Automatic Discovery of Fuzzy Synsets from Dictionary Definitions

Hugo Gonçalo Oliveira and **Paulo Gomes**
CISUC, University of Coimbra, Portugal
{hroliv,pgomes}@dei.uc.pt

Abstract

In order to deal with ambiguity in natural language, it is common to organise words, according to their senses, in synsets, which are groups of synonymous words that can be seen as concepts. The manual creation of a broad-coverage synset base is a time-consuming task, so we take advantage of dictionary definitions for extracting synonymy pairs and clustering for identifying synsets. Since word senses are not discrete, we create fuzzy synsets, where each word has a membership degree. We report on the results of the creation of a fuzzy synset base for Portuguese, from three electronic dictionaries. The resulting resource is larger than existing hancrafted Portuguese thesauri.

1 Introduction

Information systems are becoming more and more dependent on natural language processing (NLP) tasks, including the determination of similarities [Agirre *et al.*, 2009], word sense disambiguation (WSD) [Gomes *et al.*, 2003] or question-answering [Pasca and Harabagiu, 2001], where broad-coverage lexical resources, as WordNet [Fellbaum, 1998], play a crucial role. In opposition to formal languages, natural language is ambiguous – different words may have the same meaning and the same word may have different meanings, commonly referred to as word senses. Therefore, WordNet-based lexical resources represent concepts as synsets, which are groups of words that, in some context, have the same meaning and are thus synonyms.

However, from a linguistic point of view, word senses are not discrete and cannot be separated with clear boundaries [Kilgarriff, 1997] [Hirst, 2004]. They are typically complex and overlapping structures. So, sense division in dictionaries and lexical resources is most of the times artificial. On the other hand, this trade-off is often needed to increase the usability of computational lexical resources.

A more realistic approach for coping with this fact is to represent synsets as models of uncertainty, such as fuzzy sets, where each word has a membership degree. The fuzzy membership of a word in a synset can be interpreted as the confidence level about using this word to indicate the meaning of the synset. It can also have a probabilistic interpretation, and denote the likelihood of a word conveying the meaning of each synset it belongs to.

We propose to obtain fuzzy synsets from dictionaries, which are broad-coverage resources structured on word senses, though not fully ready to be used as computational lexical resources. First, we exploit specific textual patterns in definitions for extracting synonymy pairs (hereafter synpairs). Then, the similarity between words in synpairs is computed and fuzzy clusters (hereafter fuzzy synsets) are discovered, in a graph formed by these pairs. The result is a fuzzy thesaurus, Padawik, which, after applying a cut point to the membership degrees, can also be used as a simple thesaurus, similar to the WordNet synset-base.

This work is in the scope of a project [Gonçalo Oliveira and Gomes, 2010] aiming to create a lexical ontology for Portuguese semi-automatically, from existing textual resources. It is an alternative to the manual creation of lexico-semantic resources and, even though it may be applied to other languages, intends to be a contribution towards the development of Portuguese NLP. Moreover, Padawik is not only the largest free Portuguese thesaurus, but also the first attempt to create a Portuguese thesaurus based on uncertainty.

We start by introducing some related work and move on to the description of our approach, including the extraction of synpairs from dictionaries and the discovery of clusters in synonymy graphs. Then, we present our experimentation and some results. Fuzzy synsets were discovered from a graph extracted from three Portuguese dictionaries. Before concluding, we compare the resulting thesaurus, Padawik, with handcrafted Portuguese thesauri and show the results of the manual synset evaluation, where more than 73% of the synsets were classified as correct.

2 Related Work

During the last decades, dictionaries have been exploited in the automatic creation of computational lexico-semantic resources (e.g. [Chodorow *et al.*, 1985]). Semantic relations (e.g. synonymy, hyponymy), represented as triples (a relation b) were extracted, but words (a and b) can denote different concepts, thus making this representation impractical for tasks such as the inference of new knowledge. As a result, popular lexical resources (e.g. WordNet) organise words in synsets, which are groups of synonymous words that can be seen as the concepts of a lexical ontology.

Despite structured on word senses, dictionaries are not ready to be used as computational lexical resources nor to produce synsets, because they list word senses and describe them by words, not senses. Therefore, ambiguity prevents a straightforward match between the occurrences of words and their possible senses. Furthermore, word senses in different dictionaries do not always match, because there is not a well-defined criteria for the division of meanings into word senses [Dolan, 1994] [Peters *et al.*, 1998]. Senses of the same word can go from tightly related (e.g. in polysemy or metonymy) to completely unrelated (e.g. homonymy).

Regarding that information in dictionaries is usually incomplete, using more than one dictionary [Ide and Veronis, 1995], or using alternative sources of knowledge, such as corpora [Lin and Pantel, 2002], can minimise this problem. However, in opposition to other kinds of relation, synonymous words, despite sharing similar neighbourhoods, may not co-occur frequently in corpora text [Dorow, 2006], which leads to few textual patterns connecting this kind of words. Most of the works on synonymy (or near-synonymy) extraction from corpora rely on the application of mathematical models (e.g. [Turney, 2001]), including graphs, clustering algorithms, or both (e.g. [Dorow, 2006]). Claiming that the so called broad-coverage lexical resources do not cover many concepts found in text, [Lin and Pantel, 2002] propose an algorithm for discovering concepts described by words belonging to some class (e.g. firearms, cells), after clustering words occurring in similar contexts.

Co-occurrence graphs extracted from corpora are useful for identifying not only synonymous words, but also word senses [Dorow, 2006]. As for syonymy graphs extracted from dictionaries, clusters tend to express concepts [Gfeller *et al.*, 2005] and can be exploited for the establishement of synsets. Methods for the improvement of synonymy graphs, extracted from different resources, were presented by [Navarro *et al.*, 2009].

From a linguistic point of view, word senses are not discrete, so their representation as crisp objects does not reflect the human language. Therefore, it is more realistic to adopt models of uncertainty, including fuzzy logic, to handle word senses and natural language concepts. For instance, [Velldal, 2005] describes a similar work to [Lin and Pantel, 2002], but he represents word sense classes as fuzzy clusters, where each word has an associated membership degree. Furthermore, [Borin and Forsberg, 2010] present an ongoing work on the creation of Swedish fuzzy synsets. They propose two methods for achieving their purpose using a lexicon with word senses and a set of synpairs. The fuzzy membership values are based on human judgements of the synpairs.

3 From dictionary definitions to fuzzy synsets

The process of identifying fuzzy synsets, described in this section, consists of two stages: i) extraction of a synonymy graph from a dictionary; ii) clustering words in synsets, based on the configuration of the graph.

3.1 Extraction of synpairs from dictionaries

As referred in Section 2, sense divisions in dictionaries are almost arbitrary, so we do not consider them. Synpairs can be extracted from plain definitions consisting of only one word, enumerations or definitions using synonymy textual patterns as the following:

- **mind**, n: *brain, head, intellect*
  - (*brain*, *mind*) (*head*, *mind*) (*intellect*, *mind*)
- **machine**, n: *the same as computer*
  - (*computer*, *machine*)

In addition, dictionaries such as the Wiktionary provide synonymy lists for some of its entries.

3.2 Discovery of fuzzy synsets in synonymy graphs

Synonymy graphs are structures $G = (N, E)$, with $|N|$ nodes and $|E|$ edges, $E \subset N^2$. Each node $n_a \in N$ represents a word and each edge connecting n_a and n_b, $E(n_a, n_b)$, indicates that, in some context, words n_a and n_b may have the same meaning and are thus synonymous.

In our case, each synpair defines an edge of the synonymy graph, G. Synsets with fuzzy membership are then identified after running the following graph clustering algorithm on G:

1. Create an empty sparse matrix M, $|N| \times |N|$.
2. Fill each cell M_{ij} with the similarity between the adjacency vectors of the words n_i and n_j, $\vec{n_i}$ and $\vec{n_j}$.
3. Normalise the columns of M, so that the values in each column, M_j, sum up to 1.
4. Extract a fuzzy cluster F_i from each row M_i, consisting of the words n_j where $M_{ij} > 0$. The value in M_{ij} is used as the membership degree of the word n_j to F_i, $\mu_{F_i}(n_j)$.
5. For each cluster F_i with all elements included in a bigger cluster F_j ($F_i \cup F_j = F_j$ and $F_i \cap F_j = F_i$), F_i and F_j are merged, giving rise to a new cluster F_k with the same elements of F_j, where the membership degrees of the common elements are summed, $\mu_{F_k}(n_j) = \mu_{F_i}(n_j) + \mu_{F_j}(n_j)$.

This algorithm is simpler than fuzzy c-means [Bezdek, 1981]. In opposition to the latter, there is no need to keep two matrixes, one with the weights and another with the centroids, which is important because synonymy graphs can be very large. Moreover, there is no need to specifiy the number of clusters – words are organised into m clusters, where m is never higher than the number of unique words, $|N|$.

If $\mu_{F_i}(n_a) > 0$, the word n_a has a sense with a common meaning to the other words in F_i. The membership degree $\mu_{F_i}(n_a)$ is thus the confidence level on the usage of the word n_a with the meaning of the synset F_i. Looking at all the membership degrees of the same word, they all sum up to 1, $\sum \mu_{F_i}(n_j) = 1$. As a result, membership degrees of n_a can also be interpreted as the possible senses of the word n_a and the likelihood of the word n_a conveying their meanings.

Any measure for computing the similarity of two vectors can be used in item 2 of the algorithm. If the adjacencies are binary vectors, measures typically used for computing the similarity between sets, such as the Jaccard coefficient, are a suitable alternative.

However, if it is possible to extract a synpair more than once, we can take advantage of redundancy. So, the number of edges between nodes n_a and n_b can be considered for calculating the adjacency vectors, $\vec{n_a}$ and $\vec{n_b}$. Each edge of

the graph thus becomes a triple $E(n_i, n_j, w_{ij})$, where a synpair (n_a, n_b) has an associated weight, w_{ab}, relative to the number of times it was extracted[1]. Furthermore, inspired by [Lin and Pantel, 2002], each position of the vectors $\vec{n_i}$ can be filled with the pointwise mutual information (pmi) between the word n_i and all the other words, computed using expression 1. As the pmi is biased towards infrequent words, it should be multiplied by the discounting factor in expression 2, also suggested by [Lin and Pantel, 2002]. Finally, the similarity of two words is given, for instance, by the cosine similarity between their vectors (expression 3).

$$pmi(a,b) = \frac{\frac{w_{ab}}{W}}{\frac{\sum_{j=0}^{|N|} w_{aj}}{W} \times \frac{\sum_{i=0}^{|N|} w_{ib}}{W}}, W = \sum_{i=0}^{|N|}\sum_{j=0}^{|N|} w_{ij} \quad (1)$$

$$df(a,b) = \frac{w_{ab}}{w_{ab}+1} \times \frac{min\left(\sum_{j=0}^{|N|} w_{aj}, \sum_{i=0}^{|N|} w_{ib}\right)}{min\left(\sum_{j=0}^{|N|} w_{aj}, \sum_{i=0}^{|N|} w_{ib}\right)+1} \quad (2)$$

$$sim(n_a, n_b) = \frac{\vec{n_a}.\vec{n_b}}{|\vec{n_a}||\vec{n_b}|} = \frac{\sum_{i=0}^{|N|} pmi(a,n_i) \times pmi(b,n_i)}{\sqrt{\sum_{i=0}^{|N|} pmi(a,n_i)^2 \times \sum_{i=0}^{|N|} pmi(b,n_i)^2}} \quad (3)$$

4 A fuzzy thesaurus for Portuguese

In this section, we describe our experimentation towards the creation of a fuzzy Portuguese thesaurus, Padawik, and analyse some results. As information in dictionaries is often incomplete [Ide and Veronis, 1995], three dictionaries were used to maximise both quality and coverage.

4.1 Resources used

Synpairs were acquired from two public domain dictionaries of Portuguese and from a public domain lexical network, PAPEL 2.0 [Gonçalo Oliveira *et al.*, 2010], extracted from a proprietary dictionary. Both public domain dictionaries were processed by the grammars available through PAPEL's website[2], created specifically to extract several semantic relations, including synonymy, from definitions in Portuguese.

One of the dictionaries used was Dicionário Aberto [Simões and Farinha, 2010], a resource based on the transcription of a Portuguese dictionary from 1913. It contains approximately 227K definitions, some of them written in old Portuguese. Therefore, for each pair with words containing disused sequences, several replacements were made according to the suggestions in [Simões *et al.*, 2010]. If the new words were in PAPEL, we kept the pair, otherwise we discarded it. The other dictionary was the 25th October 2010 dump of the Portuguese Wiktionary[3], a collaborative dictionary by the Wikimedia foundation. Despite several parsing issues due to the lack of standartisation of the wiki text (also referred by [Navarro *et al.*, 2009]), 54K definitions and synonymy lists were collected from Wiktionary.

Table 1: Data of the synonymy graphs.

POS	$\|N\|$	$\|E\|$	$\overline{deg}(G)$	$\|N_{lcs}\|$	$\overline{CC_{lcs}}$
Nouns	39,355	57,813	2.93	25,828	0.26
Verbs	11,502	28,282	4.92	10,631	0.29
Adj.	15,260	27,040	3.54	11,006	0.31

4.2 Synonymy graph data

Table 1 has the properties of the graphs established by our synpairs: number of nodes $|N|$ and edges $|E|$, average degree $\overline{deg}(G)$ of the graph, average clustering coefficient $\overline{CC_{lcs}}$ and the number of nodes of the largest connected subgraph $|N_{lcs}|$. Weights were not considered in the construction of this table.

The average degree (expression 4) is the ratio between the number of nodes $|N|$ and the number of edges $|E|$ in the graph. The average clustering coefficient (expression 5) measures the degree to which nodes tend to cluster together as a value in [0-1]. In random graphs, this coefficient is close to 0. The local clustering coefficient $CC(n_i)$ (expression 6) of a node n_i quantifies how connected its neighbours are.

$$\overline{deg}(G) = \frac{1}{|N|} \times \sum_{i=1}^{|N|} deg(n_i) = \frac{1}{|N|} \times \sum_{i=1}^{|N|} |E(n_i, n_j)| : n_i, n_j \in N \quad (4)$$

$$\overline{CC} = \frac{1}{|N|} \times \sum_{i=1}^{|N|} CC(n_i) \quad (5)$$

$$CC(n_i) = \frac{2 \times |E(n_j, n_k)|}{K_i(K_i-1)} : n_j, n_k \in neigh(n_i) \wedge K_i = |neigh(n_i)| \quad (6)$$

Despite having a large number of nodes, all the graphs are quite sparse. It is thus possible to represent them as sparse matrixes and minimise memory consumption.

An interesting fact is that the largest connected subgraph lcs contains always more than half of the total nodes. If there was no ambiguity, this would mean all the words in lcs were synonymys of each other, which is not true. This points out the need of additional organisation of synonymy automatically extracted from dictionaries. Clustering coefficients are slightly higher than those of graphs extracted from several Wiktionaries (between 0.2 and 0.28) [Navarro *et al.*, 2009].

4.3 Qualitative results

Figure 1 presents one subgraph and the fuzzy synsets obtained after clustering. It shows words denoting a person who rules and divides them in two slightly different concepts. A ceaser/emperor is someone who rules an empire, while a king rules a kingdom. Nevertheless, several words can denote both concepts, with different membership degrees.

To have and idea on how ambiguity was handled in the establishment of fuzzy synsets, we selected two polysemic Portuguese words, *pasta* and *cota*, looked at some of the synsets

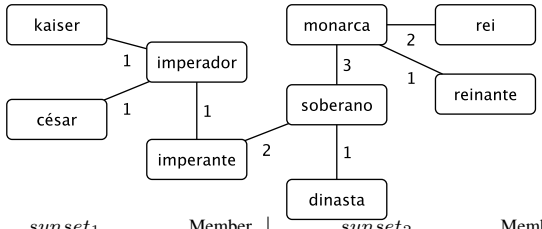
[1] If the order of the words in a synpair is considered (e.g. (a,b), (b,a)) at most two equivalent synpairs can be extracted from each dictionary.

[2] Available through http://www.linguateca.pt/PAPEL/

[3] Available through http://pt.wiktionary.org/

Table 2: Fuzzy synsets of polysemic words

Word	Concept	Fuzzy synsets
pasta	money	*arame*(0.6774), *zerzulho*(0.6774), *metal*(0.6774), *carcanhol*(0.6774), *pecunia*(0.6774), *bagarote*(0.6774), *pecuniária*(0.6774), *cunques*(0.6774), *matambira*(0.6774), *jan-da-cruz*(0.6774), *bagalho*(0.6774), *cacau*(0.6774), *boro*(0.6774), *calique*(0.6774), *marcaureles*(0.6774), *teca*(0.6774), *níquel*(0.6774), *mussuruco*(0.6774), *massaroca*(0.6774), *baguines*(0.6774), *bilhestres*(0.6774), *parrolo*(0.6774), *pastel*(0.6774), *cum-quibus*(0.6774), *dieiro*(0.6774), *pilim*(0.6774), *gimbo*(0.6735), *chelpa*(0.6735), *pecúnia*(0.6735), *patacaria*(0.6735), *pataco*(0.6347), *bagalhoça*(0.62), *bago*(0.6181), *china*(0.6178), *cobre*(0.6173), *numo*(0.616), *maco*(0.5971), *jimbo*(0.5953), *guines*(0.5903), *pasta*(0.5657), *maquia*(0.5243), *gaita*(0.5242), *grana*(0.5226), *painço*(0.517), *jibungo*(0.517), *numerário*(0.5145), *dinheiro*(0.5139), *fanfa*(0.4617), *posses*(0.4604), *finanças*(0.4425), *ouro*(0.4259), ...
	file	*diretório*(1.0), *dossier*(0.9176), *pasta*(0.1118), ...
	mixture	*amálgama*(0.09279), *dossier*(0.08130), *landoque*(0.05162), *angu*(0.04271), *pot-pourri*(0.03949), *marinhagem*(0.03722), *mosaico*(0.03648), *cocktail*(0.03480), *mixagem*(0.02688), *cacharolete*(0.02688), *macedónia*(0.02688), *comistão*(0.02374), *colectânea*(0.02317), *anguzada*(0.02205), *caldeação*(0.02108), *mistura*(0.02032), *moxinifada*(0.01976), *imisção*(0.01917), *massamorda*(0.01845), *pasta*(0.01827), *incorporação*(0.01800), *farragem*(0.01779), *matalotagem*(0.01397), *misto*(0.01280), *salsada*(0.01262), *ensalsada*(0.01050)
	briefcase	*maleta*(0.0759), *saco*(0.0604), *maco*(0.054), *bagalhoça*(0.0263), *fole*(0.0154), ..., *pasta*(0.0128), ...
cota	mother	*mamãe*(0.8116), *mamã*(0.8116), *nai*(0.7989), *malúrdia*(0.7989), *darona*(0.7989), *mamana*(0.7989), *velha*(0.7989), *mãe-de-famílias*(0.7989), *ti*(0.7989), *mare*(0.6503), *naia*(0.5549), *uiara*(0.5549), *genetriz*(0.5549), *mãe*(0.5221), *madre*(0.2749), *cota*(0.2407), ...
	father	*palúrdio*(0.6458), *dabo*(0.6458), *genitor*(0.6458), *painho*(0.6458), *benfeitor*(0.6458), *papaí*(0.6183), *papá*(0.6169), *tatá*(0.4934), *paí*(0.3759), *primogenitor*(0.3543), *velhote*(0.2849), *velho*(0.2817), ... , *cota*(0.1463), *progenitor*(0.08416015), *ascendente*(0.062748425)
	quota	*colecta*(0.6548), *quota*(0.5693), *contingente*(0.309), *pagela*(0.2304), *prestação*(0.1723), *cota*(0.1655), *mensalidade*(0.0908), *quinhão*(0.0605),...



$synset_1$	Member	$synset_2$	Member
kaiser ($kaiser$)	1.0	reinante ($regnant$)	1.0
césar ($caesar$)	1.0	rei ($king$)	1.0
imperador ($emperor$)	0.95	monarca ($monarch$)	0.86
imperante ($dominant$)	0.57	soberano ($sovereign$)	0.85
dinasta ($dynast$)	0.24	dinasta ($dynast$)	0.76
soberano ($sovereign$)	0.15	imperante ($dominant$)	0.43
monarca ($monarch$)	0.14	imperador ($emperor$)	0.05

Figure 1: Weighted synonymy graph and resulting synsets

containing them, and divided them into possible senses of these words, as shown in Table 2.

4.4 Thesaurus data for different cut points

Different cut-points (θ) applied to a fuzzy thesaurus result in different simple thesauri. Table 3 is an overview on the noun thesauri obtained from Padawik. It includes the number of words and ambiguous words, the average number of senses per word, the number of senses of the most ambiguous word, the number of synsets, the average synset size in terms of words, synsets of size 2 and size larger than 25, which are less likely to be useful [Borin and Forsberg, 2010], and the largest synset. This table does not consider synsets of size 1.

Before collecting the data in Table 3, we followed one of the clustering methods for word senses proposed for EuroWordNet [Peters *et al.*, 1998], which suggests that synsets with three members in common can be merged. However, the design of our clustering algorithm and the configuration of our synonymy graphs are prone to create synsets sharing more than one word. So, to minimise the possibility of merging synsets denoting different concepts, we made sure that merged synsets had at least 75% overlap (expression 7).

$$Overlap(S_a, S_b) = \frac{S_a \cap S_b}{min(|S_a|, |S_b|)} \quad (7)$$

As expected, as θ grows, ambiguity drops. This is observed not only from the number of ambiguous words, but also from the average number of word senses and the number of synsets.

Out of curiosity, the largest synset in Padawik with $\theta \geq 0.075$ denotes the concept of *money*. We manually checked that every word in this synset could have this meaning, which indicates there are many ways of referring to *money* in Portuguese: informal (e.g. *pastel*, *pasta*, *carcanhol*, *pilim*), popular (e.g. *massaroca*, *cacau*, *guita*), Brazilian (e.g. *grana*, *tutu*) or Mozambican Portuguese variant (e.g. *mussuruco*, *matambira*), figurative senses (*ouro*, *metal*) and older forms (*dieiro*), amongst others. Another large synset in most thesauri refers to alcoholic intoxication and contains words such as *piela*, *touca*, *pifo*, *chiba*, *carraspana*, *bezana* or *bebedeira*.

5 Evaluation

Evaluating clustering results is usually a difficult task so, in order to make it simpler for humans, we selected Padawik with $\theta = 0.075$, whose size was compared against existing Portuguese thesauri and then manually evaluated.

5.1 Comparison with existing Portuguese thesauri

There are two open domain handcrafted Portuguese thesauri, organised in synsets: OpenThesaurus[4], a small collaborative thesaurus whose main purpose is to suggest writing alternatives in OpenOffice[5]; and TeP [Maziero *et al.*, 2008], a thesaurus for Brazilian Portuguese. Table 4 puts side-by-side the former thesauri, fuzzy Padawik and Padawik with $\theta = 0.075$, which is between the cut-points that lead to the closest average senses ($\theta = 0.05$) and the closest average synset size ($\theta = 0.15$) as compared to TeP's. Both automatically created thesauri are larger than the handcrafted ones. For nouns, the former have two times more words than TeP.

We did not use these thesauri as gold standards for evaluation because, as other authors analysed [Teixeira *et al.*, 2010], they have low overlaps with PAPEL and Wiktionary.

[4]Available through http://openthesaurus.caixamagica.pt/

[5]Available through http://www.openoffice.org/

Table 3: Noun thesauri data, using different cut points θ

θ	Words				Synsets				
	Total	**Ambig.**	**Avg(senses)**	**Most ambig.**	**Total**	**Avg(size)**	$size = 2$	$size > 25$	$max(size)$
0.025	39,350	21,730	3.18	18	13,344	9.39	3,921	576	80
0.05	39,288	17,585	1.86	9	12,416	5.89	4,224	119	62
0.075	38,899	12,505	1.44	7	12,086	4.64	4,878	47	58
0.1	38,129	8,447	1.26	6	11,748	4.10	5,201	34	58
0.15	35,772	4,198	1.12	4	11,044	3.64	5,248	16	58
0.25	30,266	1,343	1.04	3	9,830	3.22	5,095	10	58
0.5	22,203	0	1.0	1	8,004	2.77	5,011	3	47

Table 4: Thesaurus comparison

Thesaurus	**POS**	**Words**				**Synsets**				
		Quant.	**Ambig.**	**Avg(senses)**	**Most ambig.**	**Quant.**	**Avg(size)**	$size = 2$	$size > 25$	$max(size)$
OpenThesaurus.PT	Nouns	6,110	485	1.09	4	1,969	3.38	778	0	14
	Verbs	2,856	337	1.13	5	831	3.90	226	0	15
	Adjectives	3,747	311	1.09	4	1,078	3.80	335	0	17
TeP 2.0	Nouns	17,158	5,805	1.71	20	8,254	3.56	3,079	0	21
	Verbs	10,827	4,905	2.08	41	3,978	5.67	939	48	53
	Adjectives	14,586	3,735	1.46	19	6,066	3.50	3,033	19	43
Padawik-fuzzy	Nouns	39,354	24,343	7.78	46	20,102	15.23	3,885	3,756	109
	Verbs	11,502	10,411	14.31	42	7,775	21.17	307	2,411	89
	Adjectives	15,260	10,636	10.36	43	8,896	17.77	1,326	2,157	109
Padawik-0.075	Nouns	38,899	12,505	1.44	7	12,086	4.64	4,878	47	58
	Verbs	11,070	5,717	1.76	7	4,198	4.63	1,189	14	49
	Adjectives	14,964	6,644	1.69	6	5,666	4.45	1,980	11	46

5.2 Manual evaluation

In order to make manual evaluation faster and less tedious, we selected a subset of the noun synsets in Padawik-0.075. First, we removed all the words without occurrences in the frequency lists[6] of AC/DC [Santos and Bick, 2000], which compile word frequencies in several Portuguese corpora. Then, we selected only the 834 synsets with all words with AC/DC frequencies higher than 100. We were left with a thesaurus of 1,920 words, 227 of those ambiguous, and 1.13 senses per word. Synsets had an average of 2.61 words and the largest had 10 words.

From this thesaurus, we created 22 random samples: 11 with 40 synsets and 11 with 40 synpairs, established by two words selected randomly from the same synset. Synpairs can be handled as a synset of two words. So, given a sample, judges classified each synset as: correct (1), if, in some context, all the words could have the same meaning, or incorrect (0), if at least one word could not have the same meaning as the others. Judges were advised to look for possible word senses in different dictionaries. If they still did not know how to classify the synset, they had a third option, N/A (2).

The evaluation results, in Table 5, show that the average correction of Padawik's synsets is higher than 73%, with agreements higher than 80%, which we believe to be a good quality indicator. When we decided to evaluate our data as synsets and also as synpairs, we intended to give two different perspectives on its quality. However, both kinds of evaluation yielded similar results, as the correction of synpairs is 75%.

Table 5: Results of manual evaluation of synsets and synpairs.

	Synsets sample = 440 × 2 sets		**Synpairs** sample = 440 × 2 pairs	
Correct	646	(73.4%)	660	(75.0%)
Incorrect	231	(26.3%)	218	(24.8%)
N/A	3	(0.3%)	2	(0.2%)
Agreement	364	82.7%	366	83.2%

6 Concluding remarks

Having in mind that word senses are not discrete, representing natural language concepts as fuzzy synsets is closer to reality than using simple synsets. We have shown that the former structures can be acquired from dictionary definitions by identifying clusters in synonymy graphs.

Future directions include learning individual cut-points for each fuzzy synset, and also exploring different strategies to capture word senses not described by synonymous words.

Padawik will be publicly available. For instance, it will be possible to integrate Padawik in the OpenOffice tools for Portuguese as an alternative to OpenThesaurus, which is more than four times smaller. Moreover, if Padawik is merged with handcrafted thesauri, an even larger thesaurus can be obtained. Still, since size is not the only important property of a thesaurus, we are devising the creation of smaller thesauri, after filtering less common words.

[6] Available through http://www.linguateca.pt/ACDC/

Acknowledgements

We would like to thank the CMS group for the manual evaluation of synsets and Leticia Antón Pérez for developing the Wiktionary parser. Hugo Gonçalo Oliveira is supported by the FCT grant SFRH/BD/44955/2008 co-funded by FSE.

References

[Agirre *et al.*, 2009] Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proc. Human Language Technologies: 2009 Annual Conf. of the North Ameri-*

can Chapter of ACL (NAACL-HLT), pages 19–27, Stroudsburg, PA, USA, 2009. ACL.

[Bezdek, 1981] James C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Kluwer Academic Publishers, Norwell, MA, USA, 1981.

[Borin and Forsberg, 2010] Lars Borin and Markus Forsberg. From the people's synonym dictionary to fuzzy synsets - first steps. In *Proc. LREC 2010 workshop Semantic relations. Theory and Applications.*, pages 18–25, Malta, 2010.

[Chodorow *et al.*, 1985] M. S. Chodorow, R. J. Byrd, and G. E. Heidorn. Extracting semantic hierarchies from a large on-line dictionary. In *Proc. 23rd annual meeting of Association for Computational Linguistics (ACL)*, pages 299–304, Morristown, NJ, USA, 1985. ACL.

[Dolan, 1994] William B. Dolan. Word sense ambiguation: clustering related senses. In *Proc. 15th Conf. on Computational Linguistics (COLING)*, pages 712–716, Morristown, NJ, USA, 1994. ACL.

[Dorow, 2006] Beate Dorow. *A Graph Model for Words and their Meanings*. PhD thesis, Institut fur Maschinelle Sprachverarbeitung der Universitat Stuttgart, 2006.

[Fellbaum, 1998] Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, May 1998.

[Gfeller *et al.*, 2005] David Gfeller, Jean-Cédric Chappelier, and Paulo De Los Rios. Synonym Dictionary Improvement through Markov Clustering and Clustering Stability. In *Proc. Intl. Symposium on Applied Stochastic Models and Data Analysis (ASMDA)*, pages 106–113, 2005.

[Gomes *et al.*, 2003] Paulo Gomes, Francisco C. Pereira, Paulo Paiva, Nuno Seco, Paulo Carreiro, José Luís Ferreira, and Carlos Bento. Noun sense disambiguation with wordnet for software design retrieval. In *Proc. Advances in Artificial Intelligence, 16th Conf. of the Canadian Society for Computational Studies of Intelligence*, pages 537–543, Halifax, Canada, 2003.

[Gonçalo Oliveira and Gomes, 2010] Hugo Gonçalo Oliveira and Paulo Gomes. Onto.PT: Automatic Construction of a Lexical Ontology for Portuguese. In *Proceedings of 5th European Starting AI Researcher Symposium (STAIRS 2010)*. IOS, 2010.

[Gonçalo Oliveira *et al.*, 2010] Hugo Gonçalo Oliveira, Diana Santos, and Paulo Gomes. Extracção de relações semânticas entre palavras a partir de um dicionário: o PAPEL e sua avaliação. *Linguamática*, 2(1):77–93, 2010.

[Hirst, 2004] Graeme Hirst. Ontology and the lexicon. In Steffen Staab and Rudi Studer, editors, *Handbook on Ontologies*, International Handbooks on Information Systems, pages 209–230. Springer, 2004.

[Ide and Veronis, 1995] Nancy Ide and Jean Veronis. Knowledge extraction from machine-readable dictionaries: An evaluation. In *Machine Translation and the Lexicon*, number 898 in LNAI, pages 19–34. Springer, 1995.

[Kilgarriff, 1997] Adam Kilgarriff. "I don't believe in word senses". *Computing and the Humanities*, 31(2):91–113, 1997.

[Lin and Pantel, 2002] Dekang Lin and Patrick Pantel. Concept discovery from text. In *Proc. 19th Intl. Conf. on Computational Linguistics (COLING)*, pages 577–583, 2002.

[Maziero *et al.*, 2008] Erick G. Maziero, Thiago A. S. Pardo, Ariani Di Felippo, and Bento C. 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 da Linguagem Humana (TIL)*, pages 390–392, 2008.

[Navarro *et al.*, 2009] Emmanuel Navarro, Franck Sajous, Bruno Gaume, Laurent Prévot, ShuKai Hsieh, Tzu Y. Kuo, Pierre Magistry, and Chu R. Huang. Wiktionary and nlp: Improving synonymy networks. In *Proc. 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, pages 19–27, Suntec, Singapore, 2009. ACL.

[Pasca and Harabagiu, 2001] Marius Pasca and Sanda M. Harabagiu. The informative role of WordNet in open-domain question answering. In *Proc. NAACL 2001 Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, pages 138–143, Pittsburgh, USA, 2001.

[Peters *et al.*, 1998] Wim Peters, Ivonne Peters, and Piek Vossen. Automatic sense clustering in EuroWordnet. In *Proc. 1st Intl. Conf. on Language Resources and Evaluation (LREC)*, pages 409–416, Granada, May 1998.

[Santos and Bick, 2000] Diana Santos and Eckhard Bick. Providing Internet access to Portuguese corpora: the AC/DC project. In *Proc. 2nd Intl. Conf. on Language Resources and Evaluation (LREC)*, pages 205–210, 2000.

[Simões and Farinha, 2010] A. Simões and R. Farinha. Dicionário Aberto: Um novo recurso para PLN. *Vice-Versa*, September 2010.

[Simões *et al.*, 2010] Alberto Simões, José João Almeida, and Rita Farinha. Processing and extracting data from dicionário aberto. In *Proc. Intl. Conf. on Language Resources and Evaluation (LREC)*, Malta, 2010.

[Teixeira *et al.*, 2010] J. Teixeira, L. Sarmento, and E. Oliveira. Comparing verb synonym resources for portuguese. In *Computational Processing of the Portuguese Language, 9th Intl. Conf. Proc. (PROPOR)*, pages 100–109, 2010.

[Turney, 2001] Peter D. Turney. Mining the web for synonyms: PMI–IR versus LSA on TOEFL. In *Proc. 12th European Conf. on Machine Learning (ECML)*, volume 2167 of *LNCS*, pages 491–502. Springer, 2001.

[Velldal, 2005] Erik Velldal. A fuzzy clustering approach to word sense discrimination. In *Proc. 7th Intl. Conf. on Terminology and Knowledge Engineering*, Copenhagen, Denmark, 2005.