CONTO.PT: Groundwork for the Automatic Creation of a Fuzzy Portuguese Wordnet

Hugo Gonçalo Oliveira(✉)

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

Abstract. There are several lexical resources available for the computational processing of Portuguese, organised differently and created by different people with different approaches and limitations. This paper presents the first experiments towards the exploitation of seven of those resources in the automatic creation of a large wordnet, where numerical scores are assigned to the inclusion of words in synsets and to the connection of synsets by semantic relations. Experiments confirm that a large wordnet can indeed be created and, to some extent, computed scores can be used as a confidence measure, which will enable the users to select only a portion of the resource, depending on the needs of their application on quantity and quality of lexical-semantic knowledge.

Keywords: Wordnet · Semantic relations · Confidence · Redundancy · Fuzzy

1 Introduction

Wordnets are lexical-semantic knowledge bases, modelled after Princeton WordNet (PWN) [1]. They group synonyms in synsets, which represent concepts by their possible lexicalisations. Together with the synset glosses, different types of semantic relation, including hypernym and meronymy, are established between synsets and help to describe their meaning. As the same meaning might be transmitted by different words, the same word might be in more than one synset, one for each of its senses.

Due to its machine-friendly structure, wordnet became the standard model of a lexical knowledge base. We have seen the creation of wordnets for many languages, including Portuguese [2], though none is as consensual as PWN is for English. Given the overwhelming task of populating a wordnet from scratch, the open Portuguese wordnets are created automatically or semi-automatically, and rely heavily on the contents of other lexical resources, including wordnets of other languages. On the one hand, automatic processes enable a faster creation but, at the same time, existing noise leads to less reliable resources.

In order to tackle existing limitations, we aim go further on leveraging the advantages of automatic approaches, and to give the users some control on coverage and reliability, depending on their needs. We believe in the potential of

J. Silva et al. (Eds.): PROPOR 2016, LNAI 9727, pp. 283–295, 2016.
DOI: 10.1007/978-3-319-41552-9_29

redundant information across open Portuguese lexical-semantic resources, which should enable the creation of a new broad-coverage wordnet where confidence degrees are assigned to the decisions taken, including the membership of words in synsets or the connection of two synsets by a semantic relation. This should enable users to select their own confidence cut-points, which will set either large but less reliable or smaller and more reliable wordnets. The result can be seen as a fuzzy wordnet, an idea that is not completely new (see [3]), but has not been much explored. Moreover, the fuzzy representation is less artificial, as we know that word senses are not discrete [4], but complex and overlapping structures, so their representation as crisp objects does not reflect the human language.

This paper presents the first experiments towards the creation of a fuzzy Portuguese wordnet. Next section overviews the current Portuguese wordnet initiatives. Resources exploited in this work are then enumerated, and their contents and redundancy analysed. After that, the proposed approach for discovering fuzzy synsets and fuzzy semantic connections is described, together with some results and their evaluation. It follows the steps of ECO [5] – extraction, clustering and ontologising –, an abstract model tailored for the automatic creation of Onto.PT, one of the open Portuguese wordnets, but flexible enough to the creation of other resources of the same kind. This is also why this new wordnet is baptised as CONTO.PT – as in *Confidence-enriched* Onto.PT. The paper ends with the first conclusions of this approach and some lines for further work.

2 Portuguese Wordnets

There are at least six Portuguese lexical-semantic knowledge bases structured according to the wordnet model [2], created by independent teams, following different approaches, and with different licenses and usage restrictions. WordNet.PT Global [6] is the most recent instantiation of the first Portuguese wordnet, in development since 1998. It is essentially handcrafted and created from scratch, for Portuguese, it can be browsed online, but it is not available for download. WordNet.Br is a wordnet project for Brazilian Portuguese where synsets and antonymy relations were first manually produced, based on dictionaries and corpora, and released under the name TeP [7]. Synsets were then manually aligned with PWN and semantic relations between Portuguese synsets with English equivalents were inherited [8]. To our knowledge, this part is not publicly available. MultiWordNet.PT[1] is a Portuguese wordnet with synsets derived from the translation of PWN synsets. It can be browsed online and used under the payment of a license.

Besides the previous, there are four open Portuguese wordnets. Onto.PT [5] is created in a completely automatic fashion – both synset boundaries and the attachment of semantic relations are learned from the exploitation of available lexical semantic resources, without any human supervision. Its development follows ECO, a three-step approach to integrate words and relations from different sources: (i) relation extraction between words; (ii) synset discovery from

[1] See http://mwnpt.di.fc.ul.pt/.

the synonymy relations; (iii) mapping of words in remaining relations to discovered synsets. OpenWordNet-PT [9] was originally developed as a syntactic projection of the Universal WordNet [10] for Portuguese. Its development is thus based on the translation of lexical information in PWN, across multiple languages of Wikipedia, open dictionaries, and also some information from corpora. It is aligned to PWN and a manual curation process is currently undergoing. PULO [11] is based on the probabilistic translation of open wordnets of other languages, with special focus to those included in the MCR project [12], where wordnets of the Iberian languages are aligned to PWN. UfesWN [13] is another Portuguese wordnet, based on the automatic translation of PWN.

With more than 168 k lexical items, 248 k word senses, 117 k synsets, and 340 k relation instances, Onto.PT is the largest Portuguese wordnet [2], which additionally covers a broad range of relation types. On the other hand, it is not aligned to PWN nor any other wordnet and it is far from being 100 % reliable. In a manual evaluation [5], 74 % of synsets were labelled as correct, in 18 % there was no agreement between two judges, and the remaining had at least one incorrect word. Moreover, considering that relations between incorrect synsets are also wrong, between 78 %–82 % were labelled as correct. This highlights the need for incorporating confidence information in large automatically-created wordnets, such as Onto.PT, which may allow users to, depending on their needs, define their coverage *vs* reliability trade-off.

3 Redundancy in Portuguese Lexical-Semantic Resources

This section overviews the contents of the lexical-semantic resources exploited in the reported work and analyses their redundancy, which can be useful for the computation of confidence measures, as shown in the following section.

3.1 Open Portuguese Lexical-Semantic Resources Used

Seven Portuguese lexical-semantic resources are exploited. All of them, listed here, are freely available for download:

- Semantic relation instances of the network PAPEL [14], extracted automatically from a commercial Portuguese dictionary;
- Additional semantic relation instances extracted from **two** dictionaries – Dicionário Aberto (DA) [15] and Wiktionary.PT[2] (Wikt.PT) – using the same grammars as PAPEL, and included in the network CARTÃO [16];
- Synonymy and antonymy instances from **two** handcrafted synset-based thesauri: TeP 2.0 [17] and OpenThesaurus.PT[3] (OT.PT);
- Semantic relation instances acquired from **two** open Portuguese wordnets: OpenWordNet-PT (OWN.PT) [9] and PULO [11].

[2] http://pt.wiktionary.org.
[3] http://paginas.fe.up.pt/~arocha/AED1/0607/trabalhos/thesaurus.txt.

All the obtained lexical-semantic information was converted to a suitable input format for the second and third steps of ECO – term-based triples *(a related-to b)*, where words *a* and *b* are connected by a predicate (*related-to*) that is the name of a semantic relation. For that purpose, thesauri and wordnets synsets had to be deconstructed. For instance, a part-of relation between the synsets {*porta, portão*} and {*automóvel, carro, viatura*} would result in the triples: (*porta* synonym-of *portão*), (*automóvel* synonym-of *carro*), (*automóvel* synonym-of *viatura*), (*carro* synonym-of *viatura*), (*portão* part-of *automóvel*), (*porta* part-of *carro*), (*porta* part-of *viatura*), (*porto* part-of *automóvel*), (*portão* part-of *carro*), (*portã* part-of *viatura*). Relation types used were those covered by PAPEL, with a minor extension to include wordnet relations not extracted from dictionaries, such as hypernymy between verbs (`hiperonimoAccaoDe`) or entailment (`accaoQueCausaAccao`). Other wordnet relation names were adapted to the equivalent names in PAPEL. For instance, `hypernymOf` became `hiperonimoDe` and `substanceHolonymOf` became `materialDe`.

From all the resources, a lexical-semantic network was established with 355,026 lexical items and 1,139,243 triples (excluding inverse relations in the wordnets) respectively distributed according to Table 1.

3.2 Redundancy

As expected, although most triples in the network occurred in only one resource, about 109 k were in more than one, and 192 in all the seven. Table 2 distributes the triples of covered types according to the number of resources they occur at.

A key intuition behind this work is that the more resources a triple is in, the more likely it is to transmit a consensual and useful relation, which is confirmed by selected examples in Table 3. On the other hand, triples that only occur in one resource are more likely to either be incorrect, resulting from noise on the automatic process, or to involve very specific meanings, though less useful.

4 Computing Confidence from Redundancy

We aim at exploiting the potential of redundancy for computing confidence towards the creation of a fuzzy Portuguese wordnet. For this purpose, triples acquired from the seven resources might be the input of a new implementation of the second and third steps of the ECO [5] that should encompass the assignment of scores that transmit confidence. In the second step, fuzzy synsets are discovered from synonymy triples and, in the third, they are connected by different semantic relations, based on the exploitation of all available triples.

4.1 Discovering Fuzzy Synsets

Though not very explored, the idea of fuzzy synsets is not new. Fuzzy memberships of words to synsets have been obtained from manual judgements [18] or from the structure of synonymy networks [19]. In order to integrate domain

Table 1. Number of lexical items and triples used from each exploited resource.

Lexical items							
POS	PAPEL	DA	Wikt.PT	TeP	OT.PT	OWN.PT	PULO
Nouns	56,660	61,334	30,170	17,149	6,110	32,509	5,149
Verbs	21,585	16,429	8,918	8,280	2,856	3,626	1,573
Adjectives	22,561	18,892	9,536	14,568	3,747	4,401	1,316
Adverbs	1,376	3,160	610	1,095	143	1,120	153
Total	102,182	99,815	49,234	41,092	12,856	41,656	8,191
Relations							
Type	PAPEL	DA	Wikt.PT	TeP	OT.PT	OWN.PT	PULO
Synonymy	83,432	52,278	35,330	388,698	51,410	35,597	9,189
Antonymy	388	440	1,263	92,234	–	5,774	2,818
Hypernymy	49,210	46,079	22,931	–	–	78,854	26,596
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
Producer	1,336	913	500	–	–	–	–
Purpose	9,144	5,220	4,227	–	–	–	–
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	–	–	–	–
Place	832	487	1,159	–	–	–	–
Total	191,497	139,404	80,071	480,932	51,410	151,731	44,198

knowledge, PWN has been extended with fuzzy memberships of words to synsets, as well as fuzzy semantic relations [3]. Fuzzy sets of highly related words have also been discovered from text, to represent word senses [20].

Despite its similarities with word sense disambiguation [21], this part of the work can be seen as a kind of word sense induction [22] because, instead of assigning words to senses in an inventory, word senses are drawn from scratch, based on the structure of the synonymy network.

Method: We have recently proposed an alternative approach for discovering fuzzy synsets from synonymy networks, in two steps [23]: (i) centroid discovery; (ii) fuzzy memberships computation. It is applied to a weighted synonymy network $N = (W, P)$, where W is a set of words and P a set of weighted synonym pairs, with a weight reflecting the number of times a synonym pair, $P(W_i, W_j)$,

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

Relation	1	2	3	4	5	6	7
Synonymy	262,325	38,495	14,945	6,035	2,301	792	192
Antonymy	48,444	1,257	345	96	22	7	–
Hypernymy	165,484	23,320	3,188	413	66	–	–
Part	22,620	1,883	146	6	1	–	–
Member	13,200	638	48	3	–	–	–
Material	1,735	159	6	–	–	–	–
Contains	635	65	3	–	–	–	–
Cause	9,286	3,354	927	–	–	–	–
Producer	2,225	217	33	–	–	–	–
Purpose	15,657	1,272	130	–	–	–	–
Property	45,431	6,057	798	76	3	–	–
State	1,031	81	6	1	–	–	–
Quality	1,760	631	72	–	–	–	–
Manner	3,845	551	47	–	–	–	–
Place	1,609	286	99	–	–	–	–
Total	595,287 (85 %)	78,266 (11 %)	20,793 (3 %)	6,630 (0.9 %)	2,393 (0.3 %)	799 (0.1 %)	192 (0.0 %)

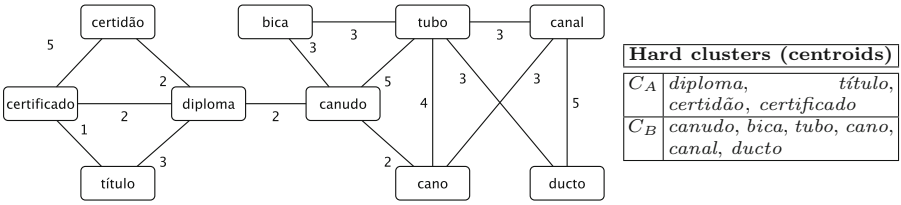
Table 3. Examples of redundant triples.

#	Relation triples
7	*gruta* `sinonimoNDe` *caverna*, *vulgar* `sinonimoAdjDe` *ordinário*, *agarrar* `sinonimoVDe` *pegar* *porventura* `sinonimoAdvDe` *talvez*
6	*público* `antonimoAdjDe` *privado*, *fácil* `antonimoAdjDe` *difícil*, *parcial* `antonimoAdjDe` *imparcial*
5	*pessoa* `hiperonimoDe` *artista*, *mudança* `hiperonimoDe` *mutação*, *degrau* `parteDe` *escada*, *convencional* `dizSeSobre` *convenção*, *sexual* `dizSeSobre` *sexo*, *humanitário* `dizSeSobre` *humanidade*
4	*feliz* `devidoAEstado` *felicidade*, *gendarme* `membroDe` *gendarmaria*, *carta* `membroDe` *baralho*, *letra* `membroDe` *alfabeto*, *decisivo* `dizSeDoQue` *decidir*

occurs in the exploited sources. In the first step, Chinese Whispers [24] (CW), an efficient graph clustering algorithm, is run in the network. This results in a set of hard words clusters, used as centroids. In the second step, the membership degree of each word W_i to each centroid C_k is computed by Eq. 1, which considers the number of synonym pairs between W_i and each word in C_k.

$$\mu(W_i, C_k) = \frac{\sum_{j=0}^{|C_k|} \#(W_i \text{ synonym-of } [C_k]_j)}{|C_k|} \quad (1)$$

Example: The synset discovery approach is illustrated in Fig. 1, with the help of a weighted graph where two senses of the Portuguese word *canudo* arise: a tube/pipe, or, more informally, a diploma. If CW identifies the hard clusters C_A and C_B, to compute the membership of *canudo* to the fuzzy cluster C'_A, the weights of the connections between this word and words in C_A are summed and divided by the size of C_A. Since $\#(canudo \text{ synonym-of } diploma) = 2$,



Hard clusters (centroids)	
C_A	*diploma*, *título*, *certidão*, *certificado*
C_B	*canudo*, *bica*, *tubo*, *cano*, *canal*, *ducto*

Fuzzy clusters (synsets)	
C'_A	*certificado*(3.75), *certidão*(3.5), *diploma*(3.5), *título*(2.75), *canudo*(0.5)
C'_B	*tubo*(3.5), *canal*(3.0), *canudo*(2.83), *cano*(2.66), *ducto*(2.5), *bica*(2.16), *diploma*(0.33)

Fig. 1. Weighted lexical network, resulting hard clusters, and fuzzy synsets.

$\mu(canudo, C'_A) = \frac{2}{4} = 0.5$. For the membership of *canudo* to C'_B, the three connections between this word and words in C_B are considered, plus the word *canudo* itself, which belongs to C_B and has the maximum weight (7, if seven sources are exploited). So $\mu(canudo, C'_B) = \frac{3+5+2+7}{6} = \frac{17}{6} = 2.83$

Results: A total of 20,315 fuzzy synsets (13,735 noun, 4,827 adjective, 1,126 verb, 627 adverbs) were discovered from the synonymy network obtained from the seven exploited resources. On average, noun synsets had 9.4 words, adjectives 11.9 and verbs 59.3, because their network has more connections, which can be interpreted as a higher ambiguity and/or more synonyms for the Portuguese verbs. The resulting fuzzy thesaurus was baptised as CLIP 2.1 [23].

Evaluation: To assess the quality of the fuzzy synsets and computed memberships, random pairs of words from the same synset (240 nouns, 150 verbs, 150 adjectives), organised in sets of ten, were uploaded to the Crowdflower platform[4], where Portuguese-speaking volunteer contributors, living in Portuguese-speaking countries, manually labelled each pair either as possible synonyms or not[5]. In the end, 59 % of the noun pairs, 46 % verb and 55 % adjective pairs were labelled as correct. Each pair was labelled by two judges, respectively with an agreement (IAA) of 87 %, 85 % and 75 %. At first, quality does not look very promising. However, it improves for increasing membership degrees. Figure 2 plots the evolution of the proportion of correct pairs for different cut-points – if the membership of one of the words in the pair is below the cut-point, the pair is ignored – and confirms that the computed memberships behave as a confidence measure, because they are positively correlated with the quality. For instance, for a cut-point of 1.0, the proportion of correct noun and adjective pairs is 85 % and for verbs 89 %. Moreover, there is a point after which all the pairs are correct. Also in Fig. 2, the total number of words and their average number of senses is presented for each cut-point.

[4] https://crowdflower.com/.
[5] The same contributor was not allowed to label more than two sets of pairs.

Cut	Correct pairs N	V	Adj	Size #words	*senses*
0.00	59%	46%	55%	94,835	2.74±3.93
0.25	64%	52%	63%	73,958	1.32±0.85
0.50	79%	67%	78%	63,116	1.13±0.45
0.75	87%	65%	84%	50,897	1.08±0.31
1.00	89%	85%	89%	45,163	1.05±0.24
1.25	89%	89%	96%	21,401	1.04±0.20
1.50	87%	90%	95%	16,949	1.02±0.15
1.75	85%	91%	100%	7,581	1.01±0.11
2.00	83%	94%	100%	5,389	1.01±0.08
2.25	90%	100%	100%	2,378	1.00±0.06
2.50	100%	100%	100%	1,546	1.00±0.04
IAA	87%	85%	75%		

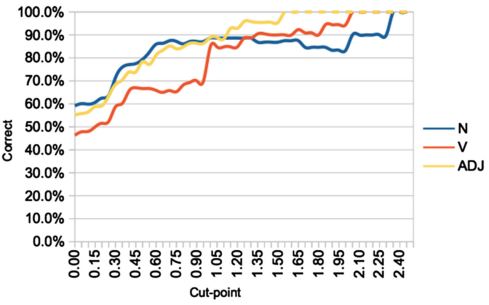


Fig. 2. Evolution of the correct synonymy pairs while increasing the cut-point. (Color figure online)

4.2 Discovering Fuzzy Synset Connections

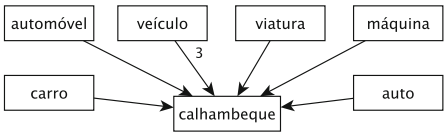
After discovering the fuzzy synsets, some of them may be automatically connected by semantic relations. Possible attachment points can be discovered by exploiting the non-synonymy triples, which is done in this step.

Method: Each pair of synsets, S_a and S_b, is analysed to set attachment points with a fuzzy score, computed by Eq. 2. For each relation type R, this equation considers the: (i) number of triples of type R between a word from each synset, a_i and b_j; (ii) number of resources where each of the previous triples occurs, $\#(a_i, R, b_j)$; (iii) membership of each word in the previous triples to their synset, $\mu(a_i, S_a)$ and $\mu(b_j, S_b)$.

$$c(S_a, R, S_b) = \frac{\sum_{i=0,j=0}^{|S_a|,|S_b|} \left(\#(a_i, R, b_j) \times (\mu(a_i, S_a) + \mu(b_j, S_b))\right)}{|S_a| + |S_b|} : a_i \in S_a, b_j \in S_b \tag{2}$$

Example: Figure 3 illustrates the computation of the proposed measure in two synsets with several hypernymy triples between their words. Hypernymy triples used are represented in a graph, where the only redundant triple has weight 3.

S_1	*carro*(1.2); *automóvel*(1.0); *veículo*(0.94); *viatura*(0.71);*máquina*(0.63); *auto*(0.6)
S_2	*calhambeque*(1.25); *maxambomba*(0.75); *zambeque*(0.5);*caximbeque*(0.5)



$$c(S_1, hiperDe, S_2) = \frac{(1.2+1.25)+(1.0+1.25)+3\times(0.94+1.25)+(0.71+1.25)+(0.63+1.25)+(0.6+1.25)}{6+4} = 1.696$$

Fig. 3. Computing the confidence of the connection S_1 `hiperonimoDe` S_2.

condição(0.97);disposição(0.92);situação(0.88) `hiperonimoDe` crispação(0.8);tensão(0.73);contração(0.6)	**Confidence:** 0.82 **Rendering:** *crispação é um tipo/género de condição*
origem(1.10);princípio(0.81);começo(0.70) `antonimoNDe` término(1.0)	**Confidence:** 2.28 **Rendering:** *origem é o contrário de término*
pressentir(1.73);prognosticar(1.73);prever(1.61) `accaoQueCausa` prognóstico(2.0);presságio(1.77);vaticínio(1.74)	**Confidence:** 0.45 **Rendering:** *pressentir pode levar a prognóstico*
educativo(1.75);doutrinal(1.75);educacional(1.25) `dizSeDoQue` ensinar(2.24);instruir(1.91);doutrinar(1.44)	**Confidence:** 0.23 **Rendering:** *pode ser educativo por ensinar*
desordenar(1.92);destemperar(1.73);desarranjar(1.67) `hiperonimoAccaoDe` contrabalançar(3.75);compensar(3.6);equilibrar(3.0)	**Confidence:** 0.25 **Rendering:** *desordenar é uma forma de contrabalançar*
pitoresco(2.25);pictórico(1.875);pinturesco(1.25) `dizSeSobre` novidade(3.33);nova(3.0);notícia(2.75)	**Confidence:** 0.24 **Rendering:** *pitoresco pode qualificar novidade*
incumbir(3.08);encarregar(2.85);confiar(2.54) `finalidadeDe` mística(2.5);misticismo(2.5);misticidade(0.67)	**Confidence:** 0.32 **Rendering:** *mística pode servir para incumbir*
lado(1.08);ilharga(1.08);flanco(1.0) `parteDeAlgoComPropriedade` trilateral(1.5);trilátero(1.5)	**Confidence:** 0.52 **Rendering:** *lado pode fazer parte de algo que é trilateral*
enfeite(1.0);adorno(0.98);ornato(0.80) `fazSeCom` jarro(1.71);jarra(1.29);vaso(0.63)	**Confidence:** 0.42 **Rendering:** *enfeite pode fazer-se com jarro*
loureiro(2.33);louro(2.0);papagaio(0.75) `membroDe` lauráceas(1.0)	**Confidence:** 1.11 **Rendering:** *loureiro pode ser um membro de lauráceas*

Fig. 4. Examples of discovered synset connections, their computed confidence, and their rendering, used in the crowdsourced evaluation.

Results: The previous measure was computed between all pairs of discovered fuzzy synsets, with a cut-point of 0.1, for relation triples of any type that were in at least two resources. A total of 52,504 synset connections were discovered, with a score higher than 0. As those did not include triples between words without synonyms, and thus not in the discovered fuzzy synsets, in a second step, when a word w involved in a triple was not in any synset, a new synset S_w containing just that w was created, with $\mu(w, S_w) = 1.0$. In the end, 406,751 additional synset connections were made, with at least one synset with a single word. Moreover, 13,542 new single-word synsets were added to the 20,315 multiword synsets discovered earlier.

Evaluation: To assess the quality of the discovered synset connections and the suitability of their computed confidence, we relied once again on Crowdflower, where a random selection of 930 synset connections were uploaded. These included only connections where at least one synset had more than one word. To make labelling faster for the contributors, the following was done before uploading: (i) only the first word of each synset was used, as we noticed that they are often the most representative for the underlying concept; (ii) each triple was rendered to a natural language sentence, depending on the relation type. Contributors could label each rendering as either: (i) correct; (ii) incorrect; or (iii) unsure.

Figure 4 illustrates, at the same time, the output of the fuzzy attachments and of the evaluation samples. It includes the first three words and respective memberships of several synset connections in the sample, their computed confidence, and the textual rendering shown to the contributors.

Figure 5 shows the results of the crowdsourced evaluation and the evolution of the correct connections for increasing cut-points. It also presents the proportion of answers where the contributors were unsure and insights on the size of the fuzzy wordnet for the same cut-points, namely the number of synsets and connections between them. Once again, the initial quality is far from impressive: 49.5 % renderings were labelled as correct and 44.3 % as incorrect. Agreement was also lower, 70 %. It should still be noted that connections between two single-word synsets, with a higher chance of being correct, were not used. Not to mention that, in some cases, the used renderings might be too limitative and they show just one word per synset. Moreover, though less consistently than for the synonyms evaluation, quality still increased for higher cut-points, which indicates that the computed score behaves as a confidence measure. At the same time, the number of connections is drastically reduced each time the cut-point increases, especially from 0 to 0.25.

Cut	Triples		Size	
	Correct	Unsure	#synsets	#conns
0.00	49.5%	6.2%	33,857	459,255
0.25	55.0%	6.0%	33,857	25,966
0.50	62.3%	4.5%	33,844	7,657
0.75	64.1%	6.4%	33,692	3,347
1.00	72.1%	4.7%	33,229	1,724
1.25	79.2%	4.2%	9,117	1,212
1.50	72.0%	4.0%	7,770	590
1.75	64.3%	0.0%	3,781	336
2.00	69.2%	0.0%	3,042	153
2.25	71.4%	0.0%	1,537	127
2.50	83.3%	0.0%	1,061	66
IAA	70%			

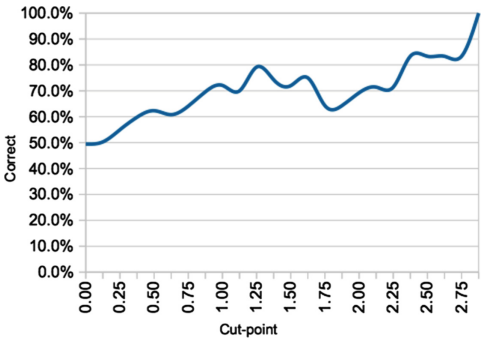


Fig. 5. Evolution of the correct triples while increasing the cut-point.

After a shallow error analysis, we noticed that there were several renderings that should have been labelled as correct, but were not. Those included connections with confidence higher than 1.8, such as (*origem* `antonimoDe` *término*), (*dicéfalo* `dizSeDoQue` *ter_duas_cabeças*), or (*planta* `hiperonimoDe` *bisnaga*). Although we asked the contributors to confirm their answers in electronic dictionaries and check for less known senses, or to mark unknown answers as unsure, most of them were probably less experienced or have answered the questions too fast, thus not following the instructions strictly.

5 Conclusion and Further Work

The first experiments towards the automatic creation of a fuzzy Portuguese wordnet, through the exploitation of redundancy in available lexical-semantic

resources, were presented. The projected wordnet combines the advantages of an automatic creation approach, including lower creation effort for a broad-coverage resource, with the option of controlling the quantity-quality trade-off, with a confidence cut-point. Synsets, discovered from synonymy networks, have words with variable memberships, and they can be connected, by semantic relations of different types, to other synsets, also with variable degrees.

A preliminary version of the resulting wordnet is available, in a non-standard format, from http://ontopt.dei.uc.pt, under the option CONTO.PT. We are still studying alternatives for representing CONTO.PT with standard formats, such as RDF/OWL.

Besides dealing with the previous issue, there is additional work to do. Alternative ways of computing confidence from redundancy should be explored, especially on the synset attachment, where the current measure seems to be biased towards smaller synsets. In order to measure progress, we can use the annotated data collected from crowdsourcing or, given the limitations of the previous, a more controlled evaluation might be performed by more experienced and trustful judges. It should also be analysed whether the synset memberships can be adjusted when connecting synsets. For instance, if several words of the same synset share a relation with another word, their memberships may increase.

It should be added that, although applied to Portuguese, this approach can be used to create fuzzy wordnets in other languages, as long as there are available computational lexical resources, whether they are dictionaries, thesauri, wordnets or even relations extracted from corpora.

References

1. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database (Language, Speech, and Communication). The MIT Press, Cambridge (1998)
2. Gonçalo Oliveira, H., de Paiva, V., Freitas, C., Rademaker, A., Real, L., Simões, A.: As wordnets do Português. In: Simões, A., Barreiro, A., Santos, D., Sousa-Silva, R., Tagnin, S.E.O. (eds.) Linguística, Informática e Tradução: Mundos que se Cruzam, pp. 397–424. OSLa: Oslo Studies in Language, University of Oslo (2015)
3. Araúz, P.L., Gómez-Romero, J., Bobillo, F.: A fuzzy ontology extension of WordNet and EuroWordnet for specialized knowledge. In: Proceedings of Terminology and Knowledge Engineering Conference, TKE 2012, Madrid, Spain, June 2012
4. Kilgarriff, A.: Word senses are not bona fide objects: implications for cognitive science, formal semantics, NLP. In: Proceedings of 5th International Conference on the Cognitive Science of Natural Language Processing, pp. 193–200 (1996)
5. Gonçalo Oliveira, H., Gomes, P.: ECO and Onto.PT: a flexible approach for creating a Portuguese wordnet automatically. Lang. Resour. Eval. **48**(2), 373–393 (2014)
6. Marrafa, P., Amaro, R., Mendes, S.: WordNet.PT global - extending WordNet.PT to Portuguese varieties. In: Proceedings of 1st Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties, Edinburgh, Scotland, pp. 70–74. ACL Press (2011)
7. Dias-da-Silva, B.C., de Oliveira, M.F., de Moraes, H.R.: Groundwork for the development of the Brazilian Portuguese wordnet. In: Ranchhod, E., Mamede, N.J. (eds.) PorTAL 2002. LNCS (LNAI), vol. 2389, pp. 189–196. Springer, Heidelberg (2002)

8. Dias-da-Silva, B.C.: Wordnet.Br: an exercise of human language technology research. In: Proceedings of 3rd International WordNet Conference (GWC), GWC 2006, South Jeju Island, Korea, pp. 301–303, January 2006
9. de Paiva, V., Rademaker, A., de Melo, G.: OpenWordNet-PT: an open Brazilian wordnet for reasoning. In: Proceedings of 24th International Conference on Computational Linguistics, COLING (Demo Paper) (2012)
10. 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), pp. 513–522. ACM, New York (2009)
11. Simões, A., Guinovart, X.G.: Bootstrapping a Portuguese wordnet from Galician, Spanish and English wordnets. In: Navarro Mesa, J.L., Ortega, A., Teixeira, A., Hernández Pérez, E., Quintana Morales, P., Ravelo García, A., Guerra Moreno, I., Toledano, D.T. (eds.) IberSPEECH 2014. LNCS, vol. 8854, pp. 239–248. Springer, Heidelberg (2014)
12. 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 (LREC 2012), pp. 2525–2529. ELRA (2012)
13. Gomes, M.M., Beltrame, W., Cury, D.: Automatic construction of Brazilian Portuguese WordNet. In: Proceedings of X National Meeting on Artificial and Computational Intelligence, ENIAC 2013 (2013)
14. Gonçalo Oliveira, H., Santos, D., Gomes, P., Seco, N.: PAPEL: a dictionary-based lexical ontology for Portuguese. In: Teixeira, A., de Lima, V.L.S., de Oliveira, L.C., Quaresma, P. (eds.) PROPOR 2008. LNCS (LNAI), vol. 5190, pp. 31–40. Springer, Heidelberg (2008)
15. Simões, A., Sanromán, Á.I., Almeida, J.J.: Dicionário-Aberto: a source of resources for the Portuguese language processing. In: Caseli, H., Villavicencio, A., Teixeira, A., Perdigão, F. (eds.) PROPOR 2012. LNCS, vol. 7243, pp. 121–127. Springer, Heidelberg (2012)
16. Gonçalo Oliveira, H., Antón Pérez, L., Costa, H., Gomes, P.: Uma rede léxico-semântica de grandes dimensões para o português, extraída a partir de dicionários electrónicos. Linguamática **3**(2) 23–38, 2011
17. 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: VI Workshop em Tecnologia da Informação e da Linguagem Humana (TIL), pp. 390–392 (2008)
18. Borin, L., Forsberg, M.: From the people's synonym dictionary to fuzzy synsets - first steps. In: Proceedings of LREC 2010 Workshop on Semantic Relations. Theory and Applications, La Valleta, Malta, pp. 18–25 (2010)
19. Gonçalo Oliveira, H., Gomes, P.: Automatic discovery of fuzzy synsets from dictionary definitions. In: Proceedings of 22nd International Joint Conference on Artificial Intelligence, IJCAI 2011, Barcelona, Spain, pp. 1801–1806. IJCAI/AAAI, July 2011
20. Velldal, E.: A fuzzy clustering approach to word sense discrimination. In: Proceedings of 7th International Conference on Terminology and Knowledge Engineering, Copenhagen, Denmark (2005)
21. Navigli, R.: Word sense disambiguation: a survey. ACM Comput. Surv. **41**(2), 1–69 (2009)

22. Nasiruddin, M.: A state of the art of word sense induction: a way towards word sense disambiguation for under resourced languages. In: Proceedings of Traitement Automatique des Langues Naturelles and Rencontres des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues, TALN/RECITAL 2013 (2013)
23. Gonçalo Oliveira, H., Santos, F.: Discovering fuzzy synsets from the redundancy in different lexical-semantic resources. In: Proceedings of 10th Language Resources and Evaluation Conference, LREC 2016, Portorož, Slovenia. ELRA, May 2016
24. Biemann, C.: Chinese whispers: an efficient graph clustering algorithm and its application to natural language processing problems. In: Proceedings of 1st Workshop on Graph Based Methods for Natural Language Processing, TextGraphs-1, New York City, pp. 73–80. ACL Press (2006)