Comparing and Combining Portuguese Lexical-Semantic Knowledge Bases

Hugo Gonçalo Oliveira

CISUC, Department of Informatics Engineering, University of Coimbra, Coimbra, Portugal
hroliv@dei.uc.pt

Abstract

There are currently several lexical-semantic knowledge bases (LKBs) for Portuguese, developed by different teams and following different approaches. In this paper, the open Portuguese LKBs are briefly analysed, with a focus on size and overlapping contents, and new LKBs are created from their redundant information. Existing and new LKBs are then exploited in the performance of semantic analysis tasks and their performance is compared. Results confirm that, instead of selecting a single LKB to use, it is worth combining all the open Portuguese LKBs.

1998 ACM Subject Classification I.2.7 Natural Language Processing

Keywords and phrases Lexical Knowledge Bases, Portuguese, WordNet, Redundancy, Semantic Similarity

Digital Object Identifier 10.4230/OASIcs.SLATE.2017.16

1 Introduction

Lexical-semantic knowledge bases (LKBs) are computational resources that organize words according to their meaning, typically used in natural language processing (NLP) tasks at the semantic level. Princeton WordNet [11] is the paradigmatic resource of this kind, for English, with a model adapted to many languages, including Portuguese. However, the first Portuguese WordNet [21] was not available to be used by the research community and the first open alternatives were only developed in the last decade.

Several open Portuguese LKBs are currently available, developed by different teams, following different approaches. Due to the difficulties inherent to crafting such a broad resource manually, most LKBs have some degree of automation in their creation process, which increases the chance of noise. Furthermore, none of them is as consensual as WordNet, created manually and with a large community of users, is for English. In fact, while some Portuguese LKBs are not large enough, others have an interesting size but include several incorrect, unfrequent or unuseful relations or lexical items.

In this paper, nine open Portuguese LKBs are characterised in terms of covered lexical items and 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. Besides confirming our intuition that there are advantages in combining different LKBs, this can be seen as the first systematic comparison of the Portuguese LKBs.

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 [11] is for

6th Symposium on Languages, Applications and Technologies (SLATE 2017).
Editors: R. Queirós, M. Pinto, A. Simões, J. P. Leal, and M. J. Varanda; Article No. 16; pp. 16:1–16:15
OASICS Open Access Series in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

English. This includes several wordnets [8] and other simpler LKBs that, in some cases, may replace a wordnet. For many languages, there is generally 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 [20]), as well as alignments with other lexical resources (e.g. FrameNet and VerbNet [29], or Wikipedia and Wiktionary [17]). WordNet is also the "core" of most multilingual wordnets (e.g. EuroWordNet [32], MultiWordNet [25], Open Multilingual WordNet [4]) and of multilingual knowledge bases that cover linguistic and encyclopaedic knowledge (e.g. Universal WordNet [6], BabelNet [24]). Furthermore, authors working on the automatic acquisition of semantic relations from English text often mention their utility for enriching WordNet [18].

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 [10]. 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

Nine open Portuguese LKBs were explored in this work, namely:

- Three wordnets: WordNet.Br [9], OpenWordNet-PT (OWN.PT) [7] and PULO [30];
- Two synset-based thesauri: TeP [22] and OpenThesaurus.PT[1] (OT.PT);
- Three lexical-semantic networks extracted from Portuguese dictionaries: PAPEL [15] and relations from Dicionário Aberto (DA) [31] and Wiktionary.PT[2];
- The semantic relations available in Port4Nooj [3], a set of linguistic resources.

As these resources do not share exactly the same structure, to enable 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 a relation name. For synset-based LKBs, synsets had to be deconstructed. For example, the instance

$$\{\textit{porta, portão}\} \text{ partOf } \{\textit{automóvel, carro, viatura}\}$$

resulted in:

$$(\textit{porta} \text{ synonymOf } \textit{portão}), (\textit{automóvel} \text{ synonymOf } \textit{carro})$$
$$(\textit{automóvel} \text{ synonymOf } \textit{viatura}), (\textit{carro} \text{ synonymOf } \textit{viatura})$$
$$(\textit{porta} \text{ partOf } \textit{automóvel}), (\textit{porta} \text{ partOf } \textit{carro})$$
$$(\textit{porta} \text{ partOf } \textit{viatura}), (\textit{portão} \text{ partOf } \textit{automóvel})$$
$$(\textit{portão} \text{ partOf } \textit{carro}), (\textit{portão} \text{ partOf } \textit{viatura})$$

Adopted relation names were those defined in the project PAPEL [15], which covered the relation types in all the LKBs, though some names had to be normalized. Table 1 characterises each explored LKB according to the number of lexical items – for each part-of-speech (POS) and total distinct (not considering POS) – and relation instances, grouped by their broader type.

[1] `http://paginas.fe.up.pt/~arocha/AED1/0607/trabalhos/thesaurus.txt` (April 2017)
[2] `http://pt.wiktionary.org` (2015 dump)

Table 1 Number of lexical items and triples extracted from each LKB.

Lexical items									
POS	**PAPEL**	**DA**	**Wikt.PT**	**TeP**	**OT.PT**	**OWN.PT**	**PULO**	**WN.Br**	**Port4Nooj**
Nouns	56,660	61,334	30,170	17,244	6,110	32,509	5,149	0	8,109
Verbs	21,585	16,429	8,918	8,343	2,856	3,626	1,573	5,857	3,161
Adjectives	22,561	18,892	9,536	14,979	3,747	4,401	1,316	0	1,055
Adverbs	1,376	3,160	610	1,138	143	1,120	153	0	475
Distinct	94,165	95,188	45,345	40,499	12,782	40,940	7,943	5,857	12,641

Relations									
Synonymy	83,432	52,278	35,330	388,698	51,410	35,597	9,189	88,488	559
Antonymy	388	440	1,263	92,234	–	5,774	2,818	–	–
Hypernymy	49,210	46,079	22,931	–	–	78,854	26,596	73,302	15,303
Part	5,491	4,367	1,574	–	–	14,275	1,146	–	–
Member	6,585	1,057	1,578	–	–	5,153	259	–	–
Material	336	518	192	–	–	958	67	–	–
Contains	391	263	120	–	–	–	–	–	–
Cause	7,700	7,211	3,278	–	–	295	291	–	3,325
Producer	1,336	913	500	–	–	–	–	–	–
Purpose	9,144	5,220	4,227	–	–	–	–	–	303
Property	23,354	15,732	7,020	–	–	10,825	3,327	–	–
State	394	237	79	–	–	–	505	–	–
Quality	1,636	1,221	381	–	–	–	–	–	–
Manner	1,268	3,381	439	–	–	–	–	–	850
Place	832	487	1,159	–	–	–	–	–	–
Total	191,497	139,404	80,071	480,932	51,410	151,731	44,198	161,790	20,340

Table 2 Occurrences of the same triples in different resources, per type.

Relation	1	2	3	4	5	6	7	8
Synonymy	230,030	65,778	17,592	7,506	3,212	1,166	377	76
Antonymy	48,444	1,257	345	96	22	7	–	–
Hypernymy	247,349	25,145	4,050	516	82	2	–	–
Part	22,620	1,883	146	6	1	–	–	–
Member	13,200	638	48	3	–	–	–	–
Material	1,735	159	6	–	–	–	–	–
Contains	635	65	3	–	–	–	–	–
Cause	10,668	3,115	1,158	432	–	–	–	–
Producer	2,216	217	33	–	–	–	–	–
Purpose	15,938	1,276	132	2	–	–	–	–
Property	45,431	6,057	798	76	3	–	–	–
State	1,031	81	6	1	–	–	–	–
Quality	1,760	631	72	–	–	–	–	–
Manner	4,274	683	98	1	–	–	–	–
Place	1,609	286	99	–	–	–	–	–
Total	646,940	107,271	24,586	8,639	3,320	1,175	377	76
	(81.6%)	(13.5%)	(3.1%)	(1.1%)	(0.4%)	(0.1%)	(0.0%)	(0.0%)

Although the LKB with more lexical items is the one extracted 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. 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. WN.Br only covers verbs and relations between them, but the other wordnet-based LKBs cover all four open POS. Besides synonymy, antonymy and hypernymy, they also cover additional relation types (e.g. part, cause, property), but some types are only found in the LKBs extracted from dictionaries.

4 Redundancy in Portuguese LKBs

Despite originally organised in different models, LKBs were created with different approaches, most of which involving automatic or semi-automatic steps. Therefore, although they try to cover the whole language, they end up having different granularities and contents, in terms of covered relations and lexical items, some of which less useful for some tasks, or even incorrect. Table 2 shows the number of relation instances grouped by relation type and number of LKBs they were found in.

The majority of relation instances (≈81%) 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 is confirmed by observed examples, including those in Table 3, which contains relation instances that are in eight to three LKBs. Each redundancy level includes only instances of types that were not present in the previous level, or were but with arguments with a different POS.

Table 3 Examples of redundant relation instances.

#	**Examples of relation instances**
8	*agarrar* synonymOf *pegar* (grab, catch)
	apressar synonymOf *acelerar* (rush, hasten)
	punir synonymOf *castigar* (punish, discipline)
7	*pedinte* synonymOf *mendigo* (beggar, mendicant)
	espesso synonymOf *grosso* (thick)
	porventura synonym *talvez* (perhaps, possibly)
6	*público* antonymOf *privado* (public, private)
	fácil antonymOf *difícil* (easy, hard)
	árvore hypernymOf *carvalho* (tree, oak)
5	*degrau* partOf *escada* (step, stairs)
	sexual propertyOf *sexo* (sexual, sex)
4	*investir* causes *investimento* (invest, investment)
	feliz stateOf *felicidade* (happy, happiness)
	carta memberOf *baralho* (card, deck)
	votar purposeOf *voto* (vote, vote)
	habilmente mannerOf *habilidade* (ably, ability)
	dependente propertyOf *depender* (dependable, depend)
3	alterar hypernymOf afetar (change, affect)
	impertinente qualityOf *impertinência* (impertinent, impertinence)
	vinho containedIn *galheta* (wine, cruet)
	coqueiro producerOf *coco* (coconut tree, coconut)
	fio materialOf *meada* (thread, hank)
	Brasil placeOf *brasileiro* (Brazil, Brazilian)

On the other hand, instances that only occur in one LKB are more likely to either be incorrect, resulting from noise on the automatic process, or to involve very specific meanings, thus less useful. Observed examples also confirm this. Some of them are presented in Table 4, which shows a list of relation instances that are in a single LKB, selected randomly for different relation types.

Following the aforementioned intuition, new LKBs were created, based on the redundancy level: one with all the relation instances in all LKBs (*All*) and seven more with the relation instances in at least two to eight LKBs (*Redun2-8*). The resulting LKBs are characterized in Table 5. From those, the largest three (*All*, *Redun2*, *Redun3*) were used to perform the same tasks as the original LKBs, as reported in the following section.

5 Comparing Portuguese LKBs indirectly

Our first attempt to compare the Portuguese LKBs relied on their extrinsic evaluation, when exploited to solve semantic similarity-related tasks, for which datasets, here used as benchmarks, are available. This section reports this attempt, which covers four different tasks: selecting the most similar word from a small set (Section 5.1); computing the semantic similarity between pairs of words (Section 5.2); selecting the most suitable word for a blank in a sentence (Section 5.3); and computing the semantic similarity between pairs of sentences (Section 5.4). Table 6 organizes the benchmark tests according to their type.

Table 4 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?), coisa `hasState` clima (thing, climate), lugar-tenente `hasQuality` lugar-tenência (lieutenant, lieutenancy?), satanizar `causes` satanização (demonize, demonization), pressão `causes` depressão (pressure, depression), cobre `containedIn` hemocianina (copper, hemocyanin), 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), guarnecer `purposeOf` cacundê (garnish, ?), transformar `hypernymOf` colorir (transform, coloring), atitude `hypernymOf` anticomunismo (attitude, anticomunism), Abissínia `placeOf` abissínio (Abyssinia, Abyssinian), parabolicamente `mannerOf` parábola (paraborically?, parable), imunoglobina `materialOf` plasma (immunoglobulin, plasma), pessoa `memberOf` lobby (person, lobby), kibibyte `partOf` megabyte, caju `producerOf` castanha (cashew, chestnut)

Table 5 Size of the redundancy-based LKBs.

Redundancy	**1** (All)	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Lexical items	178,903	56,565	23,468	11,431	5,557	2,292	764	144
Relation instances	791,182	145,429	38,173	13,587	4,948	1,628	453	76

Table 6 Characterization of the benchmark tests.

	Word level	**Sentence level**
Multiple choice	B^{2}SG	Cloze questions
Similarity score	SimLex-999	ASSIN

Table 7 First entries of each file of the B^{2}SG 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

5.1 Selecting the most similar word from a small set

The B^2SG [33] test is similar to the WordNet-Based Synonymy Test [13], but based on the Portuguese part of BabelNet [24] 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 organized in six files: two for synonymy, two for hypernymy, and two for antonymy, for nouns and for verbs. Table 7 illustrates the B^2SG test with the first line of each file. The correct answer is always the first candidate, followed by three distractors.

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 [1] algorithm in each LKB, for 30 iterations, using the target word as context.

Table 8 presents the number of covered (In) and guessed (Guess) relations for each LKB. Coverage numbers highlight known limitations of some LKBs, e.g.: 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 only covers 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 hypernymy and antonymy between nouns, for which *Redun2* gets the second highest number of guesses. Yet, 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 [19] 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 translation was originally made to assess the LX-DSemVectors [27], word embeddings learned from Portuguese corpora, and is available online[3]. Table 9 shows the first two adjectives, nouns and 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) or the cosine similarity (Adj-Cos);
- PageRank vectors, inspired by Pilehvar et al. [26]. 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 where each position contained the resulting rank of each other word in the LKB. 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 3,200.

[3] http://metashare.metanet4u.eu/ or https://github.com/nlx-group/lx-dsemvectors/ (April 2017)

Table 8 Relation instances in and guessed from the B^{2}SG test. Highest and second highest numbers are in bold.

	LKB	**Synon (1,171)**		**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%
	Redun2	45.9%	**85.9%**	20.1%	**84.7%**	39.3%	**85.5%**
	Redun3	28.4%	66.6%	5.3%	59.4%	13.8%	73.8%
	All	**80.7%**	**97.6%**	**64.6%**	**95.8%**	**71.0%**	**97.9%**
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%
	Redun2	63.4%	88.0%	43.9%	86.9%	56.9%	84.4%
	Redun3	53.6%	82.1%	9.6%	79.8%	25.7%	65.9%
	All	**92.9%**	**98.4%**	**91.9%**	**99.0%**	**94.6%**	**97.6%**

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 between the similarities in SimLex-999 and the similarities computed from each of the previous methods in each LKB. Table 10 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, while LKBs extracted from wordnets have better results with adjacency-based algorithms, using only synonymy and hypernymy relations. The best results are clearly obtained with the combination of all LKBs, using different configurations (0.56–0.60). The original LKB with the best performance is PAPEL (0.49), which performed slightly better than *Redun2* (0.48). PAPEL was followed by OWN-PT (0.44) and Wiktionary.PT (0.42), both better than *Redun3*.

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

Table 9 First two adjectives, nouns and verbs of the Portuguese SimLex-999.

Word 1	Word 2	POS	Similarity
velho	novo	A	0.00
esperto	inteligente	A	8.33
esposa	marido	N	5.00
livro	texto	N	5.00
ir	vir	V	3.33
levar	roubar	V	6.67

are required). As for the size of the vectors, there is no clear trend, except that the best result is never obtained with the larger sizes tested (1,600 and 3,200). 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 our results with the top state-of-the-art results for English, as reported in the ACL Wiki[4]. By combining distributional vectors with knowledge from Princeton WordNet, a Spearman coefficient of 0.642 was obtained for the English SimLex-999 [2], which is not very far from the results of our best configuration (0.60).

5.3 Answering Cloze Questions

Open domain cloze questions have been generated in the scope of REAP.PT [5], 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 [14] to answer a set of 3,890 of those questions, provided by the researchers involved in the REAP.PT project. Table 11 illustrates the contents of this dataset with the first two questions and the respective set of candidate words, with the correct answer in bold.

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 question, answers were guessed from the top-ranked candidate, after running Personalized PageRank, this time using all the open-class words as context.

Table 12 shows the accuracy of selecting the correct answer, for each LKB, and with a baseline that selects the most frequent alternative, based on the frequency lists of the AC/DC corpora [28]. 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 around the frequency baseline and the highest rate of correct answers (≈40%) was obtained with the *All* LKB. If using such a large LKB (≈791,000 relation instances) is not an option, PAPEL (≈191,000) or *Redun2* (≈145,000) answer ≈38% of the questions correctly.

[4] `https://www.aclweb.org/aclwiki/index.php?title=SimLex-999_(State_of_the_art)` (April 2017)

Table 10 Selection of results for the SimLex-999 test.

LKB	Relations	Algorithm	Spearman
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
Redun2	Syn+Hyper	PR-Jac$_{50}$	0.48
Redun3	Syn+Hyper	Adj-Jac	0.41
All	Syn+Hyper	PR-CosV$_{400}$	**0.60**
All	Syn+Hyper	PR-CosV$_{50}$	0.56
All	Syn+Hyper	PR-CosV$_{100}$	0.58
All	Syn+Hyper	PR-CosV$_{200}$	0.59
All	Syn+Hyper	PR-CosV$_{800}$	0.59
All	Syn+Hyper	PR-CosV$_{1600}$	0.59
All	Syn+Hyper	PR-CosV$_{3200}$	0.59
All	Syn+Hyper	Adj-Cos	0.57
All	Syn+Hyper	Adj-Jac	0.56
All	All	PR-CosV$_{400}$	0.57

Table 11 First two cloze questions of the dataset used.

Sentence	*A instalação de «superpostos» nas entradas e saídas dos grandes ______ urbanos levanta, por outro lado, algumas dúvidas à Anarec.*			
Candidates	**centros**	mecanismos	inquéritos	indivíduos
Sentence	*O artista ______ uma verdadeira obra de arte.*			
Candidates	**criou**	emigrou	requereu	atribuiu

5.4 Textual Similarity and Entailment

The ASSIN shared task targeted semantic similarity and textual entailment in Portuguese [12]. Its training data comprises 6,000 sentence pairs (t, h), half of which in Brazilian Portuguese (PTBR) and the other half in European Portuguese (PTPT). Test data comprises 4,000 pairs, 2,000 in each variant. Data is available in the task's website[5], 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 the value *Paraphrase*, *Entailment* or *None*. Table 13 shows a selection of sentence pairs in the ASSIN training collection.

LKBs were exploited to compute similarity according to equation 1. After preprocessing the sentences and computing the cosine of their stems, a bonus (γ) was added for each additional word from t directly related to a word in h (γ+=0.75) or related to a common word (γ+=0.05).

$$Sim(S_1, S_2) = \frac{|S_1 \cap S_2| + \gamma}{\sqrt{|S_1|}\sqrt{|S_2|}} \; . \tag{1}$$

[5] http://nilc.icmc.usp.br/assin/ (April 2017)

Table 12 Accuracy for answering cloze questions.

	Accuracy
Baseline (frequency)	32.83%
PAPEL	**38.53%**
DA	34.77%
Wikt.PT	36.13%
OWN-PT	33.25%
PULO	33.25%
TeP	35.53%
OT.PT	30.24%
Port4Nooj	31.93%
WN.Br	26.07%
Redun2	38.05%
Redun3	35.35%
All	**40.57%**

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 α and β were set after several experiments in the training collection.

Table 14 shows the obtained results for the PTPT and PTBR variants, with each LKB, plus a baseline that does not use a 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).

In this task, 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 other, 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.

Despite the previous remark, in opposition to the cloze questions, in this case, using the *All* LKBs leads to the lowest results is most scores, possibly due to the noise in such a large LKB, and also due to the different method applied.

6 Concluding remarks

Open Portuguese LKBs were briefly overviewed in this paper, with a focus on size and redundancy across them. 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, 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.

This comparison confirmed the limitations of some LKBs, especially those with a limited size (Port4Nooj, OT.PT), or the ones 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. This is a preliminary comparison and

Table 13 Selected examples from the ASSIN training collection.

Variant	Id	Pair		Sim	Entailment
PTPT	2675	**t**	*O Chelsea só conseguiu reagir no final da primeira parte.*	1.25	None
		h	*Não podemos aceitar outra primeira parte como essa.*		
PTBR	319	**t**	*Cerca de 10% da Grande Muralha da China já desapareceu.*	2.5	None
		h	*Em 2006, a China estabeleceu regulamentos para a proteção da Grande Muralha.*		
PTPT	315	**t**	*Todos que ficaram feridos e os mortos foram levados ao hospital.*	3.0	None
		h	*Além disso, mais de 180 pessoas ficaram feridas.*		
PTBR	2982	**t**	*Maldonado disse ainda que cerca de 125 casas foram afetadas pelo deslizamento.*	4.0	Entailment
		h	*Segundo Maldonado, mais de 100 casas podem ter sido atingidas.*		
PTBR	1282	**t**	*As multas previstas nos contratos podem atingir, juntas, 23 milhões de reais.*	5.0	Paraphrase
		h	*Somadas, as multas previstas nos contratos podem chegar a R$ 23 milhões.*		

Table 14 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
Redun3	**74.70%**	**0.47**	**0.68**	0.69	78.05%	0.46	**0.67**	0.45
Redun2	74.20%	**0.47**	0.67	0.72	77.65%	0.47	**0.67**	0.44
All	73.00%	**0.47**	0.66	0.69	75.90%	**0.48**	0.65	0.45

further analysis is needed for stronger conclusions. But results suggest that using a LKB with knowledge from all the others is generally the best solution. Due to the large size of this solution, in some cases, it might be worth using a LKB containing only relations in two or three LKBs, depending on the task. With the later solution, the negative impact 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.

All the nine LKBs compared in this work were exploited in the creation of new version of the fuzzy Portuguese wordnet CONTO.PT [16], to be released in the future, and the redundancy-based LKBs are freely available for anyone to use[6]. In the future, we aim at using these LKBs in additional tasks, or in the same but focusing on certain aspects, such as the POS. Yet, a manual evaluation might be required for stronger conclusions. It is also in our plans to compare the performance of some of these LKBs and of the algorithms used here with the performance of models of distributional similarity for Portuguese. Although created from different methods – theoretical views on the mental lexicon *vs* distribution of words in a corpus – models such as word embeddings [23] are a recent trend in many NLP tasks, including computing semantic similarity.

References

1 Eneko Agirre and Aitor Soroa. Personalizing PageRank for word sense disambiguation. In *12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 33–41, 2009.

2 Rajendra Banjade, Nabin Maharjan, Nobal B. Niraula, Vasile Rus, and Dipesh Gautam. Lemon and tea are not similar: Measuring word-to-word similarity by combining different methods. In *16th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing)*, pages 335–346, 2015.

3 Anabela Barreiro. Port4NooJ: an open source, ontology-driven portuguese linguistic system with applications in machine translation. In *2008 International NooJ Conference*, pages 19–47, 2010.

4 Francis Bond and Ryan Foster. Linking and extending an open multilingual Wordnet. In *51st Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 1352–1362, August 2013.

5 Rui Correia, Jorge Baptista, Maxine Eskenazi, and Nuno Mamede. Automatic generation of cloze question stems. In *10th International Conference on Computational Processing of the Portuguese Language (PROPOR)*, pages 168–178, April 2012.

6 Gerard de Melo and Gerhard Weikum. Towards a universal wordnet by learning from combined evidence. In *18th ACM Conference on Information and Knowledge Management (CIKM)*, pages 513–522, 2009.

7 Valeria de Paiva, Alexandre Rademaker, and Gerard de Melo. OpenWordNet-PT: An Open Brazilian WordNet for Reasoning. In *24th International Conference on Computational Linguistics*, pages 353–360, 2012.

[6] Check `http://ontopt.dei.uc.pt/index.php?sec=download_outros`.

8 Valeria de Paiva, Livy Real, Hugo Gonçalo Oliveira, Alexandre Rademaker, Cláudia Freitas, and Alberto Simões. An overview of Portuguese wordnets. In *8th Global WordNet Conference*, pages 74–81, 2016.

9 Bento C. Dias-da-Silva. Wordnet.Br: An exercise of human language technology research. In *3rd International WordNet Conference (GWC)*, pages 301–303, January 2006.

10 Douglas Downey, Oren Etzioni, and Stephen Soderland. A probabilistic model of redundancy in information extraction. In *19th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1034–1041, 2005.

11 Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database.* Language, Speech, and Communication. The MIT Press, 1998.

12 Erick Rocha Fonseca, Leandro Borges dos Santos, Marcelo Criscuolo, and Sandra Maria Aluísio. Visão geral da avaliação de similaridade semântica e inferência textual. *Linguamática*, 8(2):3–13, 2016.

13 Dayne Freitag, Matthias Blume, John Byrnes, Edmond Chow, Sadik Kapadia, Richard Rohwer, and Zhiqiang Wang. New experiments in distributional representations of synonymy. In *9th Conference on Computational Natural Language Learning*, pages 25–32, 2005.

14 Hugo Gonçalo Oliveira, Inês Coelho, and Paulo Gomes. Exploiting Portuguese lexical knowledge bases for answering open domain cloze questions automatically. In *9th Language Resources and Evaluation Conference (LREC)*, May 2014.

15 Hugo Gonçalo Oliveira, Diana Santos, Paulo Gomes, and Nuno Seco. PAPEL: A dictionary-based lexical ontology for Portuguese. In *8th International Conference on Computational Processing of the Portuguese Language (PROPOR)*, pages 31–40, September 2008.

16 Hugo Gonçalo Oliveira. CONTO.PT: Groundwork for the automatic creation of a fuzzy portuguese wordnet. In *12th International Conference on Computational Processing of the Portuguese Language (PROPOR)*, pages 283–295, July 2016.

17 Iryna Gurevych, Judith Eckle-Kohler, Silvana Hartmann, Michael Matuschek, Christian M. Meyer, and Christian Wirth. UBY – a large-scale unified lexical-semantic resource. In *13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 580–590, April 2012.

18 Marti A. Hearst. Automated discovery of WordNet relations. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*, Language, Speech, and Communication, pages 131–151. The MIT Press, 1998.

19 Felix Hill, Roi Reichart, and Anna Korhonen. Simlex-999: Evaluating semantic models with genuine similarity estimation. *Computational Linguistics*, 41(4):665–695, December 2015.

20 Bernardo Magnini and Gabriela Cavaglià. Integrating subject field codes into WordNet. In *2nd International Conference on Language Resources and Evaluation (LREC)*, pages 1413–1418, 2000.

21 Palmira Marrafa. Portuguese WordNet: general architecture and internal semantic relations. *DELTA*, 18:131–146, 2002.

22 Erick Maziero, Thiago Pardo, Ariani Di Felippo, and Bento Dias-da-Silva. A base de dados lexical e a interface web do TeP 2.0 – Thesaurus Eletrônico para o Português do Brasil. In *VI Workshop em Tecnologia da Informação e Linguagem Humana*, pages 390–392, 2008.

23 Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *Workshop Track of the International Conference on Learning Representations (ICLR)*, 2013.

24 Roberto Navigli and Simone Paolo Ponzetto. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250, 2012.

25 Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. MultiWordNet: developing an aligned multilingual database. In *1st International Conference on Global WordNet*, pages 293–302, 2002.

26 Mohammad Taher Pilehvar, David Jurgens, and Roberto Navigli. Align, disambiguate and walk: A unified approach for measuring semantic similarity. In *51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1341–1351, 2013.

27 João António Rodrigues, António Branco, Steven Neale, and João Ricardo Silva. LX-DSemVectors: Distributional semantics models for Portuguese. In *12th International Conference on the Computational Processing of the Portuguese Language (PROPOR)*, pages 259–270, 2016.

28 Diana Santos and Eckhard Bick. Providing internet access to Portuguese corpora: the AC/DC project. In *2nd International Conference on Language Resources and Evaluation (LREC)*, pages 205–210, 2000.

29 Lei Shi and Rada Mihalcea. Putting pieces together: Combining FrameNet, VerbNet and WordNet for robust semantic parsing. In *Computational Linguistics and Intelligent Text Processing (CICLing)*, pages 100–111, 2005.

30 Alberto Simões and Xavier Gómez Guinovart. Bootstrapping a Portuguese wordnet from Galician, Spanish and English wordnets. In *Advances in Speech and Language Technologies for Iberian Languages*, volume 8854 of *LNCS*, pages 239–248, 2014.

31 Alberto Simões, Álvaro Iriarte Sanromán, and José João Almeida. Dicionário-Aberto: A source of resources for the Portuguese language processing. In *10th International Conference on Computational Processing of the Portuguese Language (PROPOR)*, pages 121–127, April 2012.

32 Piek Vossen. EuroWordNet: a multilingual database for information retrieval. In *DELOS workshop on Cross-Language Information Retrieval*, 1997.

33 Rodrigo Wilkens, Leonardo Zilio, Eduardo Ferreira, and Aline Villavicencio. B2SG: a TOEFL-like task for Portuguese. In *10th International Conference on Language Resources and Evaluation (LREC)*, pages 3659–3662, May 2016.