*<br>(About 10% of the Great Wall of China has disappeared.) | 2.50 | None |
| | | **h** | *Em 2006, a China estabeleceu regulamentos para a proteção da Grande Muralha.*<br>(In 2006, China established regulations for the protection of the Great Wall.) | | |
| PTPT | 315 | **t** | *Todos que ficaram feridos e os mortos foram levados ao hospital.*<br>(All the wounded and the dead were taken to the hospital.) | 3.00 | None |
| | | **h** | *Além disso, mais de 180 pessoas ficaram feridas.*<br>(In addition, more than 180 people were injured.) | | |
| PTBR | 2982 | **t** | *Maldonado disse ainda que cerca de 125 casas foram afetadas pelo deslizamento.*<br>(Maldonado also said that about 125 homes were affected by the landslide) | 4.00 | Entailment |
| | | **h** | *Segundo Maldonado, mais de 100 casas podem ter sido atingidas.*<br>(According to Maldonado, more than 100 houses may have been hit) | | |
| PTBR | 1282 | **t** | *As multas previstas nos contratos podem atingir, juntas, 23 milhões de reais.*<br>(The penalties set in the contracts may amount to R$ 23 million.) | 5.00 | Paraphrase |
| | | **h** | *Somadas, as multas previstas nos contratos podem chegar a R$ 23 milhões.*<br>(All added up, the penalties set in the contracts may reach R$ 23 million.) | | |

LKBs were exploited to compute similarity according to Equation (3). Briefly, after preprocessing the sentences and computing the cosine of their stems, a bonus ($\gamma$) was added for each additional word from $t$ directly related to a word in $h$ ($\gamma$+ = 0.75) or related to a common word ($\gamma$+ = 0.05):

$$Sim(S_1, S_2) = \frac{|S_1 \cap S_2| + \gamma}{\sqrt{|S_1|}\sqrt{|S_2|}}. \quad (3)$$

A very simple approach was followed for the entailment task. Common words and synonyms were first removed from the longer sentence. If the proportion of remaining words was below $\alpha = 0.1$, the pairs would be classified as a Paraphrase. After this, words from the first sentence in an hypernymy relation with words from the second were also removed. If the proportion of remaining words was below $\beta = 0.45$, the pair would be classified as Entailment. Parameters $\alpha$ and $\beta$ were set after several experiments in the training collection.

Table 17 shows the obtained results for the PTPT and PTBR variants, with each LKB, plus a baseline that does not use an LKB ($\alpha = \beta = 0$), and the best official results of ASSIN. Entailment performance is scored in terms of accuracy and Macro-F1, while similarity resorts to the Pearson correlation and the mean square error (MSE).

**Table 17.** Exploiting LKBs in the ASSIN test set.

| | **PTPT** | | | | **PTBR** | | | |
|---|---|---|---|---|---|---|---|---|
| **Config** | **Entailment** | | **Similarity** | | **Entailment** | | **Similarity** | |
| | **Acc** | **F1** | **Pearson** | **MSE** | **Acc** | **F1** | **Pearson** | **MSE** |
| *Baseline (cosine)* | 74.10% | 0.43 | 0.66 | 0.66 | 78.60% | 0.43 | 0.65 | 0.445 |
| *Best PTPT* | 83.85% | 0.70 | 0.73 | 0.61 | – | – | – | – |
| *Best sim PTBR* | – | – | 0.70 | 0.66 | – | – | 0.70 | 0.38 |
| *Best entail PTBR* | 77.60% | 0.61 | 0.64 | 0.72 | 81.65% | 0.52 | 0.64 | 0.45 |
| **PAPEL** | 74.30% | 0.45 | **0.67** | 0.70 | 78.25% | 0.45 | 0.66 | 0.44 |
| **DA** | 74.10% | 0.44 | **0.67** | 0.69 | **78.50%** | 0.44 | 0.66 | **0.43** |
| **Wikt.PT** | 74.00% | 0.44 | **0.67** | **0.68** | 77.55% | 0.43 | 0.66 | **0.43** |
| **OWN-PT** | 73.80% | 0.45 | **0.67** | 0.71 | 77.30% | 0.43 | 0.66 | **0.43** |
| **PULO** | 74.00% | 0.45 | 0.66 | 0.74 | 76.80% | 0.45 | 0.66 | 0.45 |
| **TeP** | 74.55% | **0.47** | **0.67** | 0.71 | 77.90% | 0.47 | **0.67** | 0.45 |
| **OT.PT** | 74.05% | 0.44 | **0.67** | **0.68** | 78.40% | 0.44 | 0.66 | **0.43** |
| **Port4Nooj** | 73.85% | 0.43 | 0.66 | **0.68** | 78.10% | 0.43 | 0.66 | 0.44 |
| **WN.Br** | 74.20% | 0.45 | 0.66 | 0.71 | 77.50% | 0.44 | 0.66 | 0.45 |
| **ConceptNet** | 74.35% | 0.45 | **0.67** | 0.73 | 77.80% | 0.45 | 0.65 | 0.47 |
| **Redun3** | **74.80%** | **0.47** | **0.67** | 0.73 | 78.00% | 0.46 | **0.67** | 0.46 |
| **Redun2** | 74.15% | **0.47** | **0.67** | 0.73 | 77.55% | **0.48** | 0.66 | 0.44 |
| **All** | 72.95% | **0.47** | 0.66 | 0.86 | 76.00% | **0.48** | 0.65 | 0.46 |

The approach followed in this task was assumedly simplistic. In fact, the performance of using different LKBs does not vary significantly and no strong conclusions can be taken, as the cosine seems to play a greater role. To reach the best performances, LKB features would have to be combined with others, possibly in a supervised approach, where the weights for each feature would be learned during the training phase. This is how most participating systems approached ASSIN, including the best results. Further experiments made with these LKBs with additional features can be found elsewhere [35].

Despite the previous remark, in opposition to the cloze questions, in this case, using the *All* LKBs leads to the low results in most scores, possibly due to the noise in such a large LKB, and also due to the different method applied. Using the redundancy based LKBs would probably be a good option, especially for similarity.

## 6. Conclusions

Ten open Portuguese LKBs were overviewed in this paper, namely PAPEL; relations acquired from Dicionário Aberto and Wiktionary.PT; OpenWordnet-PT; PULO; TeP; OpenThesaurus.PT; semantic relations of Port4Nooj; Wordnet.Br; and the relations between Portuguese words in ConceptNet. An initial comparison focused on size, relation types covered, and redundancy across the LKBs. Despite sharing a similar goal, these LKBs were created by different teams, following different approaches, and there are significant differences in the covered lexical items, relations — more than 80% of all the relations instances are in only one LKB—, their correctness or utility. The creation of new LKBs by combining the existing ones was described and all LKBs were then compared indirectly, when exploited in different computational semantics tasks.

The limitations of some LKBs were confirmed, especially the smaller ones (Port4Nooj, OT.PT), or those focused on a single POS (WN.Br) or relation (OT.PT). Except for the expected impact of those limitations, obtained results are positive for every LKB, especially in the word-based similarity tests. However, experiments suggest that using all the available relation instances generally leads to the best results. Some of these LKBs were recently used to answer other word similarity and relatedness tests [36] and, despite different results for different tests, the claim that combining several LKBs leads to better results still holds.

This comparison should not be seen as complete and further analysis is needed for stronger conclusions. Due to the large size of the LKB with all relation instances, in some cases, it might be worth using an LKB containing only relations in two or three LKBs. In the performed experiments, the negative impact of the latter solution on performance is higher for algorithms based on the structure of the network, such as PageRank, and not so much on approaches that do not go one level further than the direct adjacencies. This happens because PageRank exploits every link in the network structure, some of which are not redundant and thus missing from the redundancy-based LKBs. Even though the aforementioned conclusions are still valid for the sentence-oriented tests, additional features and more sophisticated approaches would be required for a higher performance (see [35]).

It should be added that all the ten LKBs compared in this work were exploited in the creation of new version of the fuzzy Portuguese wordnet CONTO.PT [37], in order to be released in the future. In CONTO.PT, words are grouped together or related with a confidence measure, computed from the relations in all the exploited LKBs. This way, users may set their own cut-point on confidence and use either a smaller but more reliable LKB or a larger one, though not so reliable. All the redundancy-based LKBs are freely available for anyone to use, from http://ontopt.dei.uc.pt/index.php?sec=download_outros. We aim at using these LKBs in additional tasks, or in the same but focusing on certain aspects, such as the POS. However, a manual intervention might be required for stronger conclusions.

Following the current trend of using distributional models of words in NLP, such as word embeddings, the performance of the LKBs and algorithms used here was recently compared with the performance of some of the previous models for Portuguese [36]. On the one hand, LKBs lead to the highest results when it comes to genuine similarity. On the other hand, they are outperformed by the distributional models when computing relatedness. This is partially explained by the fact that LKBs are more theoretical views of the mental lexicon, while the distribution of words in a corpus models the way language is actually used. We are currently working on the combination of both kinds of models in a single, hopefully better, word similarity function, as others have done for English (e.g., [22,31]). Such a function might be useful for higher-level natural language tasks, such as semantic search systems or conversational agents.

**Conflicts of Interest:** The author declares no conflict of interest.

## References

1. Fellbaum, C. (Ed.) *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*; The MIT Press: Cambridge, MA, USA, 1998.
2. Marrafa, P. Portuguese WordNet: General architecture and internal semantic relations. *DELTA* **2002**, *18*, 131–146.
3. Gonçalo Oliveira, H. Comparing and Combining Portuguese Lexical-Semantic Knowledge Bases. In *Proceedings of 6th Symposium on Languages, Applications and Technologies (SLATE 2017)*; Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, OASICS: Kobe, Japan, 2017; Volume 56, pp. 16:1–16:15.
4. De Paiva, V.; Real, L.; Gonçalo Oliveira, H.; Rademaker, A.; Freitas, C.; Simões, A. An overview of Portuguese Wordnets. In Proceedings of the 8th Global WordNet Conference (GWC'16), Bucharest, Romania, 27–30 January 2016; pp. 74–81.
5. Magnini, B.; Cavaglià, G. Integrating Subject Field Codes into WordNet. In Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000), Athens, Greece, 31 May–2 June 2000; ELRA: Paris, France, 2000; pp. 1413–1418.
6. Shi, L.; Mihalcea, R. Putting Pieces Together: Combining FrameNet, VerbNet and WordNet for Robust Semantic Parsing. In *Proceedings of Computational Linguistics and Intelligent Text Processing (CICLing'05)*; Lecture Notes in Computer Science; Springer: Berlin, Germany, 2005; Volume 3406, pp. 100–111.
7. Gurevych, I.; Eckle-Kohler, J.; Hartmann, S.; Matuschek, M.; Meyer, C.M.; Wirth, C. UBY—A Large-Scale Unified Lexical-Semantic Resource. In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012), Avignon, France, 23–27 April 2012; ACL Press: Avignon, France, 2012; pp. 580–590.
8. Vossen, P. EuroWordNet: A multilingual database for information retrieval. In Proceedings of the DELOS Workshop on Cross-Language Information Retrieval, Zurich, Switzerland, 5–7 March 1997.
9. Pianta, E.; Bentivogli, L.; Girardi, C. MultiWordNet: Developing an aligned multilingual database. In Proceedings of the 1st International Conference on Global WordNet (GWC 2002), Mysore, India, 21–25 January 2002.
10. Bond, F.; Foster, R. Linking and Extending an Open Multilingual Wordnet. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria, 4–9 August 2013; ACL Press: Sofia, Bulgaria, 2013; pp. 1352–1362.
11. Gonzalez-Agirre, A.; Laparra, E.; Rigau, G. Multilingual Central Repository version 3.0. In Proceedings of the 8th International Conference on Language Resources and Evaluation (ELRA), Istanbul, Turkey, 21–27 May 2012; pp. 2525–2529.
12. De Melo, G.; Weikum, G. Towards a Universal Wordnet by Learning from Combined Evidence. In Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009), Hong Kong, China, 2–6 November 2009; ACM: New York, NY, USA, 2009; pp. 513–522.
13. Navigli, R.; Ponzetto, S.P. BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. *Artif. Intell.* **2012**, *193*, 217–250.
14. Downey, D.; Etzioni, O.; Soderland, S. A Probabilistic Model of Redundancy in Information Extraction. In Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI 2005), Edinburgh, Scotland, 30 July–5 August 2005; pp. 1034–1041.
15. Dias-da-Silva, B.C. Wordnet.Br: An exercise of human language technology research. In Proceedings of the 3rd International WordNet Conference (GWC), Jeju Island, Korea, 22–26 January 2006; pp. 301–303.
16. De Paiva, V.; Rademaker, A.; de Melo, G. OpenWordNet-PT: An Open Brazilian WordNet for Reasoning. In Proceedings of the 24th International Conference on Computational Linguistics (COLING), Mumbai, India, 8–15 December 2012.
17. Simões, A.; Guinovart, X.G. Bootstrapping a Portuguese WordNet from Galician, Spanish and English Wordnets. In *Advances in Speech and Language Technologies for Iberian Languages, Proceedings of the 2nd International Conference on IberSPEECH 2014, Las Palmas de Gran Canaria, Spain, 19–22 November 2014*; Lecture Notes in Computer Science; Springer: Berlin, Germany, 2014; Volume 8854, pp. 239–248.

18. Maziero, E.G.; Pardo, T.A.S.; Felippo, A.D.; Dias-da-Silva, B.C. A Base de Dados Lexical e a Interface Web do TeP 2.0—Thesaurus Eletrônico para o Português do Brasil. In Proceedings of the Companion Proceedings of the XIV Brazilian Symposium on Multimedia and the Web, Vila Velha, Brazil, 26–29 October 2008; ACM: New York, NY, USA, 2008; pp. 390–392.
19. Gonçalo Oliveira, H.; Santos, D.; Gomes, P.; Seco, N. PAPEL: A Dictionary-Based Lexical Ontology for Portuguese. In *Proceedings of 8th International Conference on Computational Processing of the Portuguese Language (PROPOR 2008)*; Lecture Notes in Computer Science; Springer: Berlin, Germany, 2008; Volume 5190, pp. 31–40.
20. Simões, A.; Sanromán, Á.I.; Almeida, J.J. Dicionário-Aberto: A Source of Resources for the Portuguese Language Processing. In *Proceedings of 10th International Conference on Computational Processing of the Portuguese Language (PROPOR 2012)*; Lecture Notes in Computer Science; Springer: Berlin, Germany, 2012; Volume 7243, pp. 121–127.
21. Barreiro, A. Port4NooJ: An open source, ontology-driven Portuguese linguistic system with applications in machine translation. In Proceedings of the 2008 International NooJ Conference (NooJ'08), Budapest, Hungaria, 8–10 June 2008; Cambridge Scholars Publishing: Newcastle upon Tyne, UK, 2010.
22. Speer, R.; Chin, J.; Havasi, C. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In Proceedings of the 31st AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; pp. 4444–4451.
23. Santos, D.; Bick, E. Providing Internet access to Portuguese corpora: The AC/DC project. In Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000), Athens, Greece, 31 May–2 June 2000; pp. 205–210.
24. Gonçalo Oliveira, H.; Pérez, L.A.; Costa, H.; Gomes, P. Uma rede léxico-semântica de grandes dimensões para o português, extraída a partir de dicionários electrónicos. *Linguamática* **2011**, *3*, 23–38.
25. Wilkens, R.; Zilio, L.; Ferreira, E.; Villavicencio, A. B2SG: A TOEFL-like Task for Portuguese. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia, 23–28 May 2016; ELRA: Paris, France, 2016.
26. Freitag, D.; Blume, M.; Byrnes, J.; Chow, E.; Kapadia, S.; Rohwer, R.; Wang, Z. New Experiments in Distributional Representations of Synonymy. In Proceedings of the 9th Conference on Computational Natural Language Learning (CONLL '05), Ann Arbor, MI, USA, 29–30 June 2005; ACL Press: Stroudsburg, PA, USA, 2005; pp. 25–32.
27. Agirre, E.; Soroa, A. Personalizing PageRank for word sense disambiguation. In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL'09), Athens, Greece, 30 March–3 April 2009; ACL Press: Stroudsburg, PA, USA, 2009; pp. 33–41.
28. Hill, F.; Reichart, R.; Korhonen, A. Simlex-999: Evaluating Semantic Models with Genuine Similarity Estimation. *Comput. Linguist.* **2015**, *41*, 665–695.
29. Querido, A.; Carvalho, R.; Rodrigues, J.; Garcia, M.; Silva, J.; Correia, C.; Rendeiro, N.; Pereira, R.; Campos, M.; Branco, A. LX-LR4DistSemEval: A collection of language resources for the evaluation of distributional semantic models of Portuguese. *Rev. Assoc. Port. Linguíst.* **2017**, 265–283, doi:10.26334/2183.
30. Pilehvar, M.T.; Jurgens, D.; Navigli, R. Align, Disambiguate and Walk: A Unified Approach for Measuring Semantic Similarity. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013), Sofia, Bulgaria, 4–9 August 2013; Volume 1, pp. 1341–1351.
31. Banjade, R.; Maharjan, N.; Niraula, N.B.; Rus, V.; Gautam, D. Lemon and Tea Are Not Similar: Measuring Word-to-Word Similarity by Combining Different Methods. In *Proceedings of 16th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2015)*; Lecture Notes in Computer Science; Springer: Berlin, Germany, 2015; Volume 9041, Part I, pp. 335–346.
32. Correia, R.; Baptista, J.; Eskenazi, M.; Mamede, N. Automatic generation of cloze question stems. In *Proceedings of 10th International Conference on Computational Processing of the Portuguese Language (PROPOR 2012)*; Lecture Notes in Computer Science; Springer: Berlin, Germany, 2012; Volume 7243, pp. 168–178.
33. Gonçalo Oliveira, H.; Coelho, I.; Gomes, P. Exploiting Portuguese Lexical Knowledge Bases for Answering Open Domain Cloze Questions Automatically. In Proceedings of the 9th Language Resources and Evaluation Conference (LREC 2014), Reykjavik, Iceland, 26–31 May 2014; ELRA: Paris, France, 2014.

34. Fonseca, E.R.; dos Santos, L.B.; Criscuolo, M.; Aluísio, S.M. Visão Geral da Avaliação de Similaridade Semântica e Inferência Textual. *Linguamática* **2016**, *8*, 3–13.
35. Gonçalo Oliveira, H.; Alves, A.O.; Rodrigues, R. Gradually Improving the Computation of Semantic Textual Similarity in Portuguese. In *Progress in Artificial Intelligence, Proceedings of the 18th EPIA Conference on Artificial Intelligence, Porto, Portugal, 5–8 September 2017*; Lecture Notes in Computer Science; Springer: Berlin, Germany, 2017; Volume 10423, pp. 841–854.
36. Gonçalo Oliveira, H. Unsupervised Approaches for Computing Word Similarity in Portuguese. In *Progress in Artificial Intelligence, Proceedings of the 18th Portuguese Conference on Artificial Intelligence (EPIA 2017), Porto, Portugal, 5–8 September 2017*; Springer: Berlin, Germany, 2017.
37. Gonçalo Oliveira, H. CONTO.PT: Groundwork for the Automatic Creation of a Fuzzy Portuguese Wordnet. In *Proceedings of 12th International Conference on Computational Processing of the Portuguese Language (PROPOR 2016)*; Lecture Notes in Computer Science; Springer: Berlin, Germany, 2016; Volume 9727, pp. 283–295.