• U  C •

PhD Thesis
Doctoral Program in Information Science and Technology
Artificial Intelligence

Onto.PT: Towards the Automatic Construction of a Lexical Ontology for Portuguese

Hugo Gonçalo Oliveira

Thesis Advisor
Paulo Jorge Sousa Gomes

Department of Informatics Engineering
Faculty of Sciences and Technology
University of Coimbra

September 2012

This thesis was submitted, on 21^{st} September 2012, to the University of Coimbra, for the fulfillment of the requirements for obtaining the degree of Doctor of Philosophy in Science and Information Technology.

The argument of this thesis took place on 27^{th} May 2013, in Sala dos Capelos, University of Coimbra. The candidate was unanimously approved with distinction and honors by the following jury:

- Fernando Amílcar Bandeira Cardoso, Full Professor at Faculdade de Ciências e Tecnologia, University of Coimbra, Portugal (Chair)
- Roberto Navigli, Associate Professor at the Department of Computer Science, Sapienza University of Rome, Italy
- Diana Santos, Professor at Faculty of Humanities, University of Oslo, Norway
- Maria Luísa Torres Ribeiro Marques da Silva Coheur, Assistant Professor at Instituto Superior Técnico, Technical University of Lisbon, Portugal
- Ana Alexandra Ribeiro Luís, Assistant Professor at the Faculdade de Letras, University of Coimbra, Portugal
- Paulo Jorge Sousa Gomes, Assistant Professor at Faculdade de Ciências e Tecnologia, University of Coimbra, Portugal

Acknowledgements

Although I am the only one signing this thesis, the completion of the work described here would have not been possible without many people, who helped me both scientifically and personally.

I would like to start by thanking my supervisor, Paulo Gomes, who was there every week for listening to me, for giving me his expert opinion on each detail of this work, and for encouraging me to go further.

The Master's works of Letícia Anton Pérez and Hernani Costa were key to the development of this work, so I would like to give them a special thanks for their effort, for the long and very interesting discussions, and for believing in my ideas.

I would like to thank everybody in the KIS laboratory, because they were somehow important for the development of this work. Special remarks to Ricardo Rodrigues, with whom I worked together on Págico and who developed the lemmatiser and the API for the POS-tagger used in this work; also to Bruno Antunes, for keeping everything running at the lab.

I would like to thank Alberto Simões, for the support and interactions about Dicionário Aberto; and Rui Correia and Nuno Mamede for providing me the list of cloze questions used to evaluate Onto.PT indirectly.

I would also like to thank all the reviewers involved in the manual evaluation of one or more parts of this work and who have not yet been mention, namely, Luísa Ponte, Paulo Santos, Ana Oliveira Alves, Filipe Rodrigues, Marisa Figueiredo, Jorge Santos, Diogo Silva, Raul Sá and Joel Cordeiro. Also Jorge Ávila, for proofreading this thesis.

My PhD work was financially supported by the FCT scholarship grant with the reference SFRH/BD/44955/2008, co-funded by FSE. But I would like to thank CISUC as well, for hosting me as their PhD student, and for providing me most services a PhD student needs, including the financial support for attending scientific events and present my work.

I would like to give a special word to Linguateca, and especially to Diana Santos, who always believed in me, even in the beginning, when my experience in NLP was limited. Without all the knowledge and enriching experience I acquired through Linguateca, I am absolutely sure that I would have never felt ready for a PhD.

More personally, this work would have not been possible without both the full support of Clara and of my mother, Margarida. They both take care of me so well, that they make it easier, even on those days when everything seems to be going the wrong way. To all my family, including those that did not make it to see me completing this work. And to everybody that I met along the last 5 years because, in some way, they have all contributed to setting my mood and were an additional source of inspiration for this work.

Preface

About six years ago, almost by accident, I ended up engaging in an academic research career. It all started with my Master's dissertation, the final, and probably the most important, stage of my Master's degree. Then, I was not planning to dedicate more than one year of my life to research. But even one year later, when I started working as a researcher for Linguateca, it was far from my thoughts that I would soon enroll on a PhD.

Briefly, the main goal of my Master's work was to, given a rhythmic sequence, generate matching lyrics, in Portuguese. My intention was always to work with my mother tongue – not only because I felt that the results would be more understandable and funnier for the people surrounding me, but also because I used to write a few Portuguese lyrics for my former band. I was thus very interested in investigating how far an automatic lyricist could go.

However, working with Portuguese revealed to be a challenging task. Since the beginning of the work, we noticed that there was a lack of language resources for Portuguese and it was not easy to find the few existing ones. For instance, at that time, we could not find a public comprehensive lexicon for providing words and information on their morphology and possible inflections. Not to mention a semantics-oriented lexicon. Since then, I decided I wanted to contribute with something useful, that would hopefully fulfill the aforementioned shortage of resources. More or less at the same time, I had my first contact with Linguateca, a distributed language resource centre for Portuguese, responsible not only for cataloguing existing resources, but also for developing and providing free access to them.

I was very lucky that, before the end of my Master's, Linguateca opened a position that I applied for. The main goal of this position was to develop PAPEL, a lexical-semantic resource for Portuguese, automatically extracted from a dictionary. After my Master's, I was hired for that precise task. While working for Linguateca, I started to have a deeper contact with other researchers working on the computational processing of Portuguese. I started to gain some experience on natural language processing (NLP), especially on semantic information extraction, and I became passionate for research in this area. So much that, today, I do not see myself doing something completely unrelated.

The work with Linguateca was very important for my training as a researcher in NLP. It was so enriching that I felt that, with what I had learned, I could do, and learn, more. And there is so much to do to contribute to the development of Portuguese NLP, that I wanted to continue my work, which I did, after embarking on my PhD. This thesis presents the result of a four year PhD where, starting with what we learned with PAPEL, we created a larger resource, Onto.PT, by exploiting other sources, and we developed a model for organising this resource in an alternative way, which might suit better concept-oriented NLP.

Abstract

The existence of a broad-coverage lexical-semantic knowledge base has a positive impact on the computational processing of its target language. This is the case of Princeton WordNet, for English, which has been used in a wide range of natural language processing (NLP) tasks. WordNet is, however, created manually by experts. So, despite ensuring highly reliable contents, its creation is expensive, time-consuming and has negative consequences on the resource coverage and growth.

For Portuguese, there are several lexical-semantic knowledge bases, but none of them is as successful as WordNet is for English. Moreover, all of them have limitations, that go from not handling ambiguity at the word level and having limited coverage (e.g. only nouns, or synonymy relations) to availability restrictions.

Having this in mind, we have set the final goal of this research to the automatic construction of Onto.PT, a lexical ontology for Portuguese, structured in a similar fashion to WordNet. Onto.PT contains synsets – groups of synonymous words which are lexicalisations of a concept – and semantic relations, held between synsets. For this purpose, we took advantage of information extraction techniques and focused on the development of computational tools for the acquisition and organisation of lexical-semantic knowledge from text.

Our work starts by exploring textual sources for the extraction of relations, connecting lexical items according to their possible senses. Dictionaries were our first choice, because they are structured in words and meanings, and cover a large part of the lexicon. But, as natural language is ambiguous, a lexical item, identified by its orthographical form, is sometimes not enough to denote a concept. Therefore, in a second step, we use a synset-based thesaurus for Portuguese as a starting point. The synsets of this thesaurus are augmented with new synonyms acquired in the first step, and new synsets are discovered from the remaining synonymy relations, after the identification of word clusters. In the last step, the whole set of extracted relations is exploited for attaching the arguments of the non-synonymy relations to the most suitable synsets available.

In this thesis, we describe each of the aforementioned steps and present the results they produce for Portuguese, together with their evaluation. Each step is a contribution to the automatic creation and enrichment of lexical-semantic knowledge bases, and results in a new resource, namely: a lexical network; a fuzzy and a simple thesaurus; and Onto.PT, a wordnet-like lexical ontology. An overview of the current version of Onto.PT is also provided, together with some scenarios where it may be useful. This resource, which can be further augmented, is freely available for download and can be used in a wide range of NLP tasks for Portuguese, as WordNet is for English. Despite the current limitations of an automatic creation approach, we believe that Onto.PT will contribute for advancing the state-of-the-art of the computational processing of Portuguese.

Resumo

Não há grandes dúvidas que a existência de uma base de conhecimento léxico-semântico de grande cobertura tem um impacto positivo no processamento computacional da língua a que é dedicada. É isto que acontece com a WordNet de Princeton, para o inglês que, desde a sua criação, tem sido utilizada num amplo leque de tarefas ligadas ao processamento de linguagem natural. No entanto, a WordNet é um recurso criado manualmente, por especialistas. Assim, apesar de se garantir um recurso altamente confiável, a sua criação é dispendiosa e morosa, o que se reflecte ao nível da cobertura e crescimento do recurso.

Para o português, existem várias bases de conhecimento léxico-semanântico, sem que, no entanto, nenhuma tenha alcançado o sucesso que a WordNet teve para o inglês. Além disso, todos os recursos anteriores têm limitações, tais como não lidarem com diferentes sentidos da mesma palavra ou terem uma cobertura limitada (p.e. apenas substantivos ou relações de sinonímia) até restrições ao nível da sua disponibilização e utilização.

Desta forma, definimos como o principal objectivo desta investigação a construção automática do Onto.PT, uma ontologia lexical para o português, estruturada de forma semelhante à WordNet. A Onto.PT contém *synsets* – grupos de palavras sinónimas que são lexicalizações de um conceito – e relações semânticas, entre *synsets*. Para tal, tiramos partido de técnicas de extracção de informação e focámo-nos no desenvolvimento de ferramentas computacionais para a extracção e organização de conhecimento lexico-semântico, com base em informação textual.

Começamos por explorar recursos textuais para a obtenção de relações, que ligam itens lexicais de acordo com os seus possíveis sentidos. Os dicionários foram a nossa primeira escolha, por se encontrarem estruturados em palavras e significados, e também por cobrirem uma parte considerável do léxico. Mas como a língua é ambígua, um simples item lexical, identificado pela sua forma ortográfica, é muitas vezes insuficiente para referir um conceito. Por isso, num segundo passo, utilizamos como ponto de partida um tesauro baseado em *synsets*, e criado manualmente para o português. Os *synsets* desse tesauro são aumentados com novos sinónimos obtidos no primeiro passo, e novos *synsets* são descobertos através da identificação de agrupamentos de palavras (vulgo *clusters*) nas relações de sinonímia que sobram. No último passo, tiramos partido de todas as relações extraídas para associar os argumentos de cada relação ao *synset* mais adequado, tendo em conta o sentido do argumento envolvido na relação.

Nesta tese, descrevemos cada um dos passos anteriores, e apresentamos os resultados obtidos, juntamente com a sua avaliação, quando aplicados para o português. Cada passo é uma contribuição para a construção e enriquecimento automáticos de bases de conhecimento léxico-semântico, e resulta num novo recurso, nomeadamente: uma rede lexical; um tesauro baseado em *synsets* difusos e um tesauro simples; e o

Onto.PT, uma ontologia lexical, estruturada de forma semelhante a uma *wordnet*. Além disso, fornecemos uma visão global da versão actual do Onto.PT e apresentamos alguns cenários onde este recurso pode ter grande utilidade. O Onto.PT, que poderá futuramente ser aumentado, pode ser descarregado livremente e utilizado num grande leque de tarefas relacionadas com o processamento computacional do português, tal como a WordNet é para o inglês. Acreditamos que, apesar das limitações actuais de uma abordagem automática para a sua construção, o Onto.PT poderá contribuir para um avanço no estado da arte do processamento computacional da nossa língua.

Contents

List of Figures

List of Tables

Glossary

- **AI**: Artificial Intelligence
- **CBC**: Clustering By Committee
- **ECO**: Proposed approach for creating wordnets automatically by Extraction, Clustering and Ontologising
- **IE**: Information Extraction
- **IR**: Information Retrieval
- **LKB**: Lexical Knowledge Base
- **LSA**: Latent Semantic Analysis
- **LSIE**: Lexical-Semantic Information Extraction
- **NER**: Named Entity Recognition
- **NLP**: Natural Language Processing
- **OIE**: Open Information Extraction
- **OT.PT**: OpenThesaurus.PT
- **PMI**: Pointwise Mutual Information
- **POS**: Part of Speech
- **sb-triple**: synset-based relational triple, a relational triple whose arguments are synsets
- **synpair**: synonymy pair, a pair of synonymous lexical items
- **tb-triple**: term-based relational triple, a relational triple whose arguments are lexical items
- **WSD**: Word Sense Disambiguation

Chapter 1

Introduction

A substantial amount of data produced every day is available in natural language text. Understanding its meaning involves more than recognising words and their interactions, and typically requires access to external sources of knowledge. This fact led to the creation of broad-coverage lexical-semantic resources, which can be exploited in natural language processing (NLP, Jurafsky and Martin (2009)) tasks that perform a semantic analysis of text. Given that they are structured in words and meanings, these knowledge bases are often referred to as lexical ontologies, because they have properties of a lexicon as well as properties of an ontology (Hirst, 2004; Prévot et al., 2010).

Although there are different interpretations of the concept of lexical ontology, and thus different models (see more in sections 2.2 and 3.1 of this thesis), the paradigmatic resource of this kind is the Princeton WordNet (Fellbaum, 1998). WordNet is structured in synsets – groups of synonymous words, which can be seen as possible lexicalisations of a natural language concept – and semantic relations connecting synsets, including hypernymy (a concept is a kind of another concept), part-of (a concept is part of another concept), and others.

Besides its structure, suitable for being integrated and exploited by NLP applications, the public availability of WordNet played an important role in its success. WordNet was widely accepted by the NLP community and, today, there is no doubt that the existence of such a resource has a positive impact on the computational processing of a language. This fact is evidenced for English, where WordNet has opened the range of capabilities of NLP applications. It was used in the achievement of tasks, including, but not limited to, determining similarities (Seco et al., 2004; Agirre et al., 2009a), word sense disambiguation (Resnik, 1995; Banerjee and Pedersen, 2002; Gomes et al., 2003; Agirre et al., 2009b), query expansion (Navigli and Velardi, 2003), information retrieval (Voorhees, 1998), intelligent search (Hemayati et al., 2007), question-answering (Pasca and Harabagiu, 2001; Clark et al., 2008), text summarisation (Bellare et al., 2004; Plaza et al., 2010), text categorisation (Elberrichi et al., 2006; Rosso et al., 2004), and sentiment analysis (Esuli and Sebastiani, 2007; Williams and Anand, 2009).

In addition to the resource, the WordNet model was also very successful and adopted in the creation of broad-coverage lexical-semantic resources for other languages, even though not all of those attempts resulted in public domain resources. Yet, as it happens for Princeton WordNet, a huge limitation of most wordnets is that they are manually developed from scratch. Their construction thus involves

too much human effort, which may be seen as a bottleneck for the development of the resource. Not to mention that, over time, language evolves with its users.

The truth is that as long as there is intensive labour involved in manually encoding lexical resources, lexical capabilities of NLP systems will be weak (Briscoe, 1991), and coverage limitations will always be present. The same happens for other kinds of knowledge base – handcrafting them is impractical and undesirable. We should therefore take advantage of available NLP tools in order to automate part of this task and reduce the need of manual input (Brewster and Wilks, 2004).

Having this in mind, especially before the establishment of WordNet, researchers studied how to automatise the task of acquiring lexical-semantic knowledge from text, with relative success. For instance, MindNet (Richardson et al., 1998) shown that it is possible to develop a lexical(-semantic) knowledge base (LKB) by automatic means.

Another common alternative to the manual creation of wordnets is the translation of a target wordnet (usually Princeton WordNet) to other languages (de Melo and Weikum, 2008). However, another problem arises because different languages represent different socio-cultural realities, they do not cover exactly the same part of the lexicon and, even where they seem to be common, several concepts are lexicalised differently (Hirst, 2004). Therefore, we believe that a wordnet for a language, whether created manually, semi-automatically or automatically, should be developed from scratch for that language.

As mentioned before, the manual creation of a knowledge base results in slow development and consequently in limited coverage, not only of lexical, but mostly on world knowledge. This is why, after the establishment of WordNet, researchers using this resource as their only knowledge base soon had to cope with information sparsity issues. So, apart from the work on the automatic construction of LKBs from scratch, there have been automatic attempts to enrich wordnets (e.g. Hearst (1998)) and also to link them with other knowledge bases (e.g. Gurevych et al. (2012) or Hoffart et al. (2011)), in order to create broader resources.

In the work described in this thesis, we look at the Portuguese scenario, and tackle the limitations of the LKBs for this language. We developed an automatic approach for the acquisition of lexical-semantic information and for the creation of lexical-semantic computational resources, dubbed ECO – Extraction, Clustering, Ontolgisation. The application of ECO to Portuguese resources resulted in a wordnet-like resource, dubbed Onto.PT. In the remaining of this chapter, we state the main goals of this research, briefly present the ECO approach, refer the main contributions of this work, and describe the structure of the rest of this thesis.

1.1 Research Goals

For the Portuguese language, there have been some attempts to create a wordnet or a related resource (see more in section 3.1.2), but all of them have one or more of the following main limitations:

- They are proprietary and unavailble, or their utilisation is not free;
- They are handcrafted, and thus suffer from limited coverage;

- They are not built for Portuguese from scratch, and thus have to deal with translation issues, and include problems as lexical gaps;
- They do not handle word senses, which might lead to inconsistencies regarding lexical ambiguity.

Looking at this scenario, we set our goal to the **development of computational tools for acquiring, structuring and integrating lexical-semantic knowledge from text**. Although some of these tools can be used independently, their development had in mind the exploitation of Portuguese resources and the aim of creating a new **lexical ontology for Portuguese**, where the aforementioned limitations were minimised. Consequently, the resulting resource would be:

- **Public domain** and thus free for being used by anyone, both in a research or in a commercial setting. We believe this is the best way for the resource to play its role in helping to advance the state-of-the-art of Portuguese NLP. Furthermore, a bigger community of users tends to provide important feedback, useful for improving the resource.
- **Created automatically**, which would be done by exploiting textual resources and other public LKBs, all **created from scratch for one or more variants of Portuguese**. An automatic construction enables the creation of larger and broader resources, in a trade-off for lower reliability, but still acceptable for most tasks.
- Structured according to the **wordnet model**. This option relied on the great acceptance of this model and on the wide range of algorithms that work over this kind of structure to achieve various NLP tasks.

1.2 Approach

Our flexible approach for the **acquisition, organisation and integration of lexical-semantic knowledge** involves three main automatic steps. Each step is independent of each other and can be used for the achievement of simpler tasks. Alternatively, their combination enables the integration of lexical-semantic knowledge from different heterogeneous sources and results in a wordnet-like ontology. The three steps are briefly described as follows:

1. **Extraction:** instances of semantic relations, held between lexical items, are automatically **extracted** from text. As long as the extracted instances are represented as triples (two items connected by a predicate), the extraction techniques used in this step do not affect the following steps. In the specific case of our work, we followed a pattern based extraction on dictionary definitions.
2. **Thesaurus enrichment and clustering:** if there is a conceptual base with synsets for the target language, its synsets are **augmented** with the extracted synonymy relations. For this purpose, the network established by all extracted synonymy instances (synpairs) is exploited for computing the similarities between each synset and synpair. Both elements of a synpair are then added to their most similar synset. As for synpairs with two lexical items not covered

by the conceptual base, they establish a smaller synonymy network. This network is finally exploited for the identification of word **clusters**, which can be seen as new synsets.

3. **Ontologisation:** the lexical items in the arguments of the non-synonymy relation instances are **attached** to suitable synsets. Once again, this is achieved by exploiting the network established by all extracted relations, in order to, given a relation instance, select the most similar pair of candidate synsets.

As the resulting resource is structured in synsets and semantic relations between them, it can be seen as a wordnet. Given the three aforementioned steps, this approach for creating wordnets automatically was baptised as **ECO**, which stands for **E**xtraction, **C**lustering and **O**ntologisation.

1.3 Contributions

Given our main goal, **Onto.PT** can be seen as the main contribution of this research. Onto.PT is a wordnet-like lexical ontology for Portuguese, whose current version integrates lexical-semantic knowledge from five lexical resources, more precisely three dictionaries and two thesauri. Actually, after noticing that most of the Portuguese lexical resources were somehow complementary (Santos et al., 2010; Teixeira et al., 2010), we integrated in Onto.PT those that were public.

The current version of Onto.PT contains more than 100,000 synsets and more than 170,000 labelled connections, which represent semantic relations. This new resource is a public alternative to existing Portuguese LKBs and can be used as a wordnet. This means that, for Portuguese, Onto.PT can be used in most NLP tasks that exploit the structure of a wordnet for achieving their goal, except for those that use the synset glosses, unavailable in Onto.PT.

But Onto.PT is not a static resource. It is created in a three step **flexible approach**, ECO, briefly described in the previous section. ECO enables the integration of lexical-semantic knowledge from different heterogeneous sources, and can be used to create different instances of the resource, using different parameters. Moreover, although applied only to the creation of Onto.PT, we propose ECO as an approach that may be adopted in the creation or enrichment of wordnets in other languages. It is thus another important contribution of this thesis.

Each step of ECO can also be individually seen as contribution to the fields of information extraction and automatic creation of wordnets. These steps include procedures for:

1. Enriching an existing thesaurus with new synonymys.
2. Discovering synsets (or fuzzy synsets) from dictionary definitions.
3. Moving from term-based to synset-based semantic relations, without accessing the extraction context.

On the other hand, the procedure for extracting semantic relations from dictionaries cannot be seen as novel. Still, we have compared the structure and contents in different dictionaries of Portuguese, which led to the conclusion that **many regularities are kept across the definitions of each dictionary**. This comparison,

which can be seen as another contribution of this thesis, enabled us to use the same grammars for extracting information from three dictionaries.

Together with Onto.PT, other lexical-semantic resources were developed by using each of the ECO steps independently. These resources, listed below, contribute for advancing the state-of-the-art of Portuguese LKBs:

- **CARTÃO**, the largest term-based **lexical-semantic network** for Portuguese;
- **CLIP**, as far as we know, the first broad-coverage **fuzzy thesaurus** for Portuguese;
- **TRIP**, the largest public **synset-based thesaurus** for Portuguese.

We should add that most of the work performed during the course of this thesis is reported in a total of 14 scientific papers, presented and/or published in national and international venues, such as IJCAI, ECAI, EPIA or NLDB (see more in section 9.1).

1.4 Outline of the thesis

After two chapters on background knowledge and related work, each chapter of this thesis is focused on an automatic procedure that integrates the ECO approach and performs one step towards our final goal, the creation of Onto.PT. Besides describing each procedure, one or more experiments towards its validation are reported in each chapter. Before concluding, the last version of Onto.PT, available while this thesis was written, is presented together with examples of scenarios where it might be useful. In the end of the thesis, two appendices were included, namely: (A), with an extensive list of the semantic relations in Onto.PT and their description; and (B), which shows a rough manual mapping between the Onto.PT synsets and the core concepts of Princeton WordNet. We now describe each chapter, briefly:

Chapter 2 introduces (mostly) **theoretical background knowledge** that supports this research. It starts with some remarks on lexical semantics, the subfield of semantics that deals with words and meanings. Then, different formalisms for representing lexical-semantic computational resources are described. Given that our work is related to the NLP field of information extraction, the last section is dedicated to this topic.

Chapter 3 is about **concrete work** with some relation to our research. First, it describes well-known lexical knowledge bases for English, and also for Portuguese. Second, it presents work on information extraction from dictionaries and also from corpora. Third, work on the automatic enrichment or integration of existing knowledge bases is referred.

Chapter 4 explains our work towards the **acquisition of semantic relations from dictionaries**. As many regularities are kept across definitions in different dictionaries, we reuse existing handcrafted grammars, made for the extraction of semantic relations from one dictionary, for extracting relations from other dictionaries. This results in CARTÃO, a **large lexical network for Portuguese** that integrates knowledge from three different dictionaries.

Chapter 5 describes how synonymy networks extracted, for instance, from dictionaries, may be exploited in the **discovery of synsets**. To this end, a **clustering** procedure is ran on the previous networks and the discovered clusters are used as synsets. The clustering algorithm actually discovers fuzzy synsets, where a weight is associated to the membership of a word to a synset, in a more realistic representation of words and meanings. This part of the work led to the creation of the Portuguese **fuzzy thesaurus**, CLIP, completely extracted from dictionaries.

Chapter 6 presents an approach for **enriching the synsets of a thesaurus with synonymy relations extracted from dictionaries**. Given that there are freely available synset-based resources for Portuguese, we decided to exploit them in the creation of Onto.PT. Therefore, we used a public handcrafted thesaurus as the starting point for the creation of a broader synset-base. First, synonymy instances are assigned to the most similar synset. Then, the relations not assigned to a synset are the target of clustering, to discover new synsets, later added to the synset-base. This part of the work originated TRIP, a **large synset-based thesaurus** of Portuguese.

Chapter 7 proposes several algorithms for **moving from term-based semantic relations to relations held between the synsets** of a wordnet. After establishing the synset-base of Onto.PT, we still had a lexical network with semantic relations held between terms, and not synsets. Therefore, we developed and compared a set of algorithms that take advantage of the lexical network, and of the lexical items in the synsets, to select suitable synsets for attaching each argument of the lexical network's relations. Given a synset-base and a lexical network, the result of these algorithms is a **wordnet**.

Chapter 8 summarises the work described in the previous chapters, which may be combined in ECO, the automatic approach for creating Onto.PT, a new **lexical ontology** for Portuguese. An overview of this resource is first presented, together with some details on its availability, and on its evaluation. The last section suggests possible scenarios where Onto.PT might be useful.

Chapter 9 presents a final discussion on this research and highlights its main contributions. In the end, some cues are given for further improvements and additional work.

Chapter 2

Background Knowledge

The topic of Natural Language Processing (NLP), described extensively by Jurafsky and Martin (2009), is commonly presented with the help of pop-culture futuristic visions, where robots are capable of keeping a conversation with people, using human language. Those visions are typically impersonated by movie or television characters, such as HAL9000 in the Stanley Kubrick's classic *2001: A Space Odyssey*[1], or Bender and other robots in Matt Groening's *Futurama*[2].

NLP is a field of artificial intelligence (AI, Russell and Norvig (1995)) whose main purpose is to enable machines to understand the language of people and thus to communicate with us, in our own language, as if machines were a person themselves. Given that natural language, used by humans for communication, is probably the most natural way of encoding, transmitting and reasoning about knowledge, most knowledge repositories are in written form (Santos, 1992). Therefore, the emergence of the NLP field from AI is not surprising.

One of the main problems concerning natural language is that it differs from formal languages (e.g. programming languages) because, in the latter, each symbol has only one meaning while, in the former, a symbol may have different meanings, depending on the context where it is used. Ambiguity occurs when it is not possible to assign a single meaning to a form of communication, because it can be interpreted in more than one way.

In the work described in this thesis, several NLP techniques are applied in order to obtain language resources, structured in words, which can later be used in various NLP tasks. This chapter provides background knowledge that introduces two important topics for this thesis: lexical semantics, a subfield of NLP; and information extraction, a NLP task. The representation and organisation of lexical-semantic knowledge is also discussed, between the previous topics. In the end, we add some remarks in order to connect the described background knowledge with the work developed in the scope of this thesis. We decided to keep this chapter more theoretical, while the next chapter describes practical work, including existing lexical-semantic resources as well as works on information extraction from text.

[1] See `http://www.imdb.com/title/tt0062622/` (August 2012)
[2] See `http://www.imdb.com/title/tt0149460/` (August 2012)

2.1 Lexical Semantics

In theoretical linguistics, **morphology** deals with the identification, analysis and description of the structure of words, **syntax** deals with the study of structural relationships between words in a sentence, and **semantics** studies the meaning of language. In order to interpret words, phrases, sentences and texts, natural language is mapped to a formal language. Formal models of semantics are suitable, for instance, for being handled by machines.

Our research is especially focused on semantics, more precisely on the meaning of words. **Lexical semantics** is the subfield of semantics that studies the **words** of a language and their **meanings**, and can be seen as the bridge between a language and the knowledge expressed in that language (Sowa, 1999). Lexical semantics has been the subject of studies since the early nineteenth century, so it is no surprise that several theories on this topic have been developed (see Geeraerts (2010) for an extensive overview on theories of lexical semantics).

2.1.1 Relational Approaches

Among the theories of lexical semantics, the **relational approaches** are probably the most widespread (Geeraerts, 2010). They describe the meaning of words and expressions (hereafter, **lexical items**) through relationships. Some authors call them **lexical relations** (Cruse, 1986), others **semantic relations** (Murphy, 2003), and others **sense relations** (Jurafsky and Martin, 2009; Geeraerts, 2010), because these relationships are held between meanings of the lexical items, often called **senses**. In this thesis, we have adopted the term semantic relation.

Dictionaries are probably the main human-readable source of information on vocabulary and meaning of a language, as they are organised repositories of lexical items and information about their possible senses. Dictionaries are created by lexicographers, experts in the description of meanings in natural language. Sense definitions are often explained with reference to related lexical items (e.g. category, parts, synonyms), which (implicitly) indicate one or more semantic relations.

Nevertheless, even though dictionaries may be seen as a rough approximation to the way humans structure their knowledge about language, they are different from what is usually referred to as the **mental lexicon**. To begin with, in a dictionary, it is impossible to list extensively all the possible meanings of a lexical item, as the potential uses of a word are limitless (Nunberg, 1978). The mental lexicon is structured on **concepts**, which may be represented by words in the process of lexicalisation, but are defined by much more arbitrary rules, depending on each others' experiences. As word senses are not discrete and cannot be separated with clear boundaries, sense division in dictionaries is artificial (Kilgarriff, 1996; Hirst, 2004). Furthermore, while general language dictionaries contain mainly knowledge about the language itself, in the mental lexicon it is hard to make a clear distinction between the latter and knowledge about the world, commonly referred to as encyclopedic knowledge.

In order to increase the utility of the mental lexicon as an object of study and as a valuable resource, some theoretical approaches, and most computational, adopted a more structured and practical representation for this idea, where several simplifications are made. At the same time, these models make explicit some information

that, in dictionaries, is implicit or requires additional processing to be acquired. One common simplification is the representation of concepts by one, or several, words. Another, is the explicit representation of semantic relations by a connection between word senses or concepts, as opposed to plain textual definitions[3].

2.1.2 Semantic Relations

Semantic relations are important tools for describing the interactions between words and word senses and are intrinsic to the relational approaches to the mental lexicon. They are often present in computational approaches to the mental lexicon, including earlier representations, such as Quillian (1968)'s model, as well as those materialised in lexical-semantic resources, such as Princeton WordNet (Fellbaum, 1998) and others referred in section 2.2. In this section, we describe the semantic relations most extensively studied and most mentioned in the literature.

Synonymy

The synonymy relation holds among different word senses that have the same meaning, such as:

car synonym_of *automobile*

A more practical definition states that two lexical items are synonyms if, in a sentence, we can substitute one for another without changing both the meaning and the acceptability of the sentence. This definition is however debatable, as some researchers claim that, according to it, there are no real synonyms, just near-synonyms (Edmonds and Hirst, 2002).

Antonymy

Antonymy is a relation between word senses with opposite meanings. Antonyms arise when the senses are complementary, contrary, or converse (Murphy, 2003), respectively as in:

dead antonym_of *alive*
hot antonym_of *cold*
buy antonym_of *sell*

Homonymy and polysemy

Homonymy occurs when lexical items have the same form but different meanings, as in:

bank.1: sloping land.
bank.2: financial institution.

When the meanings of two homonyms are intimately related, they are usually considered to be a single lexical item with different senses. In this case, the relation between the related senses is called polysemy (Pustejovsky and Boguraev, 1996). For instance, there is a sense of *bank* that derives from the presented *bank*.2 sense:

[3]Nevertheless, besides semantic relations, word senses in some computational resources are complemented by a textual definition.

bank.3: building where a financial institution offers services.

This is also what happens between the word referring to a person native of some country and the word referring to the language spoken in that country, as in:

Portuguese: native of Portugal.
Portuguese: the language spoken in Portugal.

Although the distinction between homonymy and polysemy is not always clear, the etymology of the lexical items and their conception by native speakers are both typically considered to define how related the items are.

The process of identifying which overall sense of a word is being used in a sentence is called word sense disambiguation (WSD, Nancy Ide (1998); Navigli (2009b)). WSD is however very dependent on the purpose (Wilks, 2000) because, as referred earlier, word senses are not discrete, and it is impossible to list all the meanings of a lexical item. Furthermore, sense division is not straightforward nor consensual – even dictionaries cannot be seen as the ultimate truth, as different lexicographers, or system developers, divide senses differently (Kilgarriff, 1996; Peters et al., 1998).

Hyponymy and hypernymy

When a concept is a subclass or a specific kind of another, we are in the presence of a hyponymy relation, sometimes also called the *is-a* relation. Hypernymy is the inverse relation of hyponymy. See, for instance:

dog hyponym_of *mammal*
mammal hypernym_of *dog*

In other words, a hyponym is a specification of its hypernym and inherits all of its properties. A true meaningful sentence should remain true if a concept is replaced by its hyponym, but it might not remain true if the concept is replaced by its hypernym. See an example in table 2.1, given that:

animal hypernym_of *mammal*

Sentence	**Value**
Mammals are warm-blooded vertebrates covered in hair or fur.	True
Dogs are warm-blooded vertebrates covered in hair or fur.	True
Animals are warm-blooded vertebrates covered in hair or fur	Not all animals

Table 2.1: Replacement of hyponyms and hypernyms.

Hypernymy and hyponymy relations can be used to organise concepts hierarchically, in taxonomies, as referred in section 2.2.3.

Meronymy and holonymy

When a concept is a part, a piece, a member, or a substance of another, a meronymy, or *part-of*, relation holds between them. In the opposite direction, a holonym is the whole that owns or has the part, as in:

wheel meronym_of *car*
car holonym_of *wheel*

There are more restrictive definitions of meronymy, as in Cruse (1986), and others more inclusive, as in Winston et al. (1987). Given that the physical relation between parts and whole vary, the latter authors consider six types of meronymy, considering the properties of functionality, homeomery[4] and separability. The following examples illustrate each of the six proposed types:

pedal component_of *bicyle*
card member_of *deck*
slice portion_of *pie*
flour stuff_of *cake*
swallowing feature_of *eating*
London place_of *England*

Besides hyponymy, it is also possible to build meronymy taxonomies.

Other semantic relations

All the relations presented above are examples of **paradigmatic relations** (Murphy, 2003). First, they are held between arguments of the same grammatical category. Second, their arguments establish a semantic paradigm, such as lexical items that may have the same meaning, or senses that have some characteristics in common but do not share others. On the other hand, **syntagmatic relations** occur between lexical items that go together in a syntactic structure (Murphy, 2003).

It is possible to think of a huge amount of semantic relations. The following, which include not only paradigmatic, but also syntagmatic relations, are examples of other semantic relations that might be useful for studying the meaning of language:

- **Causation:** one concept is caused by another:

  virus causation_of *disease*

- **Purpose:** one concept is the purpose of another:

  to_seat purpose_of *bench*

- **Property:** one concept has a certain property:

  green property_of *lettuce*

- **Manner:** one concept may be performed in a certain manner:

  quickly manner_of *to_walk*

[4] In a homeomerous meronymy relation, the part is the same kind of thing as the whole, as in {*slice* part-of *pie*} (Murphy, 2003)

2.2 Lexical Knowledge Formalisms and Resources

Language **dictionaries** are probably the primary source of lexical semantic knowledge of a language, and are of great utility for humans to acquire information on words, their possible uses and meanings. However, they are **not ready** for being used as computational lexical-semantic resources, because they do not contain explicit knowledge on the meaning of words. Figure 2.1 is an example of an entry of the Longman Dictionary of Contemporary English (LDOCE)[5].

dictionary, noun
dic-tion-a-ry, plural dictionaries [countable]

1. book that gives a list of words in alphabetical order and explains their meanings in the same language, or another language:
   [usage] a German - English dictionary
2. a book that explains the words and phrases used in a particular subject:
   [usage] a science dictionary

Figure 2.1: Entry for the word *dictionary* in the LDOCE.

Dictionaries are structured on word senses, and typically contain other kinds of information on lexical items, as their etymologies, part-of-speech, syllabic division, domain and usage examples. But they describe senses in natural language, which may sometimes be vague or ambiguous. Therefore, in order to be used by computational applications, they need to be parsed and the lexical-semantic knowledge in their definitions needs to be formalised.

In the rest of this section, we will introduce three basic formalisms for representing meaning. Then, we will describe some of the most common structures used in the development of computational lexical-semantic resources, typically known as lexical knowledge bases (LKBs).

2.2.1 Representation of Meaning

In opposition to dictionaries, there are **formal representations** for the meaning of words and sentences, following the relational approach to lexical semantics. The most popular abstractions for representing meaning in a formal language are **logical predicates** (Smullyan, 1995), **directed graphs** (Harary et al., 1965) and **semantic frames** (Fillmore, 1982). These formalisms are in many ways equivalent and can often be converted to one another. Figures 2.2, 2.3 and 2.4 illustrate, respectively, the three formalisms, by representing the meaning of the following sentences:

The bottle contains wine. Wine is a beverage.

The basic formalisms for representing meaning can be combined, reorganised and evolve into other, eventually more complex, structures, for representing the meaning of a (natural) language, in an abstract model of the mental lexicon. When materialised into LKBs, including **thesauri**, **taxonomies** and **lexical ontologies**,

[5]LDOCE is available for online queries from `http://www.ldoceonline.com/` (September 2012)

```
contains(bottle, wine)
is-a(wine, beverage)
```

Figure 2.2: Meaning representation 1: logical predicates.

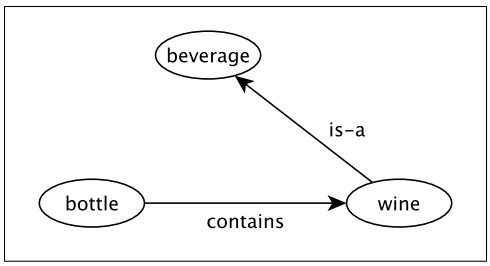


Figure 2.3: Meaning representation 2: directed graph.

```
wine
  is-a:  beverage
  container:  bottle
```

Figure 2.4: Meaning representation 3: semantic frame.

these abstractions may be used together with computational applications that, in order to achieve their goal, need to access lexical-semantic knowledge.

2.2.2 Thesauri

A **thesaurus** is a resource structured on the synonymy relation, as it groups lexical items together with their synonyms (or, according to the point of view, near synonyms (Edmonds and Hirst, 2002)). As a lexical item may have more than one meaning, it can be included in more than one group. Each inclusion of a lexical item in a group may thus be seen as a word sense, which means that ambiguous words will be included in more than one group. Moreover, a group of lexical items may be seen as the possible lexicalisations of a natural language concept.

Roget's Thesaurus (Roget, 1852), whose first edition is more than 150 years old, is the first ever and a widely-used English thesaurus. For Portuguese, TeP (Dias-Da-Silva and de Moraes, 2003; Maziero et al., 2008) is an electronic thesaurus for the Brazilian variant, created manually. In order to illustrate how the entries of a thesaurus may look like, figure 2.5 shows some of the entries for the word *thesaurus* in the online service *Thesaurus.com*[6]. Figure 2.6 shows the entries for the word *canto* in TeP. This word can either refer to a corner, or to a song.

Even though a thesaurus might represent other semantic relations besides synonymy, in the context of this thesis, the term thesaurus will be used to describe a resource that simply groups synonymous lexical items.

2.2.3 Lexical Networks

Lexical networks are graph structures, $N = (V, E)$, with $|V|$ nodes and $|E|$ edges, $E \subset V^2$, where each node $w_i \in V$ represents a lexical item and each edge between

[6]Available online from `http://thesaurus.reference.com` (August 2012)

- **Main Entry:** thesaurus
- **Part of Speech:** noun
- **Definitions:** dictionary of synonyms and antonyms
- **Synonyms:** glossary, lexicon, reference book, terminology, vocabulary, language reference book, onomasticon, sourcebook, storehouse of words, treasury of words, word list

---

- **Main Entry:** lexicon
- **Part of Speech:** noun
- **Definitions:** collection of word meanings, usage
- **Synonyms:** dictionary, glossary, terminology, thesaurus, vocabulary, word stock, word-book, wordlist

---

- **Main Entry:** vocabulary
- **Part of Speech:** noun
- **Definitions:** language of a person or people
- **Synonyms:** cant, dictionary, glossary, jargon, lexicon, palaver, phraseology, terminology, thesaurus, words, word-hoard, word-stock, wordbook

---

- **Main Entry:** reference book
- **Part of Speech:** noun
- **Definitions:** book of information
- **Synonyms:** almanac, dictionary, directory, encyclopedia, thesaurus, atlas, how-to book, source book, wordbook, work of reference

Figure 2.5: Some of the entries for the word *thesaurus* in Thesaurus.com.

canto (Substantivo)

1. canto, cantinho, recanto
2. canto, ponta
3. canto, ângulo, aresta, esquina, ponta, quina, rebarba, saliência

---

canto (Substantivo)

1. canto, música, som
2. canto, canção, melodia, poesia

Figure 2.6: Entries for the word *canto* in TeP.

nodes w_i and w_j, $E(w_i, w_j)$, indicates that lexical item w_i, or a sense of w_i, is related to lexical item w_j, or one sense of w_j.

In her thesis, Dorow (2006) works with lexical networks, extracted from corpora. Figure 2.7 is an example of such networks, where the edges represent co-occurrence of the connected nodes in the same sentence.

The edges of a lexical network can be directed and labelled according to the type of relation held by the two lexical items. Also, if the edges represent semantic

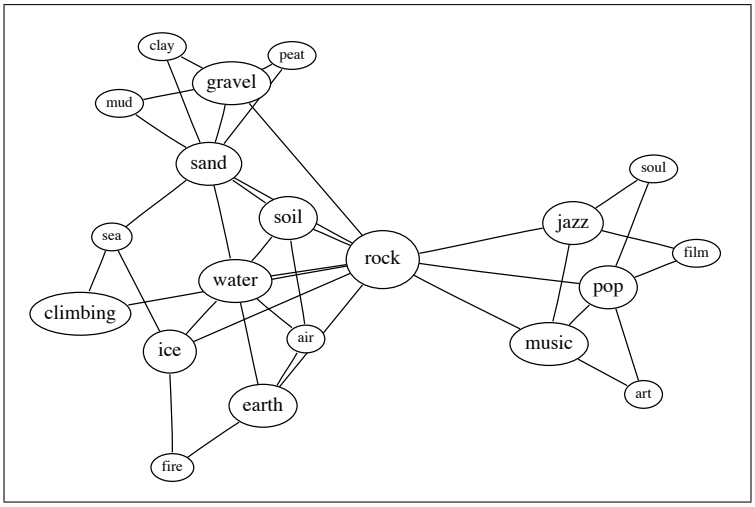


Figure 2.7: A term-based lexical network with the neighbours of the word *rock* in a corpus (from Dorow (2006)). This word might refer to a stone or to a kind of music.

relations, the nodes can be seen as word senses. In this case, an edge $E(w_i, w_j)$ indicates that one sense of w_i is related to one sense of w_j.

PAPEL (Gonçalo Oliveira et al., 2008, 2010b) and CARTÃO (Gonçalo Oliveira et al., 2011) are resources that may be used as term-based lexical networks for Portuguese. They are structured in relational triples automatically extracted from dictionary definitions. Each triple $t = \{w_1, R, w_2\}$ represents a semantic relation identified by R, which occurs between a sense of the lexical item w_1 and a sense of lexical item w_2. A lexical network is established if the arguments of the relational triples, w_1 and w_2, are used as nodes, connected by an edge labelled as R, the type of the relation. Figure 2.8 shows part of a term-based lexical network with relations.

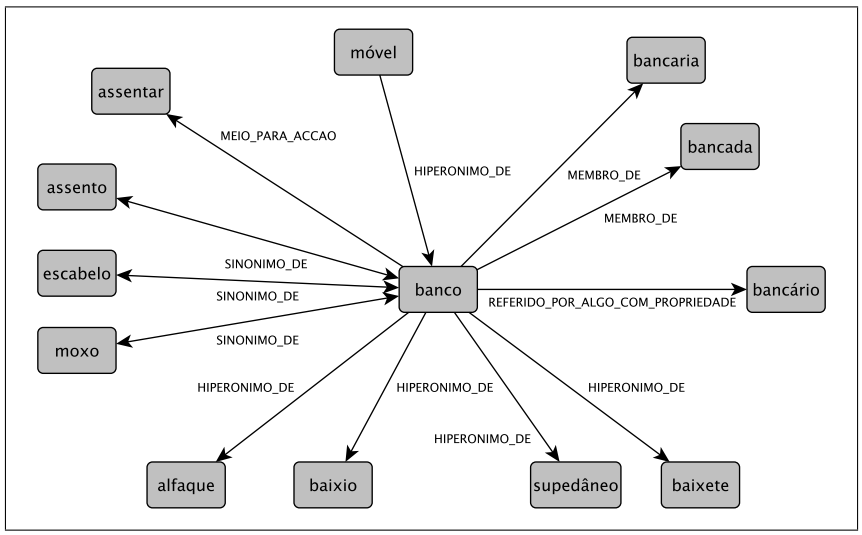


Figure 2.8: A term-based lexical network with relations where *banco* is one of the arguments (from PAPEL 2.0 (Gonçalo Oliveira et al., 2010b)). In Portuguese, besides other meanings, *banco* might refer to a bench/stool or to a financial institution.

Taxonomies

A **taxonomy** is a special kind of lexical network, structured according to certain rules. In a taxonomy, nodes represent concepts, eventually described by lexical items, and (directed) edges are relations connecting the former with their superordinates. It can be seen as a hierarchical tree where the higher nodes are more generic and the lower are more specific. In other words, a taxonomy is a classification of a certain group of entities, such as plants, academical degrees or musical genres, often used to represent hierarchical relations. For instance, in the case of hypernymy, each node inherits all the properties of its superordinate. Figure 2.9 illustrates this kind of network with a hypernymy taxonomy of animals.

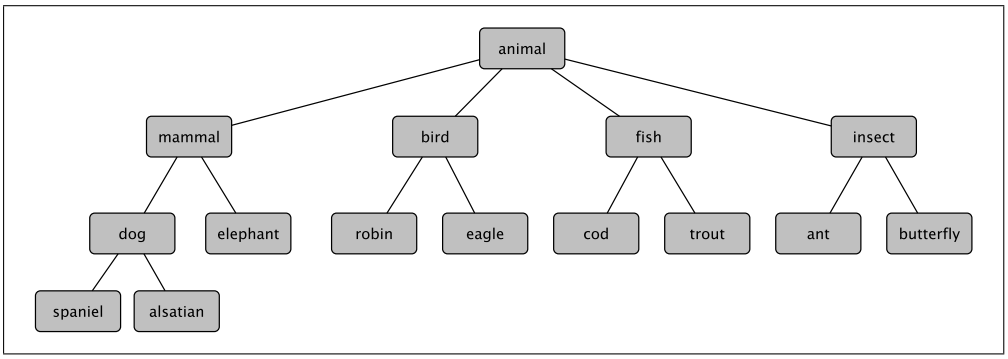


Figure 2.9: Example of a taxonomy of animals, adapted from Cruse (1986).

More on taxonomies can be found in Cruse (1986), where considerations about these structures and properties that they should hold are discussed and integrated in his view on lexical semantics. Furthermore, Smith (2001) lists the principles for taxonomy well-formedness, in the context of ontologies and information systems.

Conceptual graphs

Conceptual graphs (Sowa, 1992) are a formalism for expressing meaning, commonly used in topics that go from databases and expert systems to AI and NLP. Conceptual graphs typically represent first-order logic formulas. They contain two kinds of nodes: rectangular boxes and ovals, which denote, respectively, **concepts** and **relations**. Figure 2.10, shows a conceptual graph representation for the meaning of sentence:

John is going to Boston by bus.

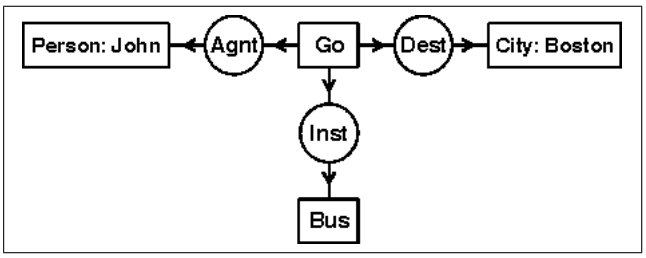


Figure 2.10: Conceptual graph, from `http://www.jfsowa.com` (September 2012).

Conceptual graphs are logically precise, humanly readable, and computationally tractable. Chein and Mugnier (2008) go further and refer that all kinds of knowledge can be seen as a labelled graph, as the they provide an intuitive and easily understandable means to represent knowledge. Therefore, most lexical networks can roughly be seen as conceptual graphs. For this purpose, the rectangles should denote the lexical nodes, while the ovals denote the labelled edges.

2.2.4 Lexical Ontologies

In the scope of computer science, despite some terminological issues, an **ontology** is commonly defined as an explicit specification of a shared conceptualisation (Gruber, 1993; Guarino, 1998). More precisely, ontologies are formalised models that typically represent knowledge of a specific domain. They are often structured on unambiguous concepts and on relations between them.

Given that information about the usage of words is irrelevant for ontologies, and that LKBs are more linguistic conventionalised than well-formed ontologies, it may seem that ontologies and LKBs are incompatible resources. The latter contain knowledge about a language and are structured in lexical items and on relations, including semantic relations. However, semantic relations are not held between words, but word meanings (word senses). Since meaning is inherently conceptual, LKBs can be seen as (lexical) ontologies with natural language concepts. **Lexical ontologies** are thus resources that have properties of a lexicon as well as properties of an ontology (Hirst, 2004; Prévot et al., 2010).

The so-called **ontolex interface** (Prévot et al., 2010) deals with the mapping between ontologies and linguistic knowledge, as represented in LKBs. This interface sees the LKB as a special ontology where, in order to represent concepts unambiguously, synonymous lexical items should be grouped under the same concept. A lexical ontology can be seen as a more formalised representation of a LKB. For instance, concepts can be seen as individuals in the ontology and can belong to classes; relations are well-defined interactions that may occur between classes or individuals; and relation instances are defined by rules that explicitly refer to the classes/individuals and to a pre-defined relation type.

A lexical ontology may be represented as a Semantic Web (Berners-Lee et al., 2001) model, using languages such as RDF (Miller and Manola, 2004) or OWL (McGuinness and van Harmelen, 2004), the W3C standards for representing information computationally as triples. An example of such a model is the W3C WordNet RDF/OWL (van Assem et al., 2006), where concepts are instances of classes that group synonymous word senses together (synsets). While the previous model deals only with linguistic knowledge, Lemon (Buitelaar et al., 2009) is a model for representing lexical information attached to ontologies. The main idea is that a lexical item may have several senses, and each one may be used to denote a concept in an ontology. Still on this context, the Lexical Markup Framework (LMF, Francopoulo et al. (2009)) is a ISO standard for representing lexicons, which covers not only the morphology and syntax, but also semantics, including words senses and semantic relations. The Lemon model can be converted to LMF.

2.2.5 The Generative Lexicon

Another important contribution to the representation of lexical-semantic knowledge is the theory of the Generative Lexicon (Pustejovsky, 1991), which accounts for the dynamic systematic polysemy of words in context. This theory argues that the word and its semantics influences heavily the compositionality mechanisms involved in explaining phenomena as synonymy, antonymy, metonymy[7] and others. According to this theory, lexical meaning is better captured by assuming the following levels of representation:

1. **Argument Structure**: the behavior of a word as a function.
2. **Event Structure**: identification of the particular event type for an expression.
3. **Qualia Structure**: the essential attributes of an object as defined by the lexical item.
4. **Inheritance Structure**: how the word is globally related to other concepts.

Qualia structures can be viewed as structured templates with semantic information that entails the compositional properties of each lexical item. In such a structure, the meaning of lexical items is described in terms of four roles, namely:

- **Constitutive**: the relation with its constituents or parts (e.g. material, weight, parts, components).
- **Formal**: aspects that distinguish it from others within a larger domain (e.g. orientation, magnitude, shape, dimensionality, color, position).
- **Telic**: its purpose or function (e.g. purpose for performing an act, aim that specifies certain activities).
- **Agentative**: what brings it into existence (e.g. creator, artifact, natural kind, causal chain).

This way, the Generative Lexicon only needs to store a single entry for every polysemous word for generating its appropriate sense in a context. Figure 2.11 is a qualia structure with the basic knowledge for the word *novel* – it is a narrative, typically in the form of a book, for the purpose of being read (whose event type is a transition T), and is an artifact created by the transition event of writing.

novel(x)			
	CONST	=	narrative(x)
	FORM	=	book(x), disk(x)
	TELIC	=	read(T, y, x)
	AGENT	=	artifact(x), write(T, z, x)

Figure 2.11: Qualia structure for the noun *novel* (from Pustejovsky (1991)).

The Brandeis Semantic Ontology (Pustejovsky et al., 2006) is a lexical ontology modelled after the generative lexicon theory.

[7]Metonymy is a figure of speech in which an object is not called by its own name, but by the name of something intimately associated with it. For instance, in the sentence "*The White House has launched a new website*", the website was not launched by the *White House* itself, but by someone working for the President of the USA, who lives in the *White House*.

2.3 Information Extraction from Text

Information extraction (IE, see Moens (2006) for an extensive overview) consists of the identification and classification of information in unstructured data sources, which this way becomes structured and ready, for instance, for populating a relational database and for being used directly by computational applications. In the specific case of IE from text, the target is text, written in natural language.

Although they are both solutions to the information overload problem, IE should not be confused with **information retrieval** (IR) (Baeza-Yates and Ribeiro-Neto, 1999), which is the task of locating required information within collections of data. In IR, the information to be searched is specified by a query, which can be, for instance, a group of keywords or a natural language question. The retrieved information is often a list of relevant documents, according to the query, which should thus contain the required information.

2.3.1 Tasks in Information Extraction from Text

According to Jurafsky and Martin (2009), a complete system for IE from text has typically four steps, where it performs the tasks of named entity recognition (NER), relation detection and classification, temporal event processing, and template filling:

1. **NER** (Chinchor and Robinson, 1997; Mota and Santos, 2008) is the task of identifying proper names mentioned in text. It can include the classification of the entities, which consists of attributing a category and, sometimes, a sub-category, to the entities, from a range including, but often not limited to, people, organizations and places. Moreover, as the entities are not always mentioned by the same name, and are sometimes referred by a pronoun, the task of NER might as well need to deal with coreference and anaphora resolution (Mitkov et al., 2000; Recasens et al., 2010). In our work, we are more interested in the identification of lexical entities, and not named entities.

2. **Relation detection and classification** (Hendrickx et al., 2010) is closely related to the scope of this thesis, and is the task of identifying semantic relations among the discovered entities, including, but not limited to, the ones presented in section 2.1.2. Semantic relations between named entities include, but are not limited to, family, employment or geospatial relations (see more in Freitas et al. (2009)).

3. As some of the relations might be true or false for different periods of time, it is sometimes important to determine when the events in the text happened. **Temporal event processing** (Verhagen et al., 2010) is related to the analysis of time expressions which include, for instance: mentions of the days of the week or months (e.g. *Sunday* or *February*), names of special days (e.g. *Christmas*, *Valentine's Day*), relative expressions (e.g. *in two months*, *next year*), clock and calendar times (e.g. *17:00 P.M.*, *2012-09-25*). This task is however out of the scope of this thesis.

4. **Template filling** is the task of searching for required data in documents that describe stereotypical information and then filling predefined slots with

the appropriate data. The slots can be, for instance, a table in a relational database.

2.3.2 Information Extraction Techniques

In order to increase the portability of IE systems, the development of techniques for IE is currently very centralised on machine learning (Moens, 2006). The alternative is to adopt approaches based on handcrafted knowledge. Existing IE techniques may be classified into three groups (Moens, 2006), namely: symbolic techniques, supervised machine learning and unsupervised machine learning.

All these techniques exploit **patterns** – recurring sequences of events, evidenced by the objects in text. **Pattern recognition** deals with the classification of objects into categories, according to the values of a selected number of their features. When using machine learning techniques, after selecting the features to be exploited, extraction rules are automatically learned from a collection of examples where the features are annotated. Alternatively, the features might be used for the manual construction of static rules. According to the types of features to explore, patterns in text can be classified into the following categories:

- **Lexical patterns**: features relative to the attributes of lexical items. For instance, if two lexical items co-occur in a context, or if a word is capitalised or not. The latter feature is especially productive for NER.
- **Syntactic patterns**: features include the part-of-speech (POS) of the lexical items (e.g. noun, verb, preposition), and the type of phrase (e.g. noun phrase, verb phrase, prepositional phrase).
- **Semantic patterns**: features that denote semantic classifications in information units. These include multi-word patterns that may be used for the extraction of semantic relations, such as *works at* (e.g. *Hugo works at CISUC*), for denoting an employment relation, or *is a* (e.g. *apple is a fruit*) for denoting hyponymy.
- **Discourse patterns**: features are values computed by using text fragments, such as the distance between two entities in a document. It is assumed that the distance is inversely proportional to the semantic relatedness.

In the rest of this section, we will briefly describe three groups of IE techniques.

Symbolic techniques

This group of IE techniques relies on handcrafted symbolic knowledge. Rules for IE are written by someone who is familiar with the formalisms of the IE systems and, especially, familiar with the data where the IE system will run on.

Symbolic techniques are often implemented through **partial parsing** and **finite state automata** (Partee et al., 1990). As the name suggests, partial parsing refers to situations when only part of the text is analysed, while the rest is skipped. The analysed part is antecipated through handcrafted patterns, which are in turn translated into the rules of a finite state automata.

Supervised learning

In supervised learning, the extraction rules are inferred automatically from a set of training data. These data consist of annotated examples, used to model the extraction process and then predict the correct handling of previously unseen examples.

Given that the set of examples is generally handcrafted and thus expensive to build, training data is usually limited. Therefore, it is important that this data is representative enough, which not always occurs, especially in natural language, where ambiguity and vagueness arise as problems, and where there is a large variety of patterns for expressing the same ideas.

Methods for supervised IE include, for instance, support vector machines (Cortes and Vapnik, 1995), maximum entropy models (Berger et al., 1996), hidden Markov models (Rabiner, 1989), or conditional random fields (Lafferty et al., 2001).

Unsupervised learning

Despite the success of supervised learning methods for IE (Moens, 2006), the cost of manual annotation is too high. While this approach might suit IE from closed and limited domains, alternatives were sough for situations when there is not enough annotated data. This problem is very challenging, especially in open-domain IE.

Having in mind that it is not hard to collect large quantities of unannotated data, unsupervised learning approaches for IE assume that it is possible to learn a classification without previous annotations, or to train classifiers with a small collection of annotated examples, the so-called seeds. This group of techniques includes **clustering**, which is completely unsupervised and relies only on unannotated data, and also **weakly supervised** approaches, where a classification is incrementally learned from a set of seeds.

Clustering techniques try to group objects together, such that objects in the same group are somehow similar, by sharing some of the selected features. There are several clustering models including, for instance, hierarchical clustering (e.g. single-linkage Sibson (1973) or complete-linkage Defays (1977)) or centroid based models (e.g. k-means (MacQueen, 1967) or fuzzy c-means (Bezdek, 1981)).

As for weakly-supervised approaches, they include self-training, co-training, and active learning. In **self-training** (McCallum et al., 1999), each iteration starts by learning one classifier from the set of seeds. Then, the unannotated data are annotated using the learned classifier. Finally, the examples where the confidence of the given annotation is higher than a threshold are added to the set of seeds. In **co-training** (Blum and Mitchell, 1998), two or more classifiers are learned from the same set of seeds, but each classifier is trained using a disjoint subset of features. At each iteration, unannotated data are annotated using the trained classifiers. Then, the examples where all classifiers agree are added to the set of seeds. Among the weakly supervised approaches, **active-learning** (Shen et al., 2004) is the one that requires more supervision, as all the training data is manually annotated. At each iteration, a set of examples is automatically selected. This set is then manually annotated and added to the training set. The selected examples tend to be those where the classifier is more uncertain.

With the argument that IE systems should be both portable and efficient enough to process huge amounts of data, as the Web, Banko et al. (2007) presented the paradigm of Open Information Extraction (OIE), a **self-supervised** approach to

IE. The input of a OIE system is a corpus and the output is a set of facts, represented as relational triples $t = \{e_1, relation\_phrase, e_2\}$. There is no need for annotated data nor need for specifying the relations to extract.

For learning a classifier, an OIE system starts by identifying the noun phrases of several thousands of sentences in the input corpus. The parsing structure of the words connecting noun phrases is also analysed. This sequence is labelled as positive or negative examples of trustworthy relations, according to predefined heuristics. Positive and negative tuples are finally used to establish triples, where a pair of noun phrases is connected by a relation phrase. Triples are mapped into feature vectors, used as the input of a classifier. For the extraction, only a single pass is needed over the input corpus. Each pair of noun phrases is used as the arguments of a triple, and the text connecting them is used as the relation phrase. Triples classified as trustworthy are extracted.

2.4 Remarks on this section

In this section, we bridge the theoretical work described in the previous sections with the work developed in the scope of this thesis. The first part targets the knowledge representation in our work and the second is about the information extraction techniques applied.

2.4.1 Knowledge representation in our work

In our work, instances of semantic relations are first extracted as relational triples $t = \{w_1, R, w_2\}$, which can both be seen as logical predicates or as the edges of a lexical network. The types of semantic relations are typical relations between word senses, including synonymy, hypernymy, several types of meronymy and most of the relations introduced in section 2.1.2. The arguments of these relations are lexical items, described by their orthographical form. Word senses are not handled.

On the other hand, the final resource of this work, Onto.PT, can be seen as lexical ontology, as we have adopted a model inspired by Princeton WordNet (see more about this resource in section 3.1.1). In order to represent natural language concepts, Onto.PT groups synonymous words in synsets, which are groups of synonymous words. This part of the resource can thus be seen as a thesaurus. As for other semantic relations, Onto.PT includes several predefined types established between synsets. Given that the presence of a lexical item in a synset defines a new possible sense of this item, different senses of the same word are recognised.

2.4.2 Information Extraction techniques in our work

We have only exploited dictionaries for the extraction of semantic relations. For this purpose, we used symbolic techniques over the dictionary definitions (see section 4).

We recall that dictionaries provide a wide coverage of the lexicon and they are structured in words and meanings. Moreover, definitions tend to use simple vocabulary and follow regularities, which makes most of them easily predictable. Therefore, after careful observation, we manually encoded a set of semantic patterns, organised in grammars, for processing them. Despite the manual labour involved in the manual creation of the grammars, we could take advantage of one of the pros of

using handcrafted knowledge over machine learning techniques – we have more control on the obtained results. But most of this work was done in the scope of the project PAPEL (Gonçalo Oliveira et al., 2008, 2010b). Given that the grammars for extracting semantic relations from one dictionary were available, the manual effort was minimised. Furthermore, as we will see in more detail in section 4, definitions follow similar regularities across different dictionaries, which gives some portability to the grammars. Other reasons for discarding machine learning techniques for this task include:

- As far as we know, there is no Portuguese dictionary (or corpus), with annotated semantic relations between lexical items, suitable, for instance, for supervised extraction approaches – the ReRelEM collection (Freitas et al., 2009) only contains relations between named entities. The creation of such a resource is more time-consuming than the manual creation of grammars.
- When it comes to the same relation instance, dictionary text has limited redundancy. So, in a few experiments that we have performed, weakly supervised bootstrapping techniques only discovered a small set of relations.
- OIE is more suitable for extracting open-domain relations from corpora, and not for extracting predefined relations. If OIE was applied, we would need to later convert the discovered predicates into one of our relation types. Moreover, the discovered relational triples typically connect generic world concepts, and not word senses.

For discovering synsets (see sections 5 and 6), we have used graph clustering techniques (Schaeffer, 2007) over the synonymy graph extracted from the dictionaries. Lexical items are grouped in synsets according the similarity of their adjacencies in the graph. Also, in order to represent ambiguity, the clusters might be overlapping. Finally, the integration of semantic relations in the thesaurus (see section 7) can also be seen as kind of clustering, as each argument of a triple is attached to most similar synset.

Chapter 3

Related Work

In this chapter, we present work related to ours. Since one of our main goals is the creation of a lexical ontology, we start by presenting the most popular lexical-semantic knowledge bases, including existing resources of this kind for Portuguese. Then, given that we acquire lexical-semantic knowledge from text, we describe work on information extraction from text. The last section of the chapter is dedicated to work on the enrichment of knowledge bases and on their integration.

3.1 Lexical Knowledge Bases

As the name suggests, **lexical knowledge bases** (LKBs) are organised repositories structured in lexical items. Both thesauri and term-based lexical networks are kinds of LKBs, but a typical knowledge base tends to contain additional information, and thus to be more complex than these resources. This should include semantic information, such as possible word senses, relations, definitions, or example sentences.

In the scope of this thesis, a LKB is defined as a broad-coverage resource, structured in lexical items and lexical-semantic relations, that tries to cover one whole language and not just a specific domain. For English, Princeton WordNet (Fellbaum, 1998) is the paradigmatic example of a LKB and its model is probably the most popular for representing such a resource. This section starts by introducing well-known LKBs, including WordNet, and proceeds by presenting existing Portuguese LKBs.

3.1.1 Popular Lexical Knowledge Bases

Besides Princeton WordNet, for English, there are other LKBs that are worth mentioning, namely MindNet, FrameNet and VerbNet. Moreover, common-sense knowledge bases are sometimes also used as LKBs, so we additionaly refer CyC and ConceptNet.

WordNet

Princeton WordNet (Miller, 1995; Fellbaum, 1998, 2010) is a handcrafted lexical resource for English based on **psycholinguistic** principles. It combines traditional lexicographic information with modern computation and may be used both as a dictionary and as a LKB. On the one hand, as in a thesaurus, WordNet is structured

in groups of synonymous lexical items (hereafter **synsets**), which are the possible lexicalisations of natural language concepts. While the synsets deal with synonymy, there are other types of lexical-semantic relations (e.g. hypernymy, part-of) between the former. On the other hand, each synset has a defined part-of-speech (POS), which indicates if the concept is a noun, verb or adjective; a gloss, which is similar to a dictionary definition; and usage example sentence(s). The inclusion of a lexical item in a synset is interpreted as a sense of that item.

Figure 3.1 contains the entries for the word *bird* in WordNet 3.0[1]. For this word, WordNet defines five different noun senses and one verb sense. Definitions follow each sense, in parenthesis, and the direct hyponyms of the first sense are expanded.

Noun

- bird (warm-blooded egg-laying vertebrates characterized by feathers and forelimbs modified as wings)
  [direct hyponym]
  - dickeybird, dickey-bird, dickybird, dicky-bird (small bird; adults talking to children sometimes use these words to refer to small birds)
  - cock (adult male bird)
  - hen (adult female bird)
  - nester (a bird that has built (or is building) a nest)
  - night bird (any bird associated with night: owl; nightingale; nighthawk; etc)
  - parrot (usually brightly colored zygodactyl tropical birds with short hooked beaks and the ability to mimic sounds)
  - ...
- bird, fowl (the flesh of a bird or fowl (wild or domestic) used as food)
- dame, doll, wench, skirt, chick, bird (informal terms for a (young) woman)
- boo, hoot, Bronx cheer, hiss, raspberry, razzing, razz, snort, bird (a cry or noise made to express displeasure or contempt)
- shuttlecock, bird, birdie, shuttle (badminton equipment consisting of a ball of cork or rubber with a crown of feathers)

Verb

- bird, birdwatch (watch and study birds in their natural habitat)

Figure 3.1: Synsets with the word *bird* in Princeton WordNet 3.0, and the first direct hyponyms of the first sense.

There has been some criticism on Princeton WordNet and some complex problems about this resource have been discussed (Sampson, 2000). For instance, it cannot be accepted as a scientific model of what it was initially supposed to be: a resource based on psycholinguistic principles. But it is undoubtedly a useful and highly complete resource, especially if we have in mind that it was created manually.

The WordNet model has been widely accepted by the NLP community. It is extensively used in NLP research and integrated in computational applications that

[1]WordNet 3.0 is downloadable through `http://wordnet.princeton.edu/wordnet/download/` (August 2012). WordNet 3.1 can be queried online, through `http://wordnetweb.princeton.edu/perl/webwn` (August 2012)

need to access lexical-semantic information. Also crucial to the huge popularity of Princeton WordNet, was its public domain character, which made it more accessible for any user willing to explore it or to develop new tools using this resource.

WordNet has been extensively used for performing various NLP and knowledge management tasks including, for instance, determination of similarities (Seco et al., 2004; Agirre et al., 2009a), word sense disambiguation (Resnik, 1995; Banerjee and Pedersen, 2002; Gomes et al., 2003; Agirre et al., 2009b)), question answering (Pasca and Harabagiu, 2001; Hovy et al., 2001; Clark et al., 2008), sentiment analysis (Williams and Anand, 2009; Esuli and Sebastiani, 2007), natural language generation (Jing, 1998; Hervás et al., 2006), intelligent search (Hemayati et al., 2007; Liu et al., 2004; Moldovan and Mihalcea, 2000), or text summarisation (Bellare et al., 2004; Plaza et al., 2010).

We can therefore consider wordnet as the paradigmatic model of a LKB. Besides Princeton WordNet, its success lead to the adaptation of the wordnet model to other languages[2], including the languages involved in multilingual wordnets, such as EuroWordNet (Vossen, 1997), MultiWordNet (Pianta et al., 2002) and BalkaNet (Stamou et al., 2002) projects. However, in opposition to Princeton WordNet, some of these resources, or part of them, are not freely available, even for research purposes.

MindNet

After working on that direction for several years (Dolan et al., 1993; Richardson et al., 1993; Vanderwende, 1994, 1995), **MindNet** (Richardson et al., 1998; Vanderwende et al., 2005) was presented as an independent LKB, initially created **automatically** from **dictionaries**. Later, it integrated as well knowledge from encyclopedias, and other kinds of text.

MindNet[3] is maintained by the Microsoft NLP research group and was created by automatic tools. In its creation, a broad-coverage parser generates syntactical trees in which logical rules for the extraction of relations between words are applied. MindNet is, therefore, not a static resource. It represents a methodology consisting of a set of tools to acquire, structure, access and explore lexical-semantic information contained in texts.

Given its initial extraction from dictionaries, the structure of MindNet is based on dictionary entries. For each word entry, MindNet contains a record for each word sense, and provides information such as their POS, and textual definition. Furthermore, each word sense is explicitly related to other words. MindNet contains a broad set of semantic (and syntactic) relations, including Attribute, Cause, Co-Agent, Color, Deep_Object, Deep_Subject, Domain, Equivalent, Domain, Goal, Hypernym, Location, Manner, Material, Means, Possessor, Purpose, Size, Source, Subclass, Synonym, Time, Modifier, Part and User.

One interesting functionality offered by MindNet, useful for determining word similarity, is the identification of **relation paths** between words. For example, there are several paths between the words *car* and *wheel*, including not only sim-

[2]The wordnet projects around the world are listed in the Global WordNet Association site: `http://www.globalwordnet.org/gwa/wordnet_table.html` (August 2012)

[3]Available for online queries through `http://stratus.research.microsoft.com/mnex/` (August 2012)

ple relations like {*car* ← Modifier ← *wheel*} but also paths of length two, like {*car* → Hypernym → *vehicle* → Part → *wheel*}, and longer. Each path is automatically weighted according to its salience. Figure 3.2 shows the ten top-weighted paths from *bird* to *parrot*.

1. bird ← **Hyp** ← parrot
2. bird → **Mod** → parrot
3. bird → **Equiv** → parrot
4. bird ← **Tsub** ← include → **Tobj** → parrot
5. bird → **Attrib** → flightless ← **Attrib** ← parrot
6. bird ← **Tsub** ← deplete → **Tsub** → parrot
7. bird → **PrepRel**(as) → kea → **Hyp** → parrot
8. bird ← **Hyp** ← macaw → **Equiv** → parrot
9. bird → **PrepRel**(as) → species → **PrepRel**(of) → parrot
10. bird → **Attrib** → flightless ← **Attrib** ← kakapo → **Hyp** → parrot

Figure 3.2: Ten top-weighted paths from *bird* to *parrot* in Mindnet.

FrameNet

Berkeley **FrameNet** (Baker et al., 1998) is a network of semantic frames (Fillmore, 1982), manually extracted from a systematic analysis of semantic patterns in corpora text. Each frame describes an object, a state or an event, and corresponds to a concept. The frame may be connected to other frames, by means of syntactic and semantic relations of the lexical item that describes the concept. Besides Inheritance, which is more or less the hypernymy relation, the represented semantic relations include Subframe, Inchoative_of, Causative_of, Precedes, Using and See_also.

A frame can be conceived as the description of a situation with properties, participants and/or conceptual roles. An example of a semantic frame is presented in Figure 3.3 – the frame *transportation* is within the domain *motion*, which provides the elements *mover(s)*, *means of transportation* and *paths* and can be described in one sentence as: *mover(s) move along path by means*.

VerbNet

VerbNet (Schuler, 2006) is a verb lexicon, compatible with WordNet, with explicit syntactic and semantic information. Verbs are organised hierarchically into Levin (1993) classes. Each class of verbs is characterized by syntactic frames, semantic predicates and a list of typical verb arguments. The verb classes use several thematic roles, namely: Actor, Agent, Asset, Attribute, Beneficiary, Cause, Location, Destination, Source, Experiencer, Extent, Instument, Material, Product, Patient, Predicate, Recipient, Stimulus, Theme, Time, and Topic. Verbs in VerbNet are mapped to their corresponding WordNet synsets and FrameNet frames. Figure 3.4 illustrates VerbNet with the entry for the class Hit-18.1.

frame(TRANSPORTATION)
frame_elements(MOVER(S), MEANS, PATH) scene(MOVER(S) move along PATH by MEANS)

frame(DRIVING)
inherit(transportation) frame_elements(DRIVING (=MOVER), VEHICLE (=MEANS), RIDER(S) (=MOVER(S)), CARGO (=MOVER(S))) scenes(DRIVER starts VEHICLE, DRIVE controls VEHICLE, DRIVER stops VEHICLE)

frame(RIDING_1)
inherit(TRANSPORTATION) frame_elements(RIDER(S) (=MOVER(S)), VEHICLE (=MEANS)) scenes(RIDER enters VEHICLE, VEHICLE carries RIDER along PATH, RIDER leaves VEHICLE)

Figure 3.3: The frame *transportation* and two of its subframes, in FrameNet (from Baker et al. (1998)).

Class Hit-18.1	
Roles and Restrictions	Agent[+int_control] Patient[+concrete] Instrument[+concrete]
Frames:	
Name	Basic Transitive
Example	*Paula hit the ball*
Syntax	Agent V Patient
Semantics	cause(Agent, E) manner(during(E), directedmotion, Agent) !contact(during(E), Agent, Patient) manner(end(E),forceful, Agent) contact(end(E), Agent, Patient)

Figure 3.4: Simplified VerbNet entry for the *Hit-18.1* class, from `http://verbs.colorado.edu/~mpalmer/projects/verbnet.html` (September 2012).

Common-sense Knowledge Bases

Even though they do not fit in our definition of LKB, common-sense knowledge bases are sometimes used as such. These resources include CyC (Lenat, 1995; Matuszek et al., 2006) and ConceptNet (Liu and Singh, 2004; Havasi et al., 2009). CyC and its free version OpenCyC[4] are highly formalised knowledge bases, structured in first-order logic predicates. As for ConceptNet[5], it was originally created collaboratively and used to contain common-sense facts obtained from natural language statements whose gaps were filled by volunteers visiting the ConceptNet website. Those statements are illustrated by the following sentences:

- *The effect of eating food is* _____ .
- *A knife is used for* _____ .

Nowadays, ConceptNet also integrates knowledge from other sources, including the English Wikipedia (through DBPedia (Bizer et al., 2009) and ReVerb (Fader et al., 2011; Etzioni et al., 2011)) and Princeton WordNet.

Both CyC and ConceptNet are structured on concepts and relations but they are not limited to lexical knowledge. They contain as well more practical knowledge,

[4]Available from `http://www.opencyc.org/` (August 2012)
[5]Available from `http://conceptnet5.media.mit.edu/` (August 2012)

knowledge about the world and common-sense knowledge, which can be defined as knowledge that an ordinary person is expected to know. An important feature of common-sense knowledge bases is that they are typically very formalised, which eases their representation as ontologies and provides reasoning capabilities.

Popular LKBs in numbers

In order to have an idea on the size and contents of the popular LKBs, tables 3.1 and 3.2 contain quantitative information about them. Table 3.1 shows the number of lexical items included in each LKB, according to their POS. Table 3.2 indicates the core structure of each LKB, the number of instances of that structure, the number of different types of relation that may connect two core structures, and the unique types of the later relations. When the information in the table cells is missing, it is not applicable. For WordNet, the number of relations includes the direct and the inverse relations, because they have different names. On the other hand, MindNet identifies direct and inverse relations by a directed arrow (e.g. ←Hyp and Hyp→, repectively for hypernymy and hyponymy). So, in front of the number of MindNet relations, we added the information '×2'.

Due to its different structure, it is not possible to compare MindNet, created automatically, with the other LKBs, all handcrafted. The only number that shows that MindNet is larger is the number of relation instances (713k), which is significantly higher than in WordNet 3.0 (285k). Moreover, as an automatic approach, MindNet can also grow by processing more text. Richardson et al. (1998) refer that, after processing the Microsoft Encarta 98 encyclopedia, 220k additional headwords were collected for MindNet. Also, the MindNet website currently refers a total of 45 different relation types, which is more than the 32 reported in 1998.

These numbers also show that LKBs are much smaller than knowledge bases as DBPedia (more than 2.5M concepts and 250M relations) and Freebase (Bollacker et al., 2008), a collaborative knowledge base (more than 20M concepts and 300M relations). This occurs especially because LKBs are restricted to lexical knowledge, while the others are much broader and contain a wide-range of world knowledge facts.

Resource	**Lexical items**					
	Nouns	Verbs	Adjectives	Adverbs	Other	Total
WordNet 3.0 (2006)	117,097	11,488	22,141	4,601	-	155,327
MindNet (1998)	-	-	-	-	-	159,000 headwords
FrameNet (2012)	5,136	4,819	2,268	-	378	12,601
VerbNet (2012)	-	3,769	-	-	-	3,769

Table 3.1: Comparison of LKBs according to included lexical items.

Resource	**Core structure**		**Relations**	
	Type	**Instances**	**Unique Types**	**Instances**
WordNet 3.0 (2006)	synset	117k+	20	285k
MindNet (1998)	word entry	191k definitions	32 (×2)	713k
FrameNet (2012)	frame	1,674	8	-
VerbNet (2012)	class entry	274	-	-

Table 3.2: Comparison of LKBs according to core structure and relations.

3.1.2 Portuguese Lexical Knowledge Bases

There are three proprietary handcrafted wordnets for Portuguese: WordNet.PT, MWN.PT and WordNet.Br. In the context of public LKBs, other resources that should be mentioned include: TeP and OpenThesaurus.PT, two handcrafted thesauri structured in synsets; Wiktionary.PT and Dicionário Aberto, two electronic dictionaries; and PAPEL, a lexical-semantic network, automatically extracted from a dictionary. Besides those, there is a resource based on frame semantics, FrameNet.Br (Salomao, 2009; de Souza, 2010), for Brazilian Portuguese.

Moreover, during the writing of this thesis, a new public domain Portuguese wordnet was released – OpenWN-PT (de Paiva and Rademaker, 2012). This resource is the Portuguese part of an open multilingual wordnet initiative[6], where wordnets in several languages are linked to Princeton WordNet 3.0. The Portuguese synsets were obtained after translation of the most important WordNet synsets, considering their position in the hierarchy and the number of relations where they are involved. The semantic relations are inherited from the same resource.

WordNet.PT

WordNet.PT (Marrafa, 2001, 2002), recently extended to WordNet.PT Global – *Rede Léxico-Conceptual das variedades do Português* (Marrafa et al., 2011), is a resource developed by Centro de Linguística da Universidade de Lisboa in collaboration with Instituto Camões. This project aimed to develop a broad-coverage wordnet for the European Portuguese variant. The recent extension deals with the inclusion of variants from other Portuguese speaking countries.

The development of WordNet.PT started in 1998. Its version 1.5 contains a rich set of semantic relations, covering: general-specific; whole-part; equivalence; opposition; categorisation; participation in an event; and defining the event structure. The creation of WordNet.PT is manual, and its structure is based on the EuroWordNet (Vossen, 1997) model, and thus inspired by WordNet. It covers the following semantic sub-domains: professional and artistic activities, food, geographical and political regions, institutions, instruments, means of transportation, works of art, health and medical acts, living beings, and clothing.

According to the information provided by its website[7], WordNet.PT Global contains a network with 10,000 concepts, including nouns, verbs and adjectives, their lexicalisations in the different Portuguese variants, and their glosses. The concepts, which are a subset of the WordNet.PT concepts, are integrated in a network with more than 40,000 relation instances of several types.

MWN.PT

MWN.PT – MultiWordNet of Portuguese[8] is the Portuguese branch of the MultiWordNet project (Pianta et al., 2002). It is developed by the NLX-Natural Language and Speech Group at the University of Lisbon, and can be purchased from

[6] See more at `http://casta-net.jp/~kuribayashi/multi/`

[7] See `http://cvc.instituto-camoes.pt/traduzir/wordnet.html` (August 2012)

[8] Available online from `http://mwnpt.di.fc.ul.pt` (August 2012)

the ELRA catalog[9].

In MWN.PT's documentation[10], its authors refer that the first version of this resource spans over 17,200 manually validated synsets, which correspond to 21,000 word senses/word forms and 16,000 lemmas, from both European and Brazilian variants of Portuguese. The MWN.PT synsets are aligned with the translation equivalent concepts of Princeton WordNet and, transitively, to the MultiWordNets of Italian, Spanish, Hebrew, Romanian and Latin.

MWN.PT synsets are linked under the semantic relations of hypernymy/hyponymy and meronymy (part, member and substance) (Santos et al., 2010). Furthermore, MWN.PT includes the subontologies under the concepts of Person, Organization, Event, Location, and Art works. The authors of MWN.PT claim that their resource covers the top ontology with the Portuguese equivalents to all concepts in the top four layers of Princeton WordNet, to the 98 Base Concepts suggested by the Global Wordnet Association, and to the 164 Core Base Concepts, indicated by the EuroWordNet project[11]. However, MWN.PT only covers nouns, while the 164 Core Base Concepts contain not only 66 concrete and 63 abstract noun synsets, but also 35 abstract verb synsets.

WordNet.Br

WordNet.Br (Dias da Silva et al., 2002) is a wordnet resource for the Brazilian variant of Portuguese. This project consisted of two main development phases. First, a team of three linguists analysed five Brazilian Portuguese dictionaries and two corpora in order to acquire synonymy and antonymy information. This resulted in the manual creation of synsets, and antonymy relations between them, as well as the writing of synset glosses and the selection of sentences where the synset occurred.

In a second phase (Dias-da Silva et al., 2006), the WordNet.Br synsets were manually aligned with the Princeton WordNet synsets, with the help of bilingual dictionaries. A strategy similar to that suggested by the EuroWordNet project was followed in this process. After the alignment, the WordNet.Br synsets that were aligned to Princeton WordNet could inherit the relations of this resource. This means that WordNet.Br covers the relations of: hypernymy, meronymy, cause and entailment.

Portuguese thesauri

TeP (Dias-Da-Silva and de Moraes, 2003; Maziero et al., 2008) was originally the synset-base of WordNet.Br (Dias da Silva et al., 2002), created during its first development phase. It is maintained by Núcleo Interinstitucional de Lingüística Computacional (NILC) of the University of São Paulo, in São Carlos, Brazil. Its current version, TeP 2.0[12], is publicly available and contains more than 44,000 lexical items, organised in 19,888 synsets. TeP also contains 4,276 antonymy relations between synsets.

[9]The European Language Resources Association (ELRA) catalog is available from `http://catalog.elra.info/` (August 2012)

[10]Available online from `http://mwnpt.di.fc.ul.pt/features.html` (August 2012)

[11]See more about these lists of concepts in `http://www.globalwordnet.org/gwa/ewn_to_bc/topont.htm` (August 2012)

[12]Available from `http://www.nilc.icmc.usp.br/tep2/` (August 2012)

OpenThesaurus.PT[13] (hereafter, OT.PT) is the Portuguese version of a collaborative thesaurus initiative (Naber, 2004). It is approximately four times smaller than TeP. OT.PT contains 13,258 lexical items, organised in 4,102 synsets, but the project has not had any significant development since 2006. This resource is mainly used in the OpenOffice[14] word processor for suggesting synonyms.

Electronic dictionaries

There are several Portuguese dictionaries available for online queries, however, we would like to mention two of them which, besides containing some additional explicit semantic markups, are public domain and thus freely available for download and use.

Wiktionary.PT[15] is a collaborative dictionary by the Wikimedia foundation where, besides the typical dictionary information, it is possible to add information on semantic relations for each entry. For Portuguese however, this resource is still small and, besides other problems, most entries do not have information about semantic relations. On May 2012, Wiktionary.PT contained almost 180,000 entries. However, as all Wiktionaries are multilingual, not all of those entries correspond to Portuguese words.

Dicionário Aberto (hereafter DA, Simões and Farinha (2011); Simões et al. (2012)) is the electronic version of an old Portuguese dictionary from 1913, maintained by Alberto Simões in University of Minho. DA, whose orthography is currently being modernised, has 128,521 entries. Recently, some semantic relations, extracted using simple patterns, were added to the DA's interface[16], in a so called ontology view (Simões et al., 2012).

Lexical-semantic network

PAPEL (Gonçalo Oliveira et al., 2008, 2009, 2010b) is a public domain lexical resource with instances of several types of semantic relations, extracted automatically from Dicionário PRO da Língua Portuguesa (DLP, 2005), a Portuguese dictionary, property of Porto Editora. It was developed by Linguateca. The main differences between PAPEL and a wordnet is that PAPEL was created automatically and is not structured in synsets, nor sense-aware. PAPEL can be seen as a lexical network – it is structured in relational triples $t = \{w_1, R, w_2\}$ denoting instances of semantic relations R, where w_1 and w_2 are lexical items, identified by their orthographical form. Its current version, PAPEL 3.0[17], contains about 190,000 triples of different types, connecting about 100,000 unique lexical items.

Portuguese LKBs in numbers

Similarly to what we have done for the English LKBs, here, we put the Portuguese LKBs side-by-side. Table 3.3 characterises the LKBs according to their construction and availability. Table 3.4 shows the number of included lexical items, according to their POS. Especially due to its automatic construction, PAPEL is clearly the

[13] Available from `http://openthesaurus.caixamagica.pt/` (August 2012)
[14] See `http://www.openoffice.org/` (August 2012)
[15] Available from `http://pt.wiktionary.org/` (August 2012)
[16] Available from `http://www.dicionario-aberto.net/` (August 2012)
[17] Available from `http://www.linguateca.pt/PAPEL/` (August 2012)

resource that covers more lexical items. The same table shows that, excluding MWN.PT that only covers nouns, all LKBs cover also verbs and adjectives[18].

Table 3.5 presents the LKBs according to their core structure and relation types. For resources structured in synsets, synonymy is not considered a relation type, but a structural property. Moreover, the number of unique relations does not include the inverse types, even though some resources have them defined. For those, we provide additional information for considering the inverse relations and instances: '$\times 2$' if all relations have an inverse; '$+n$' if n relations have an inverse.

Table 3.6 indicates the relation types covered by each LKB. When the LKB contains more than one subtype of some relation, the number of subtypes is given as well, in parenthesis. For instance, WordNet.PT defines two types of antonymy and hypernymy, six types of meronymy, three of cause, and four of purpose. In this table, we did not consider the inverse relation types (not explicit in PAPEL). WordNet.PT and PAPEL are the resources with, at least, one subtype of each semantic relation in the table. For WordNet.Br, we are not sure about the covered relation types but, given that, besides antonymy, its relations come from Princeton WordNet, we believe that it covers exactly the same types as this resource. On the other hand, as thesauri, TeP and OT.PT are structured on synsets, and thus only handle synonymy. They do not cover other semantic relations, except for TeP, which contains also antonymy relations between synsets.

Resource	Construction	Availability
WordNet.PT	manual	proprietary
WordNet.Br	manual (synsets) + translation equivalence (relations)	proprietary
MWN.PT	manual (translation)	paid license (academic, standard, custom)
TeP	manual	free for research purposes
OT.PT	manual (collaborative)	free
PAPEL	automatic	free

Table 3.3: Portuguese LKB according to construction and availability.

Resource	Lexical items				
	Nouns	Verbs	Adjectives	Adverbs	Total
WordNet.PT 1.0	9,813	633	485	0	10,931
MWN.PT v1	16,000	0	0	0	16,000
WordNet.Br	17,000	10,910	15,000	1,000	43,910
TeP 2.0	17,276	7,660	15,001	1,138	44,325
OT.PT	6,110	2,856	3,747	143	12,856
PAPEL 3.0	74,592	32,454	28,248	1,787	137,081

Table 3.4: Portuguese LKBs according to included lexical items.

More information on some of the aforementioned LKBs and their comparison may be found in Santos et al. (2010). Furthermore, Teixeira et al. (2010) provide a comparison between the verbs in TeP, OT.PT, PAPEL and Wiktionary.PT. Both

[18]OT.PT does not really contain explicit information on the POS of the lexical items and synsets, but we inferred the POS of most of the synsets, with the help of a morphological analyser. Given a word, the latter lists all its possible POSs. So, we attributed to each synset the POS of the majority of the words it contained.

Resource	Core structure		Relations	
	Type	Instances	Unique Types	Instances
WordNet.PT 1.5	synset	12,630	35 (+26)	40,000+
MWN.PT v1	synset	17,200	5 (×2)	68,735
WordNet.Br	synset	18,200	5 (× 2)	N/A
TeP 2.0	synset	19,888	1	2,138 (×2)
OT.PT	synset	4,002	0	0
PAPEL 3.0	lexical item	137,081	44 (×2)	201,288 (×2)

Table 3.5: Portuguese LKBs according to core structure and relations.

Resource	Relations							
	Synonymy	Antonymy	Hypernymy	Meronymy	Cause	Purpose	Place	Manner
WordNet.PT	yes	yes (2)	yes (2)	yes (6)	yes (3)	yes (4)	yes	yes
MWN.PT v1	yes	no	yes	yes (3)	no	no	no	no
WordNet.Br	yes	yes	yes	yes	yes	no	no	no
TeP 2.0	yes	yes	no	no	no	no	no	no
OT.PT	yes	no	no	no	no	no	no	no
PAPEL 3.0	yes (4)	yes	yes	yes (9)	yes (5)	yes (4)	yes	yes (2)

Table 3.6: Semantic relations in Portuguese LKBs.

of these works concluded that, although all the LKBs are broad-coverage language resources for Portuguese, their contents are **more complementary than overlapping** and it would be fruitful to merge some of them in a unique broader resource. This problem is common to most languages. It is thus no surprise that there have been attempts to merge or align different knowledge bases, or to enrich knowledge bases with information extracted from other sources (see examples in section 3.3).

Limitations of Portuguese LKBs

The previous numbers show that all Portuguese LKBs presented have interesting sizes and might be useful for several tasks. At the same time, some of their limitations are highlighted. Below, four limitations are enumerated.

Limited coverage: As it happens for English, a manual creation approach limits the coverage of the resource (and the time needed for its creation). For instance, MWN.PT only covers nouns, and TeP and OT.PT only handle synonymy. Moreover, the authors of WordNet.PT refer that this resource only covers a set of semantic sub-domains (Marrafa, 2002) while, as far as we know, the other LKBs are not limited to any domain.

It is thus no surprise that PAPEL, the only resource created automatically, is much larger than the others, not only in terms of covered lexical items, but also in terms of covered types of semantic relations. On the other hand, a manual construction approach tends to create more reliable resources, which means that PAPEL has probably more inconsistencies than the other LKBs.

Translation approach: One limitation, not present in the comparison, regards the construction of MWN.PT, based on the translation of Princeton WordNet. As different languages lexicalise the same concepts differently, this approach results in several lexical gaps. For instance, in MWN.PT, some of the relations have empty arguments, or arguments filled with `GAP!` or `PSEUDOGAP!`, which might refer to

English concepts that do not have a Portuguese equivalent (Santos et al., 2010). This happens, for instance, to the Princeton WordNet concepts of *human_action* or *magnitude_relation*, aligned to a `GAP!` in MWN.PT. Moreover, the translation approach tends not to cover specific lexicalisations of the target language.

This points out a serious problem when translating a target wordnet to a different language. The particular semantics of a word in a language might be significantly different from its translation equivalent in another language (Cruse, 1986). Moreover, as different languages represent different socio-cultural realities, they do not cover exactly the same part of the lexicon and, even where they seem to be common, several concepts are lexicalised differently (Hirst, 2004).

An alternative approach is followed for WordNet.Br, where the concepts are created from scratch for Portuguese and only the relations of the translation equivalents are inherited from Princeton WordNet. Although this approach should not result in lexical gaps, in our view, it does not guarantee that, after the selection of the translation equivalents, inconsistent relations are not generated.

Not sense aware: PAPEL is the only referred LKB not structured in synsets, and not sense-aware. Since language is ambiguous, in several NLP tasks, not discriminating different senses of the same word is a limitation. Also, even though it is also the resource with more relation instances, if PAPEL were structured in synsets, this number would surely be lower, as some words would be grouped together.

Usage restrictions: Not all of these LKBs are freely available for utilisation and integration in other systems or applications. Despite the availability of part of WordNet.PT for online queries[19], at the moment of writing this thesis, it was not publicly available for download. MWN.PT is also available for queries through two interfaces[20]. It is not free, but a commercial or an academic license can be bought. Only the synset-base of WordNet.Br is freely available, through TeP. The relations are not, but it is possible to query online for its information on verbs[21].

3.2 Lexical-Semantic Information Extraction

Lexical-Semantic Information Extraction (LSIE) is a special kind of IE where, instead of concepts or named entities, relations are held between word senses, typically identified by lexical items. This means that LSIE deals mainly with the acquisition of lexical-semantic relations.

Since the 1970's, before the creation of Princeton WordNet, researchers have been exploiting textual resources and developing techniques towards the automatic extraction of lexical-semantic knowledge, which could be used in the automatic creation of a broad-coverage LKB. It is thus no surprise that electronic dictionaries were the primary resources exploited for LSIE (see Calzolari et al. (1973) and Amsler (1980)). Language dictionaries are repositories that compile words and expressions

[19] WordNet.PT can be queried online, through `http://www.clul.ul.pt/wn/` (August 2012)

[20] The Visuwords interface for MWN.PT is available from `http://mwnpt.di.fc.ul.pt/` (August 2012). The MultiWordNet interface is available from `http://multiwordnet.fbk.eu/online/multiwordnet.php` (August 2012)

[21] See `http://caravelas.icmc.usp.br/wordnetbr/index.html` (September 2012)

of a language. They are substantial sources of general lexical knowledge (Briscoe, 1991) and "authorities" of word senses (Kilgarriff, 1997), which are described in textual definitions, written by lexicographers, the experts on the field.

Despite several automatic attempts to the creation of a broad-coverage LKB, for English, Princeton WordNet, a manual effort, ended up to be the leading resource of this kind (Sampson, 2000). As discussed in section 3.1.1, the existence of a wordnet in one language has a positive impact in the development of NLP tools for that language. Nevertheless, despite the wide acceptance of WordNet, research on LSIE continues, not only from dictionaries, but especially from corpora and other unstructured resources, whether it is for the enrichment of WordNet (see section 3.3) or for the creation of alternative LKBs, including LKBs in non-English languages.

In this section, we start with a brief chronology of LSIE from dictionaries. Then, we present work on LSIE from corpora and IE from other unstructured textual resources.

3.2.1 Information Extraction from Electronic Dictionaries

In the beginning

During the 1970s, and throughout the 1980s, electronic dictionaries started to be the target of empirical studies (e.g. Calzolari et al. (1973); Amsler (1980); Michiels et al. (1980)), having in mind their exploitation in the automatic construction of a LKB. This kind of knowledge base would ease the access to morphological and semantic information about the defined words (Calzolari et al., 1973), which would then be very useful in the achievement of NLP tasks.

These earlier works confirmed that the vocabulary used in dictionaries is limited, which makes them easier to process for obtaining semantic or syntactic relations (Michiels et al., 1980). They concluded that the textual definitions are often structured on a ***genus*** and a ***differentia*** (Amsler, 1980):

- The *genus* identifies the superordinate concept of the definiendum – the definiendum is an instance or a "type of" the *genus*, which means there is a hyponymy relation between the former and the latter.
- The *differentia* contains the specific properties for distinguishing the definiendum from other instances of the superordinate concept.

Having in mind that this kind of structure is suitable for being exploited in the automatic acquisition of **taxonomies**, Amsler (1981) proposes a taxonomy for English nouns and verbs. The extracted structures, dubbed **tangled hierarchies**, were created after the analysis of dictionary definitions and manual disambiguation of the head word of each definition. Amsler (1981) concluded that dictionaries clearly represent two taxonomic relations: is-a (hypernymy) and is-part (part-of).

Calzolari (1984) suggests a set of frequent patterns in dictionary definitions, and examines the occurrence of the hyponymy and "restriction" relations. She claims that hyponymy is the most important and evident relation in the lexicon and confirms it can be easily extracted from a dictionary, after identifying the *genus* and the *differentia*.

Markowitz et al. (1986) identified a set of textual patterns that occur in the beginning of the definitions of a dictionary. Those patterns are used to denote relations

of superordination (*any*, *any of*), member-set (*member of*), human noun (*one*), and active or stative verb or adjective.

Chodorow et al. (1985) proposed "head-finding" heuristics to identify the *genus* of noun and verb definitions. Bearing in mind the structure of the definitions and assuming that a defined concept is often a hyponym of its superordinate, they took advantage of the restricted vocabulary used in the definitions to develop **semi-automatic recursive procedures** aiming at the extraction and organisation of semantic information into taxonomies. The definitions did not have to be completely parsed due to their predictability. However, the human user played an important role when it came to WSD. The authors claim a virtual 100% accuracy in the *genus* extraction for verbs, using a very simple heuristic: the head is the single verb following the word *to*. If there is an enumeration of verbs following *to*, then they are all heads. For example:

- `winter, v: to pass the winter` → head = *pass*
- `winter, v: to keep, feed or manage during the winter` → heads = {*keep*, *feed*, *manage*}

When it comes to nouns, the task is more complex due to their greater variety, but Chodorow et al. (1985) could still come up with a heuristic for the extraction of the *genus*. First, they isolate the substring containing the head, which is bounded on the left by a word like *a*, *an*, *the*, *its*, *two*, *three*, ... , *twelve*, *first*, *second*, ... and is bounded on the right by a word with the following characteristics:

- a relative pronoun (introducing a relative clause);
- a preposition not followed by a conjunction (introducing a complement to the head noun);
- a preposition-conjunction-preposition configuration (also introducing a complement);
- a present participle following a noun (introducing a reduced relative clause).

The head is typically the rightmost noun in the substring. Chodorow et al. (1985) claim 98% accuracy for the heuristic for nouns, but this heuristic was only capable of identifying the head of the definition, and decide whether that was the hypernym of the definiendum or not.

Alshawi (1987) analysed the definitions of a dictionary to identify syntactic patterns, and used them to define a set of semantic structures based on the meaning of the defined words. The structures were derived from the identification of the subordinated terms or modifiers, prepositions and other words that could indicate relations in the definition. A set of semantic relations (e.g. class, purpose, manner, has-part) and, in some cases, specific properties, were extracted and included in the semantic structures. While the aforementioned works were based on string patterns for parsing parts of the definition, Alshawi (1989) proposed a specific **semantic grammar** for the derivation of the definitions of a specific dictionary. Since they were based on the structure of a specific dictionary, the application of the grammars to unrestricted text or to other dictionaries would not be a good option.

Broad-coverage parsing of dictionaries

After some discussion about the advantages and the drawbacks of using string patterns or **structural patterns** to extract semantic information from the definitions, Montemagni and Vanderwende (1992) concluded that, although string patterns are very accurate for identifying the *genus*, they cannot capture the variations in the *differentia* as well as structural patterns. String patterns have several limitations in the extraction of relations depending on the *differentia*, including when:

- There is an enumeration of concepts at the same level, as in: `to make laws, rules or decisions`;
- There are parentheses in the middle of the definition;
- It is necessary to identify functional arguments;
- There are specific relations inside the definition as in: `pianta erbacea com bacche di color arancio`. There, the color should not be extracted as a feature of the definiendum.

They propose the use of a **broad-coverage grammar** to parse the dictionary definitions and obtain rich semantic information. Structural patterns are based on the syntactic structure of the definition, obtained after the syntactic analysis, made by a broad-coverage parser. Despite seeming excessive to use a broad-coverage parser for definitions, the authors refer that there are cases when its use is warranted.

Work using a broad-coverage parser to process dictionaries (Dolan et al., 1993; Richardson et al., 1993; Vanderwende, 1994, 1995; Richardson, 1997), led to the creation of **MindNet** (Richardson et al., 1998), which can be seen as sort of an independent LKB, in a way that previous work on LSIE from dictionaries was not. MindNet can also be seen as a methodology for acquiring, structuring, accessing, and exploiting semantic information from text. Richardson (1997) discusses the creation of MindNet in more detail. For more information on MindNet as a resource, see section 3.1.1. The creation of MindNet shown as well that much information about a word can be found in the definition of other words (Dolan et al., 1993).

A different approach for the automatic creation of a LKB from a dictionary is presented by Barriere (1997), who describes a method for transforming a children's dictionary in a LKB based on **conceptual graphs** (Sowa, 1992). Points in favor of using a children's dictionary go from the inclusion of simple knowledge where almost all words are themselves defined, and the use of complete sentences, to the naive view of things and limited number of senses for each word. Conceptual graphs were used because they are a logic-based formalism and are flexible to express the background knowledge necessary for understanding natural language.

Issues in LSIE from dictionaries

One of the problems first noticed when dictionaries started to be used for building taxonomies was **circularity**, often present in definitions (Calzolari, 1977). This phenomenon occurs when visiting the definition of the head of an entry iteratively and, at some point, ending up in an entry that had already been processed. The following is an example of circularity (*portion* → *part* → *piece* → *portion*):

- `portion, n: a part of a whole`;

- `part, n: a piece of something`;
- `piece, n: a portion of some material`;

Amsler (1981) believes that these loops are usually the evidence of a truly primitive concept, such as the set containing the words *class*, *group*, *type*, *kind*, *set*, *division*, *category*, *species*, *individual*, *grouping*, *part* and *section*. These primitives are often related with "**covert categories**" (Ide and Véronis, 1995), which are concepts that do not correspond to any particular word and are introduced to represent a specific category or group of concepts. For instance, there is no word to describe the hypernym of the concepts described by *tool*, *utensil*, *implement* and *instrument*, so a new "covert" hypernym, *instrumental-object*, is artificially created.

Chodorow et al. (1985) introduced the notion of "**empty heads**". Words belonging to this small class (e.g. *one*, *any*, *kind*, *class*, *manner*, *family*, *race*, *group*, *complex*) might occur in the beginning of the definition followed by the preposition *of*, but do not represent the superordinate concept. Guthrie et al. (1990) explored the class of "empty heads" to extract other semantic relations, besides hyponymy. For instance, the word *member* is related with the member-set relation (Markowitz et al., 1986) and the word *part* is related with the is-part relation (included by Amsler (1981) in his tangled hierarchies). Concerning this problem, Nakamura and Nagao (1988) provide a list of function nouns that appear in the beginning of dictionary definitions, and the relations they are usually associated with:

- `kind`, `type` → is-a
- `part`, `side`, `top` → part-of
- `set`, `member`, `group`, `class`, `family` → membership
- `act`, `way`, `action` → action
- `state`, `condition` → state
- `amount`, `sum`, `measure` → amount
- `degree`, `quality` → degree
- `form`, `shape` → form

Another typical issue is the **disambiguation of the *genus***, which consists on matching words that appear in the definition with their correct sense in the dictionary. In Amsler (1981) and Chodorow et al. (1985), this task requires human intervention. Some years later, Bruce and Guthrie (1992) worked on an automatic procedure to accomplish *genus* disambiguation. First, they identify the *genus* of the definition. Then, they exploit category markups (e.g. plant, solid) and frequency information to disambiguate the *genus* with 80% accuracy.

More recently, Navigli (2009a) presented an algorithm to disambiguate words in dictionary definitions. Their approach is based on the exploitation of circularity in dictionaries.

Electronic dictionaries are certainly an important source of lexical-semantic knowledge, but their organisation does not favour their **direct use as NLP tools**, since they were made to be read by humans. Wilks et al. (1988) mention several

points that should be considered in the automatic extraction of knowledge from a dictionary, and its conversion into a computational format. They refer three approaches for creating a knowledge base from a dictionary, which vary in the initial required amount of knowledge, and in the quality of the extracted information:

- Co-ocurrences enable the establishment of associations between words, without requiring initial linguistic information.
- A grammar with a collection of linguistic patterns enables, for instance, to identify the *genus* (hypernym) and the *differentia* for each dictionary entry.
- Hand-coding the lexical entries of a controlled vocabulary (about 5% of the knowledge base), and iterating through the remaining words, enables to derive a network of semantic units.

While the third approach results in a rich semantic structure, it needs a substantial amount of initial linguistic knowledge. The first approach produces a much simpler resource, but does not require hand-coded knowledge.

Ide and Véronis (1995) are very critical of the research on information extraction from dictionaries. They refer that dictionaries use **inconsistent** conventions to represent knowledge and that the definitions are **not as consistent** as they should be. Since they are the result of several lexicographers work for several years, dictionaries have many variations to transmit the same thing. Reviews and updates increase the probability of inconsistencies.

In order to assess the information extracted from dictionaries, Ide and Véronis (1995) performed a quantitative evaluation of automatically extracted hypernymy relations. As hypernymy is the least arguable semantic relation and the easiest to extract, the authors believed that, if their results were poor, they would be poorer for more complex domains and less clearly defined relations. The evaluation consisted of comparing an "ideal" hierarchy, manually created, with hierarchies extracted from five dictionaries. The extraction procedure was based on the heuristics of Chodorow et al. (1985), which resulted in tangled hierarchies, later disambiguated manually. After inspection, it was noticed that these hierarchies had serious problems of **incompleteness** and there were difficulties at higher levels:

- Some words were (relatively randomly) attached too high in the hierarchy; some heads of definitions were not the hypernym of the definiendum, but the "whole" that contains it; overlaps that should occur between concepts are sometimes missing.
- All the heads separated by the conjunction *or* are considered to be hypernyms, but sometimes, when looking at the hierarchy, problems exist; circularity tends to occur in the highest levels, possibly when lexicographers lack terms to designate certain concepts.

The authors state that hierarchies with this kind of problems are likely to be unusable in NLP systems and discuss means to refine them automatically. **Merging** the hierarchies of the five dictionaries and introducing "covert categories" drastically reduces the amount of problems from 55-70% to 6%. Other problems are minimised by considering "empty heads" and patterns occurring in the beginning of the definition that denote the part-of relation; or by using more complex grammars/broad-coverage parsers instead of static string patterns for extraction.

LSIE from dictionaries after WordNet

Since the establishment of WordNet as the paradigmatic LKB for English, less attention has been given to the exploitation of dictionaries for the automatic creation of LKBs. Nevertheless, there are recent works where electronic dictionaries are exploited for this and other NLP tasks, for English and for other languages. In some of these works, WordNet is used as a dictionary.

For instance, O'Hara (2005) worked on the extraction of semantic relations (e.g. used-for, has-size) from the WordNet glosses. Special attention was given to the information in the *differentia* to find distinctions between co-hyponyms and perform **WSD**. O'Hara (2005) used a broad-coverage dependency parser to determine the syntactic relations present in a sentence. Then, the surface-level syntactic relations determined by the parser were disambiguated into semantic relations between the underlying concepts. Isolating the disambiguation from the extraction allows flexibility over earlier approaches. After disambiguation, the relations are weighted according to their relevance to the assigned concepts, resulting in a labelled direct graph where each link has a probability value. The output is converted into a Bayesian network.

Navigli (2009a) took advantage of cycles in a dictionary graph for **disambiguating** the words in their definitions. The graph connects word senses to words referred in their glosses. For each word, a candidate sense is selected based on the number of (quasi-)cycles that include both this sense and the word sense of the gloss. This procedure was applied to two electronic dictionaries and also to WordNet.

Nichols et al. (2005) introduced a system that creates **ontologies** by extracting knowledge from dictionary glosses. Their approach combines deep and shallow parsing of the definition sentences and generates a semantic representation. They applied their procedure to a Japanese dictionary.

In the project PAPEL (Gonçalo Oliveira et al., 2008, 2010b), a Portuguese dictionary (DLP, 2005) was exploited in the creation of a resource where lexical items are connected by semantic relations. For the creation of PAPEL, the definitions of the dictionary were analysed for the manual creation of grammars that would then be used to extract relations between words in the definition and the definiendum. CARTÃO (Gonçalo Oliveira et al., 2011) resulted on the extraction of relations from other Portuguese dictionaries using the same procedure and grammars as PAPEL. See more about these resources in section 4.

More recently, the public dictionary **Wiktionary**[22] has also been the target of work on LSIE. Wiktionay is a collaborative initiative, maintained by the Wikimedia Foundation, which provides multilingual electronic dictionaries of free content. As Wiktionaries are built manually by non-professional volunteers on the Web, the provided information is usually incomplete and sometimes inconsistent. On the other hand, Wiktionary is free and constantly growing. Wiktionaries have been exploited, for instance, for acquiring synonyms (Navarro et al., 2009; Weale et al., 2009), computing semantic relatedness (Sajous et al., 2010; Zesch et al., 2008b) or for the enrichment of LKBs (Sajous et al., 2010; Henrich et al., 2011). They have also been exploited together with other resources in the automatic creation of LKBs for several languages, including German (Wandmacher et al., 2007) and, in our work, for Portuguese (Anton Pérez et al. (2011), see additional information on section 4).

[22] See `http://www.wiktionary.org/` (August 2012)

3.2.2 Information Extraction from Textual Corpora

It is difficult to set a clear boundary between typical IE from textual corpora and LSIE from the same target. This happens because there are documents with (unstructured) natural language text about almost every topic and domain.

This section presents work on the automatic acquisition of knowledge from unstructured text. It starts with those that associate terms according to their co-occurrence/similarity. It moves on to works on the extraction of lexical-semantic relations, and concludes with works that extract other types of relations from larger sources, such as the World Wide Web.

Associating similar words

Most of the work on word association relies on Harris **distributional hypothesis** (Harris, 1968), which assumes that similar words tend to occur in similar contexts. After defining the context of a word, these works generally follow a procedure to cluster words according to their distributional similarity.

Earlier approaches for this task (Riloff and Shepherd, 1997; Roark and Charniak, 1998) were **weakly supervised**. They used bootstrapping algorithms that started with a set of seed words belonging to the same category (e.g. *airplane*, *car*, *jeep*, *plane*, *truck*), in order to discover more members of this category. In Riloff and Shepherd (1997)'s work, the score of a word W in the category C is computed as follows:

$$Score(W, C) = \frac{frequency\ of\ W\ in\ C's\ context}{frequency\ of\ W\ in\ the\ corpus} \tag{3.1}$$

At each iteration, the five top-scoring nouns that are not yet used as seeds are added to the seed list. In the end, the system outputs a ranked list of nouns, supposedly members of the chosen category.

Roark and Charniak (1998) improved the precision of the previous work by focusing on **linguistic constructions**, where words of the same category often co-occur, such as:

- conjunctions, as in: `lions and tigers and bears...`
- lists, as in: `lions, tigers, bears...`
- appositives, as in: `the stallion, a white Arabian...`
- nominal compounds, as in: `Arabian stallion`.

They proposed a ranking measure that allows for the inclusion of rare occurrences and only considers words in the previous co-occurrence situations. They also select the most frequent head nouns in the corpus as initial seed words, and deal with compound nouns in a separate step.

Before the aforementioned works, Grefenstette (1994) presented a deep study on the automatic creation of thesauri, comparable to Roget's and WordNet synsets, from text. When it came to computing the similarity of words, his approach went further, because it started with a light parsing of text. This enables to compute word similarities considering the identified **syntactic relations** (e.g. object, subject,

modifier), which, according to Grefenstette (1994), give a more precise context than simple co-occurrence.

Lin (1998) proposes a fully unsupervised **clustering** approach for a similar task. In his work, each word is represented by a vector with the contexts where it occurs, while similarities are also computed after parsing and identification of syntactic dependencies. The purpose of clustering is to group together words with similar neighbourhoods. Following the previous work, Lin and Pantel (2002) present Clustering by Committee (CBC), an algorithm for automatically extracting **semantic classes**, as the following:

> *pink, red, turquoise, blue, purple, green, yellow, beige, orange, taupe, white, lavender, fuchsia, brown, gray, black, mauve, royal blue, violet, chartreuse, teal, gold, burgundy, lilac, crimson, garnet, coral, grey, silver, olive green, cobalt blue, scarlet, tan, amber, ...*

Initially, each element's top similar words are found. Then, a set of tight clusters, with representative elements of a potential class (committees), is created. The idea is to form as many dissimilar committees as possible. In the end, each word is assigned to its most similar clusters, which may be used to describe a concept. The committee members for the previous cluster, consisting of elements that unambiguously describe members of the class, would be: *blue, pink, red, yellow.*

Also using CBC, Pantel and Lin (2002) identify different **senses** of the same word. As CBC does not give an actual name to the concepts formed by the committees, Pantel and Ravichandran (2004) worked on an automatic method for labelling the classes formed by the clusters. The label would be a **hypernym** of all the words of the class (e.g. *color* for the previous class).

For Portuguese, Sarmento et al. (2008) refer that they have used a similar approach to Lin (1998)'s for building a verb thesaurus, later used in a question-answering system.

Given that, from a linguistic point of view, word senses are not discrete (Kilgarriff, 1996), their representation as crisp objects does not reflect the human language. Therefore, it is more realistic to adopt models of uncertainty, such as fuzzy logic, to handle word senses and natural language concepts. 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.

Landauer and Dumais (1997) use Latent Semantic Analysis (LSA, Deerwester et al. (1990)) to simulate the way people learn new word meanings from text. LSA is a technique for analysing relationships between sets of documents, according to the terms they contain. LSA uses a term-document matrix for describing the occurrences of terms, represented as points. According to the principle of proximity, terms related in meaning should be represented by points near to one another. A common way of weighting the elements in the matrix is term frequency inverse document frequency (TF-IDF), which gives a value proportional to the number of times a word appears in each document.

But when it comes to **discovering synonyms**, PMI-IR (Turney, 2001), a simpler alternative than LSA, seems to perform better. PMI-IR uses pointwise mutual information (PMI, Church and Hanks (1989)) to score the similarity between two words, which can be seen as a conditional probability of $word_1$ occurring, given that $word_2$ occurs. Both probabilities are calculated by querying a web search engine:

$$PMI(word_1,\ word_2) = \frac{P(word_1\ \&\ word_2)}{P(word_1)\ P(word_2)} \tag{3.2}$$

Extraction of semantic relations

The extraction of semantic relations from large corpora became the paradigm in IE after Hearst (1992)'s seminal work, where an automatic method to **discover lexical-syntatic patterns**, used later for the acquisition of **hyponyms**, is proposed. Besides indicating a hyponymy relation, the patterns must occur frequently and should be recognised with few pre-encoded knowledge. The method can be adapted to any lexical relation, and is summarised by the following steps:

1. Decide the relation to search for (e.g. hyponymy);
2. Gather a list of word pairs for which the relation is known to be held (e.g. *dog - animal*). The pair may be collected from an existing knowledge base;
3. Search for sentences in the corpus where the words of the same pair co-occur, and save the text connecting them;
4. Find similarities among the saved fragments of text and hypothesise patterns indicating the relation;
5. Once a new pattern is positively identified, use it to gather more instances of the target relations, and return to step 2.

The patterns used by Hearst (1992) to extract hyponymy relations are listed below, where <hypo> stands for hyponym and <hyper> for hypernym. An extraction example is given for each pattern. The first three patterns were collected by observation while the other three were discovered automatically.

1. <hyper> such as <hypo> {, <hypo> ... , (and | or) <hypo>}
   The bow lute, such as the Bambara ndang, ...
   ⇒ {*Bambara ndang* hyponym_of *bow lute*}
2. such <hyper> as {<hypo> ,}* {and | or} <hypo>
   ... works by such authors as Herrick, Goldsmith and Shakespeare.
   ⇒ {*Herrick* hyponym_of *author*), {*Goldsmith* hyponym_of *author*), {*Shakespeare* hyponym_of *author*}
3. <hypo> {, <hypo>}* {,} or other <hyper>
   Bruises, ..., broken bones or other injuries ...
   ⇒ {*bruise* hyponym_of *injury*}, {*broken bone* hyponym_of *injury*}
4. <hypo> {, <hypo>}* {,} and other <hyper>
   ... temples, treasuries, and other important civic buildings.
   ⇒ {*temple* hyponym_of *civic building*}, {*treasury* hyponym_of *civic building*}
5. <hyper> {,} including {<hypo> ,}* {and | or} <hypo>
   All common-law countries, including Canada and England ...
   ⇒ {*Canada* hyponym_of *common-law country*}, {*England*, hyponym_of *common-law country*}

6. <hyper> {,} especially {<hypo> ,}* {and | or} <hypo>
   ... most European countries, especially France, England, and Spain.
   ⇒ {*France* hyponym_of *European country*}, {*England* hyponym_of *European country*}, {*Spain* hyponym_of *European country*}

Inspired by the work of Hearst (1992), Freitas (2007) discusses the extraction of hypernymy relations from Portuguese corpora. In her work, some *Hearst patterns* were adapted to Portuguese, which resulted in the following:

- <hyper> {tais} como <hypo> {, <hypo> ... , (e | ou) <hypo>}
  A tentativa posterior de clonar outros mamíferos tais como camundongos, porcos, bezerros,....
  ⇒ {*camundongos* hyponym_of *mamíferos*}, {*porcos* hyponym_of *mamíferos*}, {*bezerros* hyponym_of *mamíferos*}
- <hypo> {, <hypo>}* {,} (e | ou) outros <hyper>
  ... a experiência subjetiva com o LSD-25 e outros alucinógenos.
  ⇒ {*LSD-25* hyponym_of *alucinógeno*}
- tipos de <hyper>: <hypo> { , <hypo> ... ,} (e | ou) <hypo>
  Existem dois tipos de cromossomos gigantes: cromossomos politênicos e cromossomos plumulados.
  ⇒ {*cromossomos politênicos* hyponym_of *cromossomos*}, {*cromossomos plumulados* hyponym_of *cromossomos*}
- <hyper> chamad(o|os|a|as) {de} <hypo>
  ... a alta frequência da doença mental chamada esquizofrenia.
  ⇒ {*esquizofrenia* hyponym_of *doença mental*}

Also for the extraction of hypernyms, Caraballo (1999) proposed a combination of pattern detection and a **clustering** method where noun candidates are obtained from a corpus using data on conjunctions and appositives. A co-occurrence matrix for all nouns is used. It contains a vector for each noun in the corpus, with the number of times it co-occurs, in a conjunction or appositive, with each other noun. If $\vec{v}$ and $\vec{w}$ are the vectors of two nouns, similarity between them is calculated as below, which can be see as a variant of LSA (cosine similarity):

$$\cos(\vec{v}, \vec{w}) = \frac{\vec{v}.\vec{w}}{|\vec{v}|.|\vec{w}|} \tag{3.3}$$

In a post-processing step, *Hearst-like* patterns are used for finding hypernym candidates, which, if appropriate, are placed as common parent nodes for clusters.

Cederberg and Widdows (2003) used a similar variant of LSA to improve the precision and recall of hyponymy relations, extracted from a corpus using *Hearst-like* patterns. Having in mind that a hyponym and its hypernym are expected to be similar, LSA is used to compute the similarity of terms in the extracted relations. While the precision of a random sample of extracted relations was 40%, the precision of the 100 relations with higher similarity was 58%, which suggests the effectiveness of this method for reducing errors.

Furthermore, as most of the potential hyponymy relations that could be extracted are not expressed by the six *Hearst patterns*, Cederberg and Widdows (2003) improved the recall of their method using coordination as a cue for similarity. They

give the following sentences, taken from the British National Corpus[23] (BNC), to illustrate their assumptions:

1. *This is not the case with sugar, honey, grape must, cloves and other spices which increase its merit.*
   ⇒ {*clove* hyponym_of *spice*}
2. *Ships laden with nutmeg or cinnamon, cloves or coriander once battled the Seven Seas to bring home their precious cargo.*
   ⇒ {*nutmeg* hyponym_of *spice*}
   ⇒ {*cinnamon* hyponym_of *spice*}
   ⇒ {*coriander* hyponym_of *spice*}

Using the correct relations extracted without the LSA filter, for each hyponym, the top ten most similar words were collected and tested for having the same hypernym. This resulted in a slight improvement of precision, while the number of relations obtained was ten times higher.

Berland and Charniak (1999) present work on the extraction of **part-of** relations from a corpus, using handcrafted patterns. In a similar fashion to Hearst (1992), seed instances are used to infer linguistic patterns, then used to acquire new relation instances. In the end, the extracted instances are ranked according to their log-likelihood (Dunning, 1993).

Girju and Moldovan (2002) followed Hearst's method to discover lexical-syntactic patterns expressing **causation**. Given that only some categories of nouns (e.g. states of affairs) can be associated with causation, extracted relations were later validated regarding semantic constraints on the relation arguments.

Cimiano and Wenderoth (2007) present an approach for the automatic acquisition of **qualia structures** (Pustejovsky, 1991), which aim to describe the meaning of lexical elements (earlier presented in section 2.2.5 of this thesis). Willing to decrease the problem of data sparseness, they propose looking for discriminating patterns in the Web. For each qualia term, a set of search engine queries for each qualia role is generated, based on known lexical-syntactic patterns. The first 50 snippets returned are downloaded and POS-tagged. Then, patterns, defined over POS-tags, conveying the qualia role of interest, are matched to obtain candidate qualia elements. In the end, the candidates are weighted and ranked according to well-known similarity measures (e.g. Jaccard coefficient, PMI).

The main problem of the aforementioned approaches is that they rely on a finite set of handcrafted rules, though some discovered with the help of automatic procedures, and are therefore vulnerable to data sparseness. Even though Hearst (1992) says that the six proposed patterns occur frequently, they are unlikely to capture all the occurrences of the target relation(s).

About the manual identification of semantic patterns, Snow et al. (2005) add that it is not very interesting and can be biased by the designer. They propose a **supervised** approach, trained with WordNet, to discover hyponymy patterns, and an automatic classifier that decides if a hypernymy relation holds between two nouns. Their procedure works as follows:

[23] See `http://www.natcorp.ox.ac.uk/` (August 2012)

1. Extract all hypernym-hyponym pairs from WordNet.
2. For each pair, find sentences in which both words occur.
3. Parse the sentences, and automatically extract patterns from the obtained trees, which are good cues for hypernymy.
4. Train a hypernymy classifier based on the previous features.

Besides rediscovering the six *Hearst patterns*, which gives a quantitative justification to Hearst's intuition, Snow et al. (2005) were able to discover the following additional patterns:

- `<hyper> like <hypo>`
- `<hyper> called <hypo>`
- `<hypo> is a <hyper>`
- `<hypo>, a <hyper>`

Girju et al. (2006) used a heavily supervised approach as well, based on WordNet, this time for discovering **part-of** relations. The same authors presented a similar approach for the extraction of **manner-of** relations (Girju et al., 2003). However, as WordNet does not contain this kind of relation, the classifier was trained with a corpus where these relations were manually annotated.

Despite quite successful works on supervised LSIE, when there is not an available set of reliable relations of a certain type with a considerable size, a fully supervised approach is not suitable, unless one is willing to create such a set. An alternative is to use a bootstrapping approach, as in the Espresso algorithm (Pantel and Pennacchiotti, 2006), that acquires semantic relations with **minimal supervision**. Pantel and Pennacchiotti (2006)'s main contribution is the exploitation of broad coverage noisy patterns (generic patterns), which increase recall, but have typically low precision (e.g. *X of Y* for part-of). Espresso starts with a small set of seed instances, I, and iterates through three main phases: (i) pattern induction, (ii) pattern ranking/selection, and (iii) instance extraction, briefly described below:

1. Infer a set of surface patterns, P, which are strings that, in the corpus, connect the arguments of the seed instances.
2. Rank each inferred pattern, $p \in P$, according to its reliability, $r_\pi(p)$, given by its average strength of association across each instance $i \in I$:

$$r_\pi(p) = \frac{\sum_{i \in I} \left(\frac{pmi(i,p)}{max_{pmi}} * r_l(i) \right)}{|I|}$$

Here, max_{pmi} is the maximum PMI between all patterns and all instances, given by the ratio between the frequency of p connecting terms x and y, $|x, p, y|$, and all the co-occurrences of x and y times the number of occurrences of p, $pmi(i, p)$:

$$pmi(i,p) = \log \frac{|x,p,y|}{|x,*,y||*,p,*|}$$

The reliability of instance i, $r_l(i)$, is given by:

$$r_l(i) = \frac{\sum_{p \in P'} \left(\frac{pmi(i,p)}{max_{pmi}} * r_\pi(p) \right)}{|P|}$$

3. Select the most reliable patterns.
4. Extract new instances after applying the selected patterns to the corpus. The most reliable instances are added to the seed set.

Ittoo and Bouma (2010) present a study on the acquisition of **part-whole** relations. Using an algorithm inspired by Espresso, they notice that special attention should be given when choosing the seed relations. Given that there are different subtypes of part-whole relations (e.g. member-of, contained-in, located-in), they confirm that, if the initial set of seeds mixes pairs of different subtypes, the algorithm fails to capture these subtypes. But even when they carefully select seeds of only one subtype, part-whole relations of other subtypes are discovered.

Relation extraction from the Web

In the last decade, with the explosion of available electronic contents, researchers felt the need for developing systems that acquire **open-domain facts** from large collections of text, including the Web. Given the size of the data to exploit, these systems, whose final goal was to turn the texts into a large knowledge base, should be robust and scalable enough.

An earlier approach to this problem was the Dual Iterative Pattern Expansion (DIPRE, Brin (1998)), a **weakly-supervised** technique for extracting a structured relation from the Web. DIPRE **bootstraps** from an initial set of seed examples, which is the only required training. For instance, for the extraction of locationOf(*location*, *organisation*) relations, the following seeds could be provided: {*Redmond*, *Microsoft*}, {*Cupertino*, *Apple*}, {*Armonk*, *IBM*}, {*Seattle*, *Boeing*} and {*Santa Clara*, *Intel*}. After finding all close occurrences of the related entities in the collection, patterns where they co-occur are used to extract new pairs holding the same relation. Snowball (Agichtein and Gravano, 2000) is a weakly-supervised system for extracting structured data from textual documents built on the idea of DIPRE, but extending it to incorporate automatic pattern and pair evaluation

KnowItAll (Etzioni et al., 2004) is an autonomous, domain-independent system that extracts facts, concepts, and relationships from the Web. The only domain-specific input to KnowItAll is a set of predicates that constitute its focus and a set of generic domain-independent extractions. KnowItAll uses the extraction patterns with classes (e.g. cities, movies) in order to generate extraction rules specific for each class of instances to extract. A web search engine is queried with keywords in each rule, and the rule is applied to extract information from the retrieved pages. The likelihood of each candidate fact is later assessed with a kind of PMI-IR (Turney, 2001), using an estimation of the search engine hit counts.

Due to the scalability issues of KnowItAll, its authors proposed the paradigm of Open Information Extraction (Banko et al., 2007) (OIE, see section 2.3.2 for more details). OIE systems make a single data-driven pass over a corpus and extract a large set of relational tuples, **without requiring any human input**.

TextRunner (Banko et al., 2007) is a fully-implemented OIE system. In order to get a classifier that labels candidate extractions as trustworthy or not, a small corpus sample is given as input . Then, all tuples that are potential relations are extracted from the corpus. In the last step, relation names are normalised and tuples have a probability assigned. TextRunner is more scalable than KnowItAll, has a lower error rate and, considering only a set of 10 relation types, both systems extract an identical number of relations. However, since TextRunner does not take as input the name of the relations, its complete set of extractions contains more types of relations.

More recently, ReVerb (Etzioni et al., 2011; Fader et al., 2011), a new and more efficient OIE system that **does not need a classifier** was presented. ReVerb is solely based on two constraints: (i) a syntactic constraint requires that the relation phrase matches a POS regular expression (`verb | verb prep | verb word* prep`); (ii) a lexical constraint requires that each relevant relation phrase occurs in the corpus with different arguments. The following illustrate ReVerb extractions:

- {*Calcium*, prevents, *osteoporosis*}
- {*A galaxy*, consists of, *stars and stellar remnants*}
- {*Most galaxies*, appear to be, *dwarf galaxies, which are small*}

The Never Ending Language Learner (NELL, Carlson et al. (2010a)) learns from reading contents on the Web and gets better at reading as it reads the same text multiple times. NELL's starting point is: (i) a set of fundamental categories (e.g. person, sportsTeam, fruit, emotion) and relation types (e.g., *playsOnTeam*(athlete,sportsTeam), *playsInstrument*(musician,instrument)), that constitute an ontology; and (ii) a set of 10 to 15 seed examples for each category and relation. Then, NELL reads web pages continuously, 24 hours a day, for extracting new category instances and new relations between instances, which are used to populate the ontology. The extracted contents are used as a **self-supervised** collection of training examples, used in the acquisition of new discriminating patterns. NELL employs coupled-training (Carlson et al., 2010b), which combines the simultaneous training of many extraction methods. The following are examples of NELL extractions:

- musicArtistGenre(*Nirvana*, *Grunge*)
- tvStationInCity(*WLS-TV*, *Chicago*)
- sportUsesEquip(*soccer*, *balls*)

The main difference between NELL and OIE systems is that NELL learns extractors for a fixed set of known relations, while an OIE system can extract meaningful information from any kind of corpora, on any domain, as relations are not given as a starting point (Etzioni et al., 2011). This has also an impact on the quantity of extracted knowledge. Still, recently, Mohamed et al. (2011) reported how a system like NELL can learn new relation types between already extracted categories.

Kozareva and Hovy (2010) present a **minimally-supervised** method to learn domain **taxonomies** from the Web. It starts by extracting the terms of a given domain, and then induces their taxonomic organisation, without any initial taxonomic information. The acquisition of hypernymy relations relies on two variations of *Hearst patterns*, which provide higher precision, requiring only a root concept and one seed hyponym. The following patterns are used for collecting more relations:

1. `<root> such as <seed> and *`
2. `* such as <term1> and <term2>`, where `term1` and `term2` are hyponyms acquired with the first pattern.

In the taxonomy induction stage, other *Hearst patterns* are used to find evidence on the position of each concept in the taxonomy. In order to identify the hierarchic levels, an algorithm finds the longest path between the root and the other concepts.

Still looking at the Web, a specific resource that has been receiving more and more attention by the IE community is **Wikipedia**, the free collaborative encyclopedia, which is constantly growing. Medelyan et al. (2009) present a survey on IE from Wikipedia. Among the works using Wikipedia, Zesch et al. (2008a) introduce an API for the extraction of knowledge from its English and German version and also Wiktionary; and Wu and Weld (2010) use the Wikipedia infoboxes for training an OIE classifier. Concerning LSIE, Herbelot and Copestake (2006) investigate the extraction of hypernymy relations from Wikipedia; and Veale (2006) captures neologisms from this resource. Most neologisms are hyponyms of its parts (e.g. *hero* is-a *superhero*), or, at least, can be seen as such.

Moreover, there are several public knowledge bases automatically extracted from Wikipedia, including WikiNet (Nastase et al., 2010), YAGO (Suchanek et al., 2007; Hoffart et al., 2011) and DBPedia (Bizer et al., 2009). Although created automatically, DBPedia is manually supervised by the community.

For Portuguese, the first works using Wikipedia as an external source of knowledge include Ferreira et al. (2008), who exploited the first sentences of Wikipedia articles to classify named entities. We (Gonçalo Oliveira et al., 2010a) have also made some experiments on the acquisition of synonymy, hypernym, part-of, purpose-of, and causation relations from the first sentences of the articles, using a set of predefined discriminating patterns. Recently, Págico (Mota et al., 2012; Santos et al., 2012), a joint evaluation on the retrieval of non-trivial information from the Portuguese Wikipedia, was organised. Among the seven participations, two were automatic systems (Rodrigues et al., 2012; Cardoso, 2012) and five were humans (three individuals and two teams).

3.3 Enrichment and Integration of Lexical Knowledge Bases

It was earlier noticed (Hearst, 1992; Riloff and Shepherd, 1997; Caraballo, 1999) that, even though WordNet is a broad-coverage resource, it is **limited** and **incomplete** in many domains, and therefore not enough for several NLP tasks. As a LKB, most of the information in WordNet is about the words and their meanings. Therefore, more than aiming at the creation of new knowledge bases, works on the automatic acquisition of semantic relations used WordNet as a reference for

comparison and claimed that the extracted relations could be used for augmenting it (see e.g. Hearst (1998); Lin and Pantel (2002); Snow et al. (2005); Kozareva and Hovy (2010)). WordNet has also been extended with domain knowledge (Navigli et al., 2004; Pantel, 2005), information from dictionaries (Nastase and Szpakowicz, 2003), and information in its own synset glosses (Harabagiu and Moldovan, 2000; Navigli et al., 2004).

In order to move from a textual structure towards an ontological structure, Pantel (2005) introduced the task of **ontologising**, which aims to associate terms, extracted from text, to their meanings, represented, for instance, as a synset in a wordnet. Pennacchiotti and Pantel (2006) present two methods that take advantage of the structure of WordNet to ontologise relational triples, extracted from text:

- The anchor approach assumes that terms related in the same way to a fixed term are more plausible to describe the same sense. Therefore, to select a suitable synset for a term, it exploits extracted triples of the same type sharing one term argument.
- The clustering approach selects suitable synsets using generalisation through hypernymy links in WordNet.

The output triples of a OIE system can as well be used to create an ontology (Soderland and Mandhani, 2007), with WordNet serving as a map of concepts. To this end, the term arguments are first mapped to synsets that include them and have the most similar context to the triple. The context of a triple contains the words in the sentences from where it was extracted, while the context of the synset contains the words in the synset, in sibling synsets and in direct hyponyms. After this, the relation name is normalised, the logical semantics is formalised, the meta-properties of each relation are learned, a correctness probability is given to each relation and, in the end, an inference engine combines the derived relations with the relations in WordNet.

On the integration of lexical resources with different structures, Kwong (1998) used WordNet as a mediator for bridging the gap between a dictionary and a thesaurus, and Tokunaga et al. (2001) developed a method for augmenting a LKB for Japanese with information in a Japanese dictionary. Furthermore, in the Panacea project (Padró et al., 2011), a platform has been developed to acquire lexical-semantic knowledge from Spanish text and then combining it with existing hand-crafted lexicons.

Much attention has also been given to the integration of WordNet and Wikipedia. For instance, Wikipedia categories have been aligned with WordNet synsets, in order to provide more information on named entities to WordNet (Toral et al., 2008), or to improve the taxonomy of Wikipedia categories (Ponzetto and Navigli, 2009). There is also work on the automatic alignment of Wikipedia articles with WordNet synsets (Ruiz-Casado et al., 2005), aiming to enrich the semantic relations of WordNet (Ruiz-Casado et al., 2007; Ponzetto and Navigli, 2010) or to refine and augment WordNet's sense inventory (Niemann and Gurevych, 2011). Work on linking WordNet and Wikipedia has originated new ontologies, such as YAGO (Suchanek et al., 2007; Hoffart et al., 2011).

A different alternative for increasing the coverage of a knowledge base is to link it to other knowledge bases. On this context, Tonelli and Pighin (2009) have

worked on mapping WordNet and FrameNet, and Shi and Mihalcea (2005) integrated FrameNet, VerbNet and WordNet. For Dutch, Vossen et al. (2008) combine WordNet synsets with the lexical units of a FrameNet-like resource, and map them into a formal ontology.

More recently, UBY – a large-scale unified lexical-semantic resource (Gurevych et al., 2012) has been presented. This project combines several public lexical-semantic resources of English and German in a unique resource, modelled after LMF (Francopoulo et al., 2009), a ISO standard for representing lexicons. The integrated resources include both handcrafted LKBs (WordNet, FrameNet, VerbNet and GermaNet (Kunze and Lemnitzer, 2002)) and other collaboratively created resources (e.g. Wikipedia, Wiktionary).

Multilingual wordnets, such as EuroWordNet (Vossen, 1997), MultiWordNet (Pianta et al., 2002), or BalkaNet (Stamou et al., 2002), are other examples of LKBs of different languages, **aligned** to Princeton WordNet according to an (interlingual) index. BabelNet (Navigli and Ponzetto, 2010, 2012) is another multilingual wordnet, recently made available. The main differences of BabelNet towards other wordnets is that it integrates knowledge from Wikipedia, to which Princeton WordNet is mapped automatically. Machine translation is also used to add lexical information in other languages. One of the strengths of multilingual wordnets is that they enable to perform cross-lingual knowledge based WSD, which opens the door to other cross-lingual tasks.

Finally, WordNet has been linked to **other kinds of ontology**, including the upper ontology SUMO (Pease and Fellbaum, 2010) and the descriptive ontology DOLCE (Gangemi et al., 2010). DBPedia (Bizer et al., 2009) is also linked directly or indirectly to several data sources, including WordNet, OpenCyc, and the collaborative knowledge base Freebase (Bollacker et al., 2008).

3.4 Remarks on this section

Most of the work described in this chapter motivated us and was a source of inspiration in the achievement of our final goal – the automatic construction of Onto.PT, a lexical ontology for Portuguese. The ECO approach, which we propose for creating wordnets automatically, combines three information extraction and integration steps, described in the following chapters of this thesis, namely: (i) LSIE; (ii) discovery of clusters which can be seen as the synsets; and (iii) integration of the extracted relations with the discovered synsets.

Our survey on LKBs showed that Princeton WordNet is the most successful resource of this kind. It was a great inspiration for the creation of LKBs in other languages and, today, several non-English languages have at least one wordnet related project. We have seen that the wide acceptance of the wordnet model lead to the development of a wide range of techniques for exploiting this model in the achievement of NLP tasks, that go from natural language generation to WSD. Therefore, we decided to model Onto.PT as a wordnet, in a sense that it is structured in synsets and semantic relations connecting them. Even though there are wordnet projects in Portuguese, all of them have limitations, which made us believe that we could created something significantly different. Given that ECO is an automatic approach, improvements are mostly related to less effort involved in the resource

creation and higher coverage, with a trade-off on the virtual 100% reliability.

Works on LSIE from dictionaries were another important inspiration. Together with the reasons given in the previous chapter, they reinforced our decision on using dictionaries as the primary source of knowledge. Similarly to what we did, most of those works use pattern based symbollic techniques for extracting information. On the other hand, when it comes to identifying similar words, and eventually concepts described by them, clustering techniques over the distribution of the words on text are more common. So, for discovering synsets, we use those techniques, but considering word co-occurrences in synonymy discriminating patterns.

Besides the automatic creation of knowledge bases from text, in this chapter, we have described works on information extracted for the enrichment of existing knowledge bases. Furthermore, we referred knowledge bases that, for one or another reason, have been either linked or integrated in a unique resource. These are both ways of enriching the knowledge base and improving its coverage.

On our work, we assumedly wanted to integrate as much structured or semi-structured lexical-semantic information there was freely available for Portuguese. We have thus used the synsets of TeP, an existing thesaurus, as a starting point. The thesaurus is then enriched with information from three dictionaries and another smaller thesaurus. But the ECO approach enables an easy integration of knowledge from different heterogeneous sources in Onto.PT, as long as it is represented as triples. Therefore, the future integration of more information, whether it is from other dictionaries, encyclopedias or textual corpora, is quite straightforward. Still, as in some works described in this chapter, semi-supervised learning techniques should be used for extracting knowledge from unstructured text, because, as long as there is a large quantity of text, with enough redundancy, they are more efficient.

In the following chapters of this thesis, we present each step of the ECO approach individually. As they can be used independently in the automatic creation or enrichment of lexical-semantic resources, each of the steps can be seen individually as a contribution of our research.

Chapter 4

Acquisition of Semantic Relations

Our approach for creating a lexical ontology from textual resources, ECO, consists of the combination of three automatic steps that enable the transformation of textual information into a wordnet-like resource. This approach, described by the diagram in figure 4.1, starts by extracting instances of semantic relations between words. Then, synsets are discovered from the synonymy instances. In the end, the arguments of the non-synonymy relations are attached to suitable synsets.

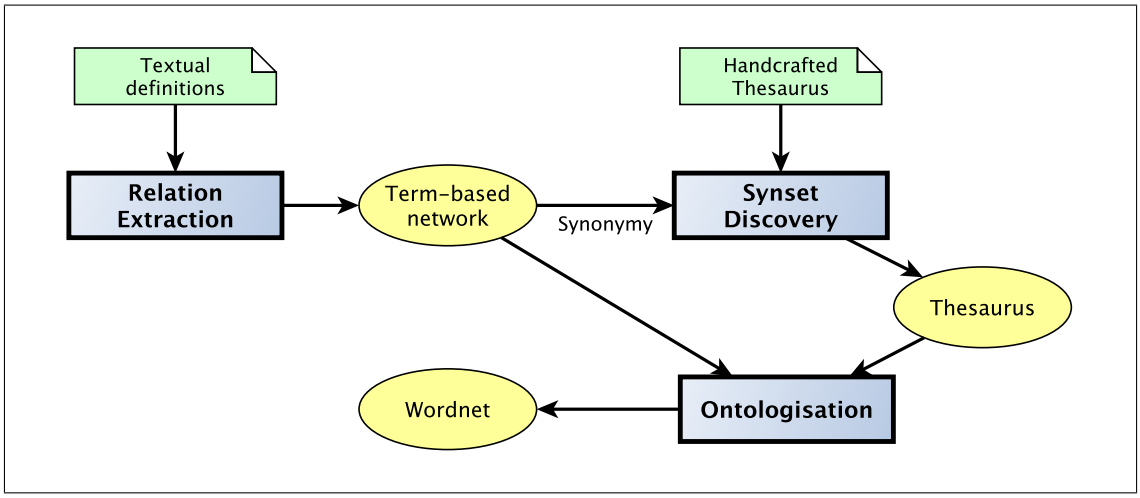


Figure 4.1: Onto.PT construction approach diagram.

At the same time, each step can individually be seen as an independent module that performs a simpler information extraction task. Therefore, in the following chapters, we present each step independently, together with the resource(s) they originate and an experimentation towards their validation:

- This chapter, 4, is dedicated to the automatic acquisition of semantic relations from dictionaries.
- Chapter 5 describes how synonymy networks extracted from dictionaries may be exploited in the discovery of synsets. Given that there is a synset-based resource for Portuguese, this step ended up being integrated in the second part of the following.
- Chapter 6 presents an approach for enriching the synsets of a thesaurus, with synonymy instances extracted from dictionaries. Instances that cannot be

added to synsets are the target of a clustering algorithm, similar to the one presented in chapter 5.

- Chapter 7 proposes several algorithms for moving from term-based semantic relations to relations held between synsets, using only the extracted term-based relations and the discovered synsets.
- After presenting all the steps, chapter 8 shows how they can be combined in the ECO approach, in order to reach our final goal, Onto.PT, a lexical ontology for Portuguese. In the same chapter, an overview of the current version of Onto.PT is provided.

It is possible to integrate any kind of information, from any source, in Onto.PT, as long as it is represented as term-based triples. Still, regarding the goal of creating a broad-coverage lexical ontology, and despite some experiments using Wikipedia (Gonçalo Oliveira et al., 2010a), electronic dictionaries were our main target for exploitation, as in the MindNet project (Richardson et al., 1998; Vanderwende et al., 2005). As referred in section 2, language dictionaries are the main source of general lexical information of a language. They are structured on words and senses and are more exhaustive on this field than other textual resources. At the same time, they are systematic and thus easier to parse.

This chapter describes the extraction of semantic relations from three Portuguese dictionaries, which resulted in the LKB named CARTÃO, a large lexical-semantic network for Portuguese. Part of the work presented here is also reported in Gonçalo Oliveira et al. (2011).

We start this chapter by introducing our approach to the acquisition of term-based relational triples from dictionary definitions. Then, we describe the work performed on the creation of CARTÃO, starting with a brief introduction about the dictionaries used, some issues about their parsing and about the structure of their definitions. After that, we present the contents of CARTÃO, we compare the knowledge extracted from each of the three dictionaries, and evaluate it using different procedures. We end this chapter with a brief discussion on the utility of a LKB structured as CARTÃO.

4.1 Semantic relations from definitions

In our work, the extraction of semantic relations from dictionaries is based on a fixed set of handcrafted rules, as opposing to state-of-the art bootstrapping algorithms that learn relations given a small set of seeds (see more in section 3.2.2). Although our approach is more time-consuming, especially in the construction of the grammars, which have to be manually adapted to new situations, this is not critical for dictionaries. As we will discuss in section 4.2.3, many regularities are preserved along definitions in the same dictionary, and even in different dictionaries. The vocabulary thus tends to be simple and easy to parse. Also, most bootstrapping algorithms rely heavily on redundancy in large collections of text, while dictionaries are smaller and much less redundant. Furthermore, our approach provides higher control over the discriminating patterns.

The extraction of semantic relations is inspired by the construction of PAPEL, reported in Gonçalo Oliveira et al. (2009, 2010b), and consists of one manual step,

where the grammars are created, and two automatic steps. Semantic relations, held between words in the definitions and the definiendum, are extracted after processing dictionary entries. Extracted relation instances are represented as term-based relational triples (hereafter, tb-triples) with the following structure:

`arg1 RELATION_NAME arg2`

A tb-triple indicates that one sense of the lexical item in the first argument (`arg1`) is related to one sense of the lexical item in the second argument (`arg2`) by means of a relation identified by `RELATION_NAME`. For instance:

`animal HIPERONIMO_DE cão` (*animal* HYPERNYM_OF *dog*)

Each step of the extraction procedure is illustrated in figure 4.2, and encompasses the following steps:

1. **Creation of the extraction grammars:** After a careful analysis of the structure of the dictionary definitions, patterns that denote semantic relations are manually compiled into grammars. The rules of the grammars are made specifically for the extraction of relations between words in dictionary definitions and their definiendum.
2. **Extraction of semantic relations:** The grammars are used together with a parser[1] that processes the dictionary definitions. Only definitions of open category words (nouns, verbs, adjectives and adverbs) are processed. In the end, if definitions match the patterns, instances of semantic relations are extracted and represented as tb-triples $t = \{w_1\ R\ w_2\}$ where w_1 is a word in the definition, w_2 is the definiendum, and R is the name of a relation established by one sense of w_1 and one sense of w_2.
3. **Cleaning and lemmatisation:** After extraction, some relations have invalid arguments, including punctuation marks or prepositions. Definitions are thus POS-tagged with the tagger provided by the OpenNLP toolkit[2], using the models for Portuguese[3]. Triples with invalid arguments are discarded[4]. Moreover, if the arguments of the triples are inflected and thus not defined in the dictionary, lemmatisation rules are applied[5].

This procedure results in a set of tb-triples of different predefined types. The resulting set may be formally seen as a term-based directed lexical network (see section 2.2.3). To this end, each tb-triple $t = \{w_1\ R\ w_2\}$ will denote an edge with label R, connecting words w_1 and w_2, which will be the nodes.

[1] We used the chart parser PEN, available from `https://code.google.com/p/pen/` (September 2012)

[2] Available from `http://incubator.apache.org/opennlp/` (September 2012)

[3] See `http://opennlp.sourceforge.net/models-1.5/` (September 2012)

[4] Definitions are not tagged before extraction because the tagger models were trained in corpora text and do not work as well as they should for dictionary definitions. Furthermore, the grammars of PAPEL do not consider tags. Tagging at this stage should only be seen as a complement to the information provided by the dictionary.

[5] The lemmatisation rules were compiled by our colleague Ricardo Rodrigues, and take advantage of the annotation provided by the OpenNLP POS tagger.

1. Part of a grammar, with rules for extracting hypernymy (HIPERONIMO_DE), part-of/has-part (PARTE_DE/TEM_PARTE), and purpose-of (FAZ_SE_COM) relations, and the definitions of an empty head (CABECA_VAZIA):

```
RAIZ ::= HIPERONIMO_DE <&> ...
...
RAIZ ::= CABECA_VAZIA
CABECA_VAZIA ::= parte
...
RAIZ ::= ...  <&> usado <&> para <&> FAZ_SE_COM
RAIZ ::= parte <&> de <&> TEM_PARTE
RAIZ ::= ...  <&> que <&> contém <&> DET <&> PARTE_DE
```

2. Dictionary entries (definiendum, POS, definition) and relations extracted using the previous rules:

```
candeia  nome  utensílio doméstico rústico usado para iluminação, com
               pavio abastecido a óleo
  →  utensílio HIPERONIMO_DE candeia
  →  com FAZ_SE_COM candeia
  →  iluminação FAZ_SE_COM candeia
espiga  nome  parte das gramíneas que contém os grãos
  →  espiga PARTE_DE gramíneas
  →  grãos PARTE_DE espiga
```

3. POS-tagging, cleaning and lemmatisation:

```
candeia  nome  utensílio#n doméstico#adj rústico#adj usado#v-pcp
               para#prp iluminação#n ,#punc com#prp pavio#n
               abastecido#v-pcp a#prp óleo#n
  →  utensílio HIPERONIMO_DE candeia
  →  iluminação FAZ_SE_COM candeia
espiga  nome  parte#n de#prp as#art gramíneas#n que#pron-indp
              contém#v-fin os#art grãos#n
  →  espiga PARTE_DE gramínea
  →  grão PARTE_DE espiga
```

Figure 4.2: Extraction of semantic relations from dictionary definitions.

4.2 A large lexical network for Portuguese

The relation acquisition procedure was used to create CARTÃO (Gonçalo Oliveira et al., 2011), a large term-based lexical-semantic network for Portuguese, extracted from dictionaries. Regarding the incompleteness of dictionaries (Ide and Véronis, 1995)), we exploited not one, but three electronic dictionaries of Portuguese, namely:

- Dicionário PRO da Língua Portuguesa (DLP, 2005), indirectly with the results of the project PAPEL;
- Dicionário Aberto (DA) (Simões and Farinha, 2011; Simões et al., 2012);
- Wiktionary.PT[6].

[6]Available from `http://pt.wiktionary.org/` (September 2012)

In this section, after introducing the dictionaries, we describe the steps towards the creation of CARTÃO, starting with parsing the dictionaries and transformation of their contents into a common data format. Then, we present some of the most common regularities across the three dictionaries. After presenting the contents of CARTÃO, the section ends with the steps performed towards its evaluation.

4.2.1 About the dictionaries

DLP is a proprietary dictionary, developed and owned by the Portuguese publisher Porto Editora[7]. DLP was exploited in the scope of the project PAPEL, through a protocol celebrated between Porto Editora and Linguateca[8], the language resource center responsible for the development of PAPEL. Our strict collaboration with Linguateca resulted in new versions of PAPEL, including the current version, PAPEL 3.0, during the work described in this thesis.

DA is the electronic version of a Portuguese dictionary, originally published in 1913. DA contains about 128,000 entries and is publicly available in several formats, including PDF, plain text, and a SQL database[9]. As the contents of DA used an old orthographic form, its orthography is currently being modernised (see details in Simões et al. (2012)). In the current version of CARTÃO, we have used the second modernisation revision, from 19th October 2011.

Wiktionary is a collaborative initiative ran by the Wikimedia foundation with the aim of providing several collaborative multi-lingual dictionaries. Besides typical information in dictionaries, such as POS, translations, pronunciation and etymology, some Wiktionary entries have as well information on semantic relations, including synonyms, antonyms or hypernyms. However, as a project dependent on volunteers, that kind of information is still very incomplete in the Portuguese version, Wiktionary.PT. Wiktionaries are freely available in XML dumps[10], where the entries are described in *wikitext*. In the current version of CARTÃO, we have used the 8th December 2011 dump of Wiktionary.PT[11], which contains about 210,000 entries, including 115,000 which were automatically identified as having at least one definition of a Portuguese word. As a multi-lingual dictionary, the rest of the entries were restricted to words in other languages.

4.2.2 Definitions format

In order to parse the definitions of the three dictionaries with the same programs and grammars, we converted all of them into a friendlier data format.

Transforming the DA's database and XML information of each entry into this format was quite straightforward. The only problem was that, even though DA's orthography is being modernised, the definienda are kept in their original form. So, after the extraction process, a decision is made automatically, in order to either keep, change or discard the extracted tb-triples. If the arguments do not match any disused sequences, they are kept. Otherwise, they are changed according to the

[7]See `http://www.portoeditora.pt/` (September 2012)
[8]See `http://www.linguateca.pt/` (September 2012)
[9]See `http://www.dicionario-aberto.net/` (September 2012)
[10]See http://dumps.wikimedia.org/
[11]Wiktionary database dumps are available through `http://dumps.wikimedia.org/`

suggestions in Simões et al. (2010). However, in order to minimise the probability of generating invalid lemmas, if they do not exist in TeP nor PAPEL, the tb-triple is discarded.

For handling the *wikitext* of the Wiktionary.PT dump, we developed a specific parser (Anton Pérez et al., 2011). Although there is an available API, JWKTL[12], for processing Wiktionary (Zesch et al., 2008a), it is only compatible with the English and German versions of the resource. The main problem is that different language editions of Wiktionary use distinct delimiter elements to represent the information of each entry, so every Wiktionary parser needs to be adapted according to the language edition. Since the source code of JWKTL was not available, we could not adapt it for Wiktionary.PT.

In the dictionary conversion process, only definitions of open-category words were used, and changed to one common notation: `nome` for nouns, `verbo` for verbs, `adj` for adjectives and `adv` for adverbs. The format adopted for representing the dictionaries contains a definition per line. Before the definition, we include the definiendum and its POS, as in the following definition for the word *coco* (coconut):

```
coco    nome    fruto gerado pelo coqueiro, muito usado para se fazer
                doces e para consumo de seu líquido
```

In this format, words with more than one definition originate more than one line. Also, since Wiktionary provides synonymy lists for some of its entries, we transformed these lists in definitions with only one word, as in the following example for the synonyms of the word *bravo* (brave):

```
Sinónimos: corajoso, destemido  ⇒ | bravo  adj  corajoso
                                 | bravo  adj  destemido
```

After the conversion of DA and Wiktionary.PT we obtained about 229,000 and 72,000 definitions, respectively for each dictionary. We do not have direct access to DLP, but we can say that it contains 176,000 definitions which gave origin to, at least, one relation.

Wiktionary.PT is the smaller resource, which resulted in the lowest number of definitions among the three dictionaries. However, before collecting these definitions, we discarded: (i) definitions corresponding to words in other languages; (ii) definitions of closed-category and inflected words (including verbal forms); (iii) definitions in entries with alternative syntaxes, not recognised by our parser. As Wiktionaries are created by volunteers, often not experts, and because there is no standard syntax for representing Wiktionary entries in *wikitext*, the structure of the entries is fairly inconsistent. It is thus impossible to develop a parser to handle all syntax variations, and thus 100% reliable. This problem seems to be common to other editions of Wiktionary, as it is referred by other authors (e.g. Navarro et al. (2009)).

4.2.3 Regularities in the Definitions

One of the main reasons for using dictionaries in the automatic acquisition of lexical-semantic relations is that they typically use simple and systematic vocabulary, suitable for being exploited in information extraction. Having this in mind, during the creation of PAPEL, we developed a set of grammars, with lexical-syntactic patterns

[12]See `http://www.ukp.tu-darmstadt.de/software/jwktl/` (September 2012)

that, in DLP, frequently denote the relations we wanted to extract[13]. The grammars were created manually, after the analysis of the structure and vocabulary of the DLP definitions, and the identification of regularities.

In order to reproduce the grammar creation procedure for extracting relations from the other dictionaries, we also analysed the structure of their definitions. This analysis showed that most of the regularities used in the DLP definitions were preserved in DA and Wiktionary.PT, which meant that the grammars of PAPEL could be reused with minor changes. Table 4.1 shows the frequency and the semantic relation usually denoted by the most productive n-grams in the three dictionaries, which are those frequent and suitable for exploitation in the automatic extraction of semantic relations. In the referred table, some patterns extract the direct relation (e.g. part-of) and others the inverse relation (e.g. has-part) but, during the extraction procedure, all relations are normalised into the type agreed as the direct (e.g. *keyboard* has-part *key* is changed to *key* part-of *keyboard*).

The few changes we made to the original grammars of PAPEL, include:

- The pattern `o mesmo que` was used in the extraction of synonymy relations.
- The keywords `natural` and `habitante` could change their order in the extraction of place-of relations.
- Brazilian Portuguese specific orthography was considered in some patterns, as they occurred in Wiktionary.PT. Words such as `gênero` and `ato` were used, respectively, for the extraction of hypernymy and causation relations.

In addition to the static patterns in table 4.1, two other productive rules were included in the grammars for extracting relations from the three dictionaries:

- Synonymy can be extracted from definitions consisting of only one word or a enumeration of words. See the following example:

  `talhar verbo gravar, cinzelar ou esculpir`
  → *gravar* synonym-of *talhar*
  → *cinzelar* synonym-of *talhar*
  → *esculpir* synonym-of *talhar*

- As most noun definitions are structured on a *genus* and *differentia* (see section 3.2.1), we identify the *genus* as a hypernym of the definiendum, which might eventually be modified by an adjective. The following are examples of the application of this rule:

  `islandês nome língua germânica falada na Islândia`
  → *língua* hypernym-of *islandês*

  `pantera nome grande felino de o gênero Panthera`
  → *felino* hypernym-of *pantera*

The second rule does not apply, however, when the definition starts by a so called "empty head" (Chodorow et al., 1985; Guthrie et al., 1990), which is usually exploited in the extraction of other relations. The list of considered "empty heads" includes words such as `acto`, `efeito` (used for the extraction of the causation relation), `qualidade` (has-quality), `estado` (has-state), `parte` (part-of),

[13]The grammars of PAPEL are freely available from `http://www.linguateca.pt/PAPEL/` (September 2012)

Pattern	POS	Frequency			Relation
		DLP	DA	Wikt.PT	
o mesmo que (the same as)	Noun	0	10,627	1,107	Synonymy
a[c]to ou efeito de (act or effect of)	Noun	3,851	2,501	645	Causation
pessoa que (person who)	Noun	1,320	47	329	Hypernymy
aquele que (one who)	Noun	1,148	3,357	545	Hypernymy
conjunto de (set of)	Noun	1,004	316	298	Member-of
espécie de (species of)	Noun	798	2,846	223	Hypernymy
género/gênero de (kind of)	Noun	29	4,148	48	Hypernymy
variedade de (variety of)	Noun	455	621	52	Hypernymy
[a] parte do/da (part of the)	Noun	445	433	107	Has-part
qualidade de (quality of)	Noun	777	775	126	Has-quality
qualidade do que é (quality of what is)	Noun	663	543	105	Has-quality
estado de (state of)	Noun	299	223	73	Has-state
natural ou habitante de/da/do (inhabitant or natural of)	Noun	536	0	79	Place-of
instrumento[,] para (instrument for)	Noun	94	284	25	Purpose-of
.. produzid[o/a] por/pel[o/a] (produced by)	Noun	155	146	60	Produtor
o mesmo que (the same as)	Verb	0	166	97	Synonymy
fazer (to do)	Verb	1,680	1,294	364	Has-cause
tornar (to make)	Verb	1,359	1,672	266	Has-cause
ter (to have)	Verb	467	519	139	Property-of
o mesmo que (the same as)	Adjective	0	2,685	197	Synonymy
relativo a/à/ao (relative to [the])	Adjective	1,236	5.554	1,063	Has-property
que se (that)	Adjective	1,602	1,599	485	Property-of
que tem (that has)	Adjective	2,698	4,291	477	Part-of/ Property-of
diz-se de (it is said about)	Adjective	2,066	738	313	Has-property
relativo ou pertencente (relative or belonging)	Adjective	1,647	9	61	Has-member/ Has-property
habitante ou natural de (inhabitant or natural of)	Adjective	0	0	189	Place-of
que não é/está (which is not)	Adjective	485	608	98	Antonymy
de modo (in a way)	Adverb	398	2.261	109	Manner-of
de maneira (in a manner)	Adverb	49	9	36	Manner-of
de forma (in a manner)	Adverb	30	3	19	Manner-of
o mesmo que (the same as)	Adverb	0	182	21	Synonymy

Table 4.1: Frequent and productive patterns in the dictionary definitions.

`membro`, `conjunto`, `grupo` (member-of), `maneira` (manner-of), `tipo`, `forma`, `género`, `espécie`, or `variedade` (hypernymy).

In order to make extraction more efficient, we separated the grammar files according to the POS of the definitions they were supposed to parse. The current version of the grammars contains 371 non-terminal symbols and 1,714 productions.

4.2.4 Contents

Using the aforementioned procedure and grammars, about 134,000 tb-triples were extracted from DA and about 57,300 from Wiktionary.PT, while PAPEL 3.0 contains about 190,000 tb-triples. This means that CARTÃO is currently the largest lexical-semantic resource of this kind, for Portuguese. It augments PAPEL 3.0 with 72% new tb-tiples and 52% new lexical items. The obtained numbers confirm that, even though dictionaries intend to cover the whole language, they are incomplete. Using more than one dictionary is thus the best way of obtaining a broader LKB, with a similar effort.

Extracted relations

Table 4.2 shows the number of extracted tb-triples, according to the source dictionary and the type of relation. As in PAPEL, each relation has different sub-types, according to the POS of its arguments. Real examples of each sub-type are presented in table 4.3. A textual description of each relation can be found in appendix A.

Table 4.2 shows that about 40% of the extracted tb-triples are instances of synonymy relations, while hypernymy tb-triples are about 33%. From the remaining relations, the highest percentage is that of property-of tb-triples, which are about 12%. All the other types of relations make up just about 15% of the resource.

To give an idea on the contribution of each dictionary in terms of tb-triples and their intersections, we present figure 4.3. It shows that there is not much redundancy across the dictionaries, as only a minority of tb-triples (about 1.8%) was extracted from all the three.

Table 4.4 gives a different perspective on the contribution of each dictionary to CARTÃO. The sets of triples extracted from each dictionary are compared in pairs, according to their similarity and novelty of one given another. For this purpose, we use the measures below:

$$Sim(A,B) = Jaccard(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{4.1}$$

$$Novelty(A,B) = \frac{|A| - |A \cap B|}{|A|} \tag{4.2}$$

As expected, given their different sizes, the highest novelties are those of DLP and DA in relation to Wiktionary.PT. Nevertheless, all the resources have high novelty (always higher than 70%) towards each other.

Covered lemmas

The lemmas in the arguments of the tb-triples were compared as well. Table 4.5 contains the number of different lemmas in the arguments of the tb-triples extracted

Relation	Args.	Quantity			
		DLP	DA	Wiktionary	Unique
Synonym-of	n,n	40,306	25,046	13,812	67,620
	v,v	18,927	11,113	4,650	28,108
	adj,adj	21,726	10,505	6,611	32,364
	adv,adv	1,178	1,199	277	2,286
Hypernym-of	n,n	62,591	44,777	17,068	97,924
Part-of	n,n	2,424	1,146	614	3,893
	n,adj	3,033	3,414	520	5,872
Member-of	n,n	5,679	928	1,161	7,328
	n,adj	77	26	25	120
	adj,n	968	80	138	1,071
Contained-in	n,n	216	124	53	381
	n,adj	176	124	34	287
Material-of	n,n	335	513	146	888
Causation-of	n,n	951	193	317	1,423
	n,adj	17	8	5	25
	adj,n	494	148	173	748
	n,v	40	17	6	60
	v,n	6,256	7,140	1,631	10,664
Producer-of	n,n	910	605	333	1,741
	n,adj	49	26	6	77
	adj,n	352	236	37	515
Purpose-of	n,n	3,659	2,353	1,442	6,978
	n,adj	56	40	9	88
	v,n	4,609	2,230	1,610	7,824
	v,adj	236	204	27	374
Has-quality	n,n	740	465	87	1,055
	n,adj	888	667	128	1,273
Has-state	n,n	265	118	44	376
	n,adj	129	102	23	220
Property-of	adj,n	6,287	5,024	1,793	10,652
	adj,v	17,718	11,076	3,569	27,902
Antonym-of	adj,adj	388	410	59	684
Place-of	n,n	834	405	601	1,483
Manner-of	adv,n	795	1,537	164	2,172
	adv,adj	345	1,624	135	1,854
Manner without	adv,n	116	147	16	250
	adv,v	6	5	3	13
Total		191,131	133,783	57,328	326,694

Table 4.2: Quantities and types of extracted relations.

from each dictionary, distributed according to their POS. The majority of the lemmas are nouns. Then, for DLP, the most represented POS are verbs, and then adjectives. On the other hand, there were more adjectives than verbs extracted from DA and Wiktionary.PT. DLP is the dictionary that provides more lemmas to CARTÃO, just slightly more than DA. But the tb-triples extracted from DA include more nouns and two times more adverbs than those extracted from DLP.

Figure 4.4 represents the contribution and overlap of each resource, regarding the covered lemmas. Table 4.6 has the same measures that were calculated for the sets of tb-triples, this time comparing the lemmas in the tb-triples extracted from each resource. If compared to the values for the tb-triples, similarities between lemmas are higher and novelties are lower. Still, novelties are always higher than 35%, which

Relation	Args.	Example	
Synonym-of	n,n	*alegria,satisfação*	(joy,satisfaction)
	v,v	*esticar,estender*	(to_extend,to_stretch)
	adj,adj	*racional,filosófico*	(rational,philosophical)
	adv,adv	*imediatamente,já*	(immediately,now)
Hypernym-of	n,n	*sentimento,afecto*	(feeling,affection)
Part-of	n,n	*núcleo,átomo*	(nucleus,atom)
	n,adj	*vício,vicioso*	(addiction,addictive)
Member-of	n,n	*aluno,escola*	(student,school)
	n,adj	*coisa,coletivo*	(thing,collective)
	adj,n	*rural,campo*	(rural,country)
Contained-in	n,n	*tinta,tinteiro*	(ink,cartridge)
	n,adj	*óleo,oleoso*	(oil,oily)
Material-of	n,n	*folha_de_papel,caderno*	(sheet_of_paper,notebook)
Causation-of	n,n	*vírus,doença*	(virus,disease)
	n,adj	*paixão,passional*	(passion,passional)
	adj,n	*horrível,horror*	(horrible,horror)
	n,v	*fogo,fundir*	(fire,to_melt)
	v,n	*mover,movimento*	(to_move,movement)
Producer-of	n,n	*oliveira,azeitona*	(olive_tree,olive)
	n,adj	*fermentação,fermentado*	(fermentation,fermented)
	adj,n	*fonador,som*	(phonetic,sound)
Purpose-of	n,n	*sustentação,mastro*	(support,mast)
	n,adj	*habitação,habitável*	(habitation,inhabitable
	v,n	*calcular,cálculo*	(to_calculate,calculation)
	v,adj	*comprimir,compressivo*	(to_compress,compressive)
Has-quality	n,n	*mórbido,morbidez*	(morbid,morbidity)
	n,adj	*assíduo,assiduidade*	(assiduous,assiduity)
Has-state	n,n	*exaltação,desvairo*	(exaltation,rant)
	n,adj	*disperso,dispersão*	(scattered,dispersion)
Place-of	n,n	*Equador,equatoriano*	(Ecuador,Ecuadorian)
Manner-of	adv,n	*ociosamente,indolência*	(idly,indolence)
	adv,adj	*virtualmente,virtual*	(virtually,virtual)
Manner without	adv,n	*prontamente,demora*	(promptly,delay)
	adv,v	*seguido,parar*	(straight,to_stop)
Antonym-of	n,n	*direito,torto*	(straight,crooked)
Property-of	adj,n	*daltónico,daltonismo*	(daltonic,daltonism)
	adj,v	*musculoso,ter_músculo*	(beefy,to_have_muscle)

Table 4.3: Examples of extracted relations.

A \ B	DLP		DA		Wikt.PT	
	Sim	Nov	Sim	Nov	Sim	Nov
DLP			0.13	0.81	0.06	0.93
DA	0.13	0.72			0.06	0.92
Wikt.PT	0.06	0.77	0.06	0.81		

Table 4.4: Similarity (Sim) and novelty (Nov) of the triples extracted from each dictionary

shows that, besides different dictionaries describing different relations, more than one third of new vocabulary was collected from each dictionary, if compared to the other dictionaries.

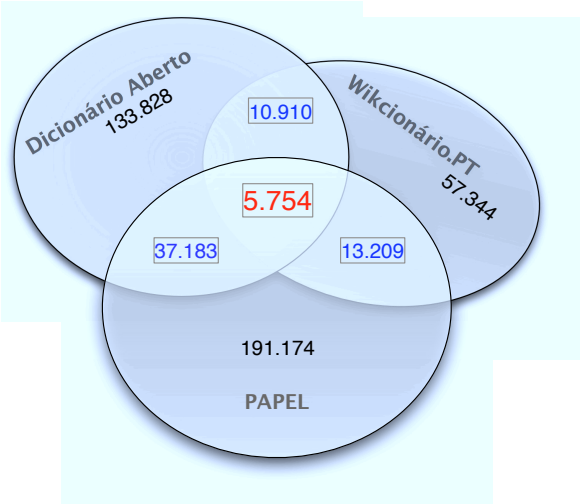


Figure 4.3: Number of tb-triples according to the source dictionary, including the intersections of those extracted from each pair of dictionaries and, in the center, those extracted from all the three dictionaries.

POS	DLP	DA	Wikt.PT	Total
Nouns	55,769	59,879	23,007	89,895
Verbs	22,440	16,672	6,932	32,572
Adjectives	22,381	18.563	7,113	29,964
Adverbs	1,376	3,073	473	3,443
Total	101,966	98,187	37,525	155,187

Table 4.5: Unique lemmas in the extracted tb-triples, according to dictionary

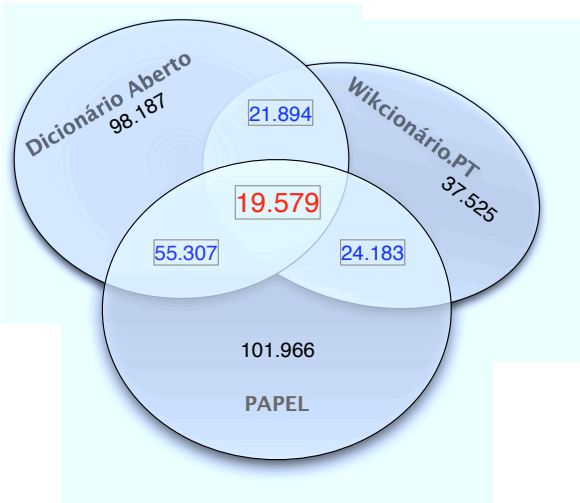


Figure 4.4: Number of lemmas in the tb-triples extracted from each dictionary, including the intersections of lemmas extracted from each pair of dictionaries and, in the center, the number of lemmas extracted from the three dictionaries.

4.2.5 Evaluation

The first validation exercises of CARTÃO were performed automatically. This option relied on the fact that manual evaluation is both time consuming and tedious. Also, it is often subjective and hard to reproduce. Nevertheless, we ended up per-

A \ B	DLP		DA		Wikt.PT	
	Sim	Nov	Sim	Nov	Sim	Nov
DLP			0.38	0.46	0.21	0.76
DA	0.38	0.44			0.19	0.78
Wikt.PT	0.21	0.36	0.19	0.42		

Table 4.6: Similarity (Sim) and novelty (Nov) of the sets of extracted tb-triples, regarding the included lemmas.

forming a manual evaluation of some of the relation types. Here, we present and discuss the three independent stages of the evaluation of CARTÃO:

- First, we took advantage of existing handcrafted Portuguese thesauri to evaluate the coverage of the lemmas and the coverage of synonymy.
- Then, we validated CARTÃO in a similar fashion to what had been done for PAPEL (see Gonçalo Oliveira et al. (2009, 2010b)) – we used a set of discriminating patterns for some of the relations in CARTÃO, and searched in a corpus for occurrences of tb-triples of those relations.
- We finally performed the manual evaluation of CARTÃO.

Coverage of lemmas by handcrafted thesauri

The coverage of the lemmas in CARTÃO was measured after comparing them with the lemmas in TeP 2.0 (Maziero et al., 2008) and OpenThesaurus.PT (OT.PT), two handcrafted Portuguese thesauri, presented in section 3.1.2. Table 4.7 shows the coverage of the lemmas in each of the resources that integrate CARTÃO by both thesauri. Between about 21% (nouns of DA) and 60% (adjectives of Wiktionary.PT) of the lemmas are covered by TeP. On the other hand, due to its size, for OT.PT these numbers are between 3% (adverbs of DA) and 31% (adjectives of Wiktionary.PT). Considering just TeP, there is a higher proportion of covered adjectives and adverbs, as compared to the proportion of covered nouns. The low proportion of DA adverbs and the high proportion of Wiktionary.PT covered nouns are the exceptions.

The tb-triples from Wiktionary.PT have a higher proportion of covered lemmas for all categories, possibly due to the collaborative nature of this resource. Wiktionary.PT is still growing and is created by volunteers, which are usually not experts, while DLP and DA are (or were) commercial dictionaries, created by lexicographers. Therefore, while DLP and DA, besides more common vocabulary, include more formal and less conventional definitions, Wiktionary.PT tends to use more conventional vocabulary. It is worth mentioning that Wiktionary.PT contains several definitions written in Brazilian Portuguese, which is the variant targeted by TeP. This also contributes to a higher proportion of lemmas covered by the TeP.

Coverage of synonymy by handcrafted thesaurus

As TeP is manually created and structured around the synonymy relation, we used it as a gold standard for evaluating the synonymy tb-triples of CARTÃO. We remind that 40% of CARTÃO are synonymy tb-triples (see section 4.2.4). OT.PT was not used because it is too small and it is a collaborative resource, not created by experts.

POS	TeP					
	PAPEL		DA		Wikt.PT	
Nouns	13,137	23.6%	12,701	21.2%	8,079	35.1%
Verbs	6,029	26.9%	5,835	35.0%	3,138	45.3%
Adjectives	9,104	40.7%	8,264	44.5%	4,265	60.0%
Adverbs	574	41.7%	683	22.2%	264	55.8%
POS	**OT**					
	PAPEL		**DA**		**Wikt.PT**	
Nouns	5,736	10.3%	5,532	9.2%	4,440	19.3%
Verbs	2,731	12.2%	2,644	15.9%	1,977	28.5%
Adjectives	3,249	14.5%	2,846	15.3%	2,256	31.7%
Adverbs	94	6.8%	94	3.1%	79	16.7%

Table 4.7: Coverage of lemmas by handcrafted Portuguese thesauri.

Table 4.8 shows the presence of each synonymy tb-triple of CARTÃO in TeP – if TeP has at least one synset that contains both arguments of a synonymy tb-triple, we consider that the tb-triple is covered by TeP. The proportion of covered tb-triples is shown for synonymy tb-triples in each of the three resources (Total), and also considering only tb-triples where both arguments exist in TeP (InTeP). Synonymy coverage according to the POS is consistent for the three resources. It is higher for synonymy among verbs, followed by synonymy between nouns and adjectives, in this order. Similarly to the coverage of lemmas, the proportion of synonymy tb-triples covered by TeP is also higher for Wiktionary.PT.

POS	PAPEL			DA			Wikt.PT		
	Covered	Total	InTeP	Covered	Total	InTeP	Covered	Total	InTeP
Nouns	11,920	30.0%	56.2%	6,821	27.2%	41.4%	4,126	29.9%	50.4%
Verbs	10,063	53.1%	83.5%	5,927	53.3%	76.2%	2,532	54.3%	78.5%
Adjs	8,506	39.2%	69.7%	4,891	46.6%	66.9%	2,903	43.9%	71.8%
Advs	267	22.7%	38.1%	208	17.3%	27.6%	131	32.9%	47.3%

Table 4.8: Synonymy coverage by TeP.

Besides giving an idea on the coverage of the CARTÃO, these numbers show that the public handcrafted thesauri are an additional source of synonymy relations. And given their manual creation, confidence on their contents is high.

As for measuring the quality of CARTÃO, on the one hand, these numbers can also be seen as cues. As we confirmed during the manual evaluation, the quality of synonymy in CARTÃO is much higher than this comparison shows. The different variants of Portuguese targeted by DLP and DA, and TeP might play an important role on this difference.

Relation support in textual corpora

The validation of parts of CARTÃO by querying corpora is the reproduction of a similar procedure that has been performed for PAPEL 1.1 (see Gonçalo Oliveira et al. (2009)) and 2.0 (see Gonçalo Oliveira et al. (2010b)). This validation is based on a set of patterns that, in corpora, typically denote the relation to validate. Those discriminating patterns are used to transform each tb-triple in several corpus queries. If there is at least one occurrence of one of the patterns for a given relation,

connecting the arguments of a tb-triple with that relation, we consider that the corpus supports the relation. Otherwise, the relation is not supported. The obtained results should however not be confused with the precision of the extracted relations. Especially considering the following reasons, which can also be seen as arguments that support the use of dictionaries for the creation of broad-coverage LKBs:

- A corpus is a resource with limited knowledge.
- There are many ways of expressing a semantic relation in text, which makes it impossible to encode all patterns and all possible variations. Although they frequently denote the same relation, some discriminating patterns may be seen as ambiguous, as they may sometimes denote a different relation.
- Several types of relations are dictionary-specific, and are thus not expected to be explicitly expressed in corpora text. This happens, for instance, for relations connecting nouns and verbs, that imply the nominalisation of the verb, as in *aumentar* causation-of *aumento* (to_augment causation-of augmentation).
- There are studies (Dorow, 2006) showing that synonymous words tend not to co-occur frequently in corpora, especially in the same sentence. This idea is consistent with the one sense per discourse assumption (Gale et al., 1992), given that, especially in domain specific texts, the author tends to use always the same word for referring to the same concept. On the other hand, synonymous words tend to occur in similar contexts.

Nevertheless, the obtained results give us an idea on the utilisation of the extracted relations in unstructured text. Furthermore, using the same set of patterns and the same corpus, the results are an indicator of the relations applicability, which may be used in the comparison of resources structured on lexical-semantic relations.

Given the aforementioned limitations, only four types of relations were validated, all of them between nouns. We used the newspaper corpus CETEMPúblico (Santos and Rocha, 2001; Rocha and Santos, 2000), where we searched for all the hypernymy, part-of, member-of, and purpose-of relations, extracted from the three dictionaries. The list of discriminating patterns used was a new version of that used for validating PAPEL 2.0, and includes the patterns used in VARRA (Freitas et al., 2012), an online service for searching for semantic relations in context.

Table 4.9 presents the results of the automatic validation. First, it shows the number, and the proportion it represents, of all tb-triples of the validated relations whose arguments co-occur in at least one sentence of CETEMPúblico (CoocArgs). It shows as well the same values for the tb-triples supported by the corpus (Supported).

The validation results show that the proportion of tb-triples with arguments co-occurring in the corpus is never higher than 37.5% (hypernymy in Wiktionary.PT), nor lower than 17.5% (hypernymy in DA). Curiously, the maximum and the minimum are obtained for the same type of relation, in a different resource. The proportion of member-of relations with co-occurring arguments is the lowest both for PAPEL and Wiktionary.PT. This might occur because this relation is the one with more supported tb-triples.

In the three resources, the proportion of supported tb-triples is always higher for the member-of relation, and lower for purpose-of. We believe that the low proportion of supported purpose-of relations is explained by the fact that this relation is not

Relation	PAPEL			
	CoocArgs		Supported	
Hypernymy	13.724	21,9%	4.098	29,7%
Part-of	573	23,6%	186	32,5%
Member-of	1.089	19,2%	464	42,6%
Purpose-of	1.017	27,8%	164	16,1%
Relation	**DA**			
	CoocArgs		**Supported**	
Hypernymy	7.846	17,5%	2.255	28,7%
Part-of	247	21,6%	81	32,8%
Member-of	303	32,7%	109	36,0%
Purpose-of	473	20,1%	65	13,7%
Relation	**Wiktionary.PT**			
	CoocArgs		**Supported**	
Hypernymy	6.405	37,5%	2.086	32,6%
Part-of	226	36,8%	94	41,6%
Member-of	317	27,3%	147	46,4%
Purpose-of	498	34,5%	75	15,1%

Table 4.9: Relations coverage by the corpus.

as semantically well-defined as the other three. Also, this relation is probably less frequently present in text, and there are more ways of expressing it. From the hypernymy and part-of tb-triples from PAPEL and DA whose arguments co-occur in CETEMPúblico, about 30% are supported. Once again, possibly due to its size and collaborative nature, higher proportions are obtained for Wiktionary.PT.

In order to give a clearer look on this validation, table 4.10 has some examples of sentences that support the extracted tb-triples. In the same sentences, the discriminating patterns are in bold.

Manual evaluation

Besides the validation based on thesaurus and corpus support (Gonçalo Oliveira et al., 2009, 2010b), as well as a simple coverage exercise against other resources (Santos et al., 2010), the only extensive human evaluation of PAPEL, which we are aware of, is that performed by Prestes et al. (2011). Before using PAPEL (presumably version 1.1), they looked at the definitions in Portuguese dictionaries and queried online search engines for obtaining real contexts where the words co-occurred. They indicate that 20,096 synonymy relations (between nouns) and 40,614 hypernymy relations were collected for their resource. While the collected relations were surely correct, they do not go deeper on the evaluation results, which prevents us from taking conclusions on the causes of discarding the remaining relations.

Therefore, we decided to perform the manual evaluation of CARTÃO, which includes PAPEL 3.0 and the tb-triples extracted from DA and Wiktionary.PT. For this purpose, we asked two human judges to independently classify tb-triples of the most frequent types of extracted relations, more precisely: synonymy between nouns, synonymy between verbs, hypernymy, member between nouns, causation between a verb and a noun, purpose between a verb and a noun, and property between an adjective and a verb. For each evaluated relation, we randomly selected a set with 300 different tb-triples – one third came from DLP, another third from DA and

Tb-triple	Supporting sentence
língua hypernym-of *alemão* (*language* hypernym-of *german*)	*As iniciativas deste gabinete passam geralmente pela promoção de conferências, exposições, workshops e aulas de* ***línguas, como o inglês, alemão*** *ou japonês.* (The initiatives of this office are generally for the promotion of conferences, exhibitions, workshops and classes in languages like English, German or Japanese.)
ciência hypernym-of *paleontologia* (*science* hypernym-of *paleontology*)	*A* ***paleontologia é uma ciência*** *que depende do que se descobre.* (Paleontology is a science that depends on its own discoveries.)
rua part-of *quarteirão* (*street* part-of *block*)	*De resto, o* ***quarteirão formado pelas ruas*** *de São João e de Mouzinho da Silveira está, por esse motivo, assente em estacas de madeira...* (Moreover, the block formed by the streets of São João and Mouzinho da Silveira is, because of that, built on wooden stacks...)
mão part-of *corpo* (*hand* part-of *body*)	*As* ***mãos são a parte do corpo*** *mais atingida (29,7%).* (*The hands are the most affected part of the body (29.7%).*)
pessoa member-of *comissão* (*person* member-of *committee*)	*A* ***comissão é constituída por pessoas*** *que ficaram marcadas pela presença de Dona Amélia: ...* (The committee consists of people who were marked by the presence of Dona Amélia: ...)
lobo member-of *alcateia* (*wolf* member-of *pack*)	*Mech e os seus colegas constataram que alguns dos cheiros contidos nas marcas de urina servem para os* ***lobos de uma alcateia*** *saberem por onde andou o lobo que deixou as marcas ...* (Mech and his colleagues found that some of the smells of urine contained in the marks are for a pack of wolves to know where the wolf that left the marks has been...)
transporte purpose-of *embarcação* (*transport* purpose-of *ship*)	*... onde foi descoberto o resto do casco de uma* ***embarcação presumivelmente utilizada no transporte*** *de peças de cerâmica ...* (... where the rest of the hull of a ship, allegedly used to transport pieces of pottery, was discovered ...)
espectáculo purpose-of *anfiteatro* (*show* purpose-of *amphitheatre*)	*Sobre a hipótese da construção de stands de artesanato e de um* ***anfiteatro para espectáculos****, a edilidade portuense diz ainda não estar nada decidido.* (About the possibility of building crafts booths and an amphitheater for performances, the Porto city council says that nothing is decided yet.)

Table 4.10: Examples of sentences supporting extracted tb-triples.

the remaining third from Wiktionary.PT. Each judge had to classify each tb-triple either as:

- Correct (2): there is at least one context where the tb-triple is valid;
- Wrong relation (1): the arguments of the tb-triple are intuitively related, but their relation (predicate) should be another (e.g. *to_eat* causation-of *spoon*, instead of *to_eat* purpose-of *spoon*);
- Wrong (0): the tb-triple is never valid.

The results of this evaluation are reported in tables 4.11 and 4.12. In the first table, the results are presented according to relation, judge and individually by resource. The second table presents the overall results per relation and the margin

of error for the correct relations, with a confidence level of 95%[14] ($ME_{(2)}$). The same table presents the agreement between the judges, quantified as the number of classification matches (IAA) and as the Kappa value (κ, see Cohen (1960); Carletta (1996)), where the amount of agreement expected by chance is removed, because the probability of each judge giving a certain classification is considered.

Relation	**Judge**	**Resource**								
		DLP			DA			Wikt.PT		
		0	1	2	0	1	2	0	1	2
n synonym-of *n*	J1	0%	0%	100%	1%	0%	99%	1%	0%	99%
	J2	0%	1%	99%	1%	0%	99%	2%	0%	98%
v synonym-of *v*	J1	0%	0%	100%	1%	0%	99%	5%	0%	95%
	J2	1%	1%	98%	01%	0%	99%	5%	2%	93%
n hypernym-of *n*	J1	2%	4%	94%	5%	3%	92%	4%	12%	84%
	J2	3%	6%	91%	9%	7%	84%	4%	8%	88%
n member-of *n*	J1	4%	0%	93%+3%	7%	5%	85%+3%	12%	5%	79%+4%
	J2	4%	3%	76%+17%	12%	3%	42%+43%	18%	14%	54%+14%
v causation-of *n*	J1	4%	1%	95%	1%	3%	96%	7%	10%	83%
	J2	4%	6%	90%	4%	5%	91%	7%	7%	86%
v purpose-of *n*	J1	31%	0%	69%	25%	1%	74%	24%	1%	75%
	J2	29%	2%	69%	20%	1%	79%	25%	0%	75%
adj property-of *v*	J1	23%	5%	72%	18%	11%	71%	26%	5%	69%
	J2	12%	1%	78%	12%	10%	78%	15%	1%	75%

Table 4.11: Results of the manual evaluation of tb-triples according to resource.

Relation	**Judge**	**Total**			**$ME_{(2)}$**	**IAA**	**κ**
		0	1	2			
n synonym-of *n*	J1	2 (1%)	0	298 (99%)	1.1%	0.99	0.66
	J2	3 (1%)	1 (≈0)	296 (99%)	1.1%		
v synonym-of *v*	J1	6 (2%)	0	294 (98%)	1.6%	0.98	0.68
	J2	7 (2%)	3 (1%)	290 (97%)	1.9%		
n hypernym-of *n*	J1	11 (4%)	19 (6%)	270 (90%)	3.4%	0.93	0.64
	J2	16 (5%)	21 (7%)	263 (88%)	3.7%		
n member-of *n*	J1	23 (8%)	10 (3%)	257+10 (86%+3%)	3.5%	0.67/0.88	0.32/0.55
	J2	34 (11%)	20 (7%)	172+74(57%+25%)	4.3%		
v causation-of *n*	J1	12 (4%)	14 (5%)	274 (91%)	3.2%	0.93	0.60
	J2	15 (5%)	18 (6%)	267 (89%)	3.5%		
v purpose-of *n*	J1	80 (27%)	2 (1%)	218 (73%)	4.9%	0.79	0.48
	J2	74 (25%)	3 (1%)	223 (74%)	4.9%		
adj property-of *v*	J1	67 (22%)	21 (7%)	212 (71%)	5.1%	0.81	0.56
	J2	39 (13%)	30 (10%)	231 (77%)	4.7%		

Table 4.12: Results of the manual evaluation of tb-triples.

Manual validation showed that synonymy relations are the most reliable in CARTÃO, as their accuracy is always close to 100%. This was somehow expected, because these relations are also the easiest to extract (see section 4.2.3). The lowest accuracy of synonymy was that of the triples extracted from Wiktionary.PT (93%-95%). This happens mainly due to parsing errors on the *wikitext* indicating the kinds of verb (e.g. transitive, intransitive). This problem has however been corrected for further versions of CARTÃO.

About 90% of the hypernymy triples are correct. Of the incorrect ones, most connect related words. This happens especially in Wiktionary, where definitions, as the following, result in the extraction of hypernyms and not members or synonymys:

[14]The calculation of the margins of error assumed that the evaluated samples were selected randomly, which is not exactly what happened. Although the triples in the samples were selected randomly, there was a constraint to make each sample contain exactly 100 tb-triples from each resource, and the resources have different sizes.

- `rua - os moradores de uma rua` (street – the residents of a street)
  → ~~*morador* hypernym-of *rua*~~ (should be *morador* member-of *rua*)
- `marinha - costa, praia, território próximo ao mar e por ele influenciado, litoral` (seascape – coast, beach, territory near the see and influenced by it, coastline)
  → ~~*costa* hypernym-of *marinha*~~ (should be *costa* synonym-of *marinha*)
  → ~~*praia* hypernym-of *marinha*~~ (should be *praia* synonym-of *marinha*)

Other incorrect hypernymy relations occur because one of the following reasons: (i) some modifiers, important for specifying the meaning of the hypernym, are not included (e.g. *figura* instead of *figura de estilo*); (ii) unconsidered empty heads (e.g. *exemplar*). We recall that the synonymy and hypernymy relations are more than two thirds of CARTÃO.

The accuracy of causation relations is similar to that of the hypernymy relations. This number would be higher if it was not for Wiktionary.PT, where definitions are less standartised with a negative impact on the accuracy of these relations (83-86%). Most of the problems in causation relations are due to underspecified verbs in the first argument, as *fazer* (to do), *realizar* (to perform), or *tornar* (to make).

There is an inherent difficulty on the identification of sub-types of meronymy, which results in different definitions for these relations (Cruse (1986) vs. Winston et al. (1987)), as well as reported difficulties on the identification of textual discriminating patterns specific for each subtype (Ittoo and Bouma, 2010). Therefore, when classifying the member-of triples, we gave the judges a fourth possible classification (3), indicating that there was clearly a meronymy relation between the arguments, but member-of was not the adequate sub-type. As the evaluation results show, most of the member-of triples are meronymy relations but, depending on the resource and on the judge, part of them is not of the member-of subtype. This happens especially for the classifications of the second judge, and more in DA than in the other dictionaries. We should add that the presented margin of error for the member-of triples considers that the triples classified as 3 are also correct (2).

The main problem about the purpose-of relations is that those with a transitive verb on the first argument tend to have that argument incomplete or underspecified. The following illustrate this problem:

- `apeadeiro - lugar, onde o comboio pára algumas vezes, só para deixar ou receber passageiros` (way station – place where the train stops a few times, just to leave or to get passengers)
  → ~~*receber* purpose-of *apeadeiro*~~ (should be *receber_passageiros* purpose-of *apeadeiro*)
- `lenimento - medicamento que serve para diminuir dores` (lenitive – drug for reducing pain)
  → ~~*diminuir* purpose-of *lenimento*~~ (should be *diminuir_dores* purpose-of *lenimento*)
- `anapnógrafo - aparelho que serve para medir a capacidade pulmonar` (anapnograph – device used to measure lung capacity)
  → ~~*medir_capacidade* purpose-of *anapnógrafo*~~ (should be *medir_capacidade_pulmonar* purpose-of *anapnógrafo*)

The same problem occurs, even more consistently, for the property-of triples. However, the property-of relation is the one to which less attention has been given, both in PAPEL, and in our work. It mixes several patterns indicating abstract and not defined relations, such as `relacionado com` (related with), `relativo a` (relative to), `diz-se de` (it is said about). As a consequence, we would like to think of the accuracy of property-of as a lower bound.

Looking at the agreement numbers, we notice that there is good (Green, 1997) or substantial agreement (Landis and Koch, 1977) in the classification of synonymy, hypernymy and causation relations. On the other hand, the relations with less classification agreement (fair and moderate) are also those less semantically well-defined. We have already mentioned the problem of judging member-of triples. We actually present two values for their agreement – the first value considers the four possible classifications, while the second considers member-of as generic meronymy, which means that the triples classified as 3 are considered as if they had been classified as 2. This way, agreement is higher, but still lower than for synonymy, hypernymy and causation. Another source of noise for the member-of relation is that, sometimes, it can be overlapping with the hypernymy. For instance, *bear* is a hyponym of mammal, but is it also a member-of of the class of mammals?

We have also mentioned the abstraction problem of the property-of relation and the underspecification problem that occurs especially for property-of and for purpose-of. Another problem that contributes to less agreement on the classification of purpose-of relations is related with the relaxed semantic constraints of its arguments. This relation may connect very different things. Just to give an idea, it relates an action (verb), which can either be a general purpose (e.g. *to_fry*, *to_desinfect*, *to_calculate*, *to_censor*, *to_dissociate*) or just something one can do with (e.g. *to_punish*, *to_transport*, *to_climb*, *to_spend*, *to_entertain*), for instance, an instrument (e.g. *frying_pan*, *desinfectant*, *whip*), a concrete object (*van*, *stairs*), an abstract means (e.g. *credit*, *calculation*, *satire*), a human entity (e.g. *clown*), or a property (e.g. *dissociation*).

We conclude by referring that these results are, to some extent, comparable, though higher, to those obtained in the creation of MindNet (Richardson et al., 1993), where a sample of 250 relations of 25 different types, extracted from the Longman Dictionary of Contemporary English, were hand-checked for correction. The overall reported accuracy was 78%. It is also referred that the highest accuracy, of about 87%, was obtained for hypernymy, and part-of was the less reliable relation, only 15% accurate.

4.3 Discussion

We have presented the first step towards the automatic creation of a wordnet-like lexical ontology for Portuguese. After explaining how semantic relations are acquired from dictionaries, we described the creation of CARTÃO, the successor of PAPEL and thus the largest term-based lexical-semantic network for Portuguese.

CARTÃO can be browsed using the interface Folheador (Gonçalo Oliveira et al., 2012b; Costa, 2011), designed for facilitating the navigation on Portuguese LKBs represented as tb-triples. Folheador is connected to the online services VARRA (Freitas et al., 2012) and AC/DC (Santos and Bick, 2000; Santos, 2011) that query corpora to provide authentic examples of the relations in context. VARRA is designed not only to search for tb-triples in context, but also to discover new discriminating patterns for each relation, and to identify good and bad examples of each tb-triple. The examples might be useful to understand and to evaluate the triple. An exercise using VARRA to validate part of PAPEL 2.0 is described in Freitas et al. (2012).

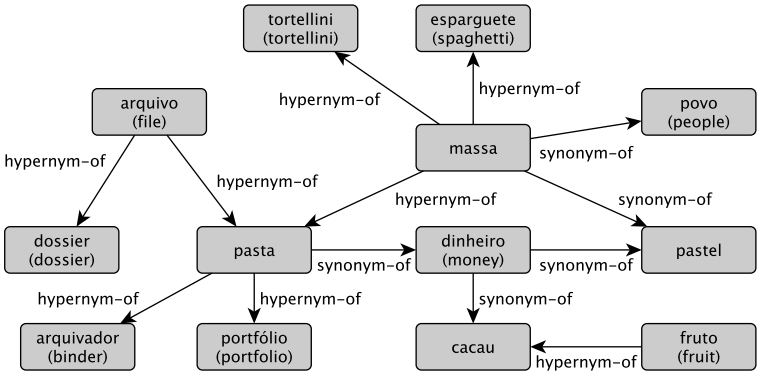


Figure 4.5: Lexical network where ambiguity arises.

Given that they share the same structure, the utility of a resource as CARTÃO is supported by the number of works using PAPEL. So far, PAPEL has been used as a gold standard for computing similarity between lexical items (Sarmento, 2010), in the adaptation of textual contents for poor literacy readers (Amancio et al., 2010), in the automatic generation of distractors for cloze questions (Correia et al., 2010), as a knowledge base for QA (Saias, 2010; Rodrigues et al., 2011) and question generation (Marques, 2011) systems, to validate terms describing places (Oliveira Santos et al., 2012), and in the enrichment (Silva et al., 2012b) and creation (Paulo-Santos et al., 2012) of sentiment lexicons. CARTÃO has already been used in the automatic generation of poetry (Gonçalo Oliveira, 2012).

On the other hand, lexical resources based on words, identified by their orthographical form, are not practical for several computational applications. This happens because words have different senses that go from tightly related, as in polysemy (e.g. bank, institution and building) or metonymy (e.g. bank, the building or its employers), to completely different, as in homonymy (e.g. bank, institution or slope). Moreover, there are words with completely different orthographical forms denoting the same concept (e.g. car and automobile).

Ambiguities may lead to serious inconsistencies in tasks where handling word senses is critical, as in inference. In figure 4.5, we present an example of a term-based lexical network with several ambiguous Portuguese words, namely:

- *pasta*, which might refer to a briefcase, paste, pasta or money (figuratively);
- *massa*, which might refer to pasta, to people or money (both figuratively);
- *pastel* might be a cake or money (figuratively);
- *cacau* might refer to cocoa (a fruit) or to money (also figuratively).

It is not hard to imagine that, if these ambiguities are not handled, erroneous inferences can be made, such as:

- *massa* synonym-of *povo* $\wedge$ *massa* hypernym-of *tortellini*
  $\rightarrow$ *povo* hypernym-of *tortellini* (people hypernym-of tortellini)
- *dinheiro* synonym-of *cacau* $\wedge$ *fruto* hypernym-of *cacau*
  $\rightarrow$ *fruto* hypernym-of *dinheiro* (fruit hypernym-of money)

A real example of these problems is presented in Gonçalo Oliveira et al. (2010b),

where transitivity was applied to the synonymy relations of PAPEL, giving rise to some inconsistencies as the following:

- *queda* synonym-of *ruína* ∧ *queda* synonym-of *habilidade*
  → *ruína* synonym-of *habilidade*

The problem occurs because one sense of *queda* is the result of falling, while another means to have some skill. Therefore, combining those two, we obtain that *ruína* (ruin) is the same as *habilidade* (ability, skill), which are almost opposites.

Nevertheless, since the beginning of the project PAPEL, our option was to build a lexical resource where lexical items were not divided into word senses. That early option relied on the following:

- From a linguistic point of view, word senses are not discrete and cannot be separated with clear boundaries (Kilgarriff, 1996; Hirst, 2004). Sense division in dictionaries and lexical ontologies is most of the times artificial.
- Following the previous point, the sense granularity in dictionaries and lexical ontologies is often different from lexicographer to lexicographer. As there is not a well-defined criteria for the division of meanings, word senses in different resources do not always match (Dolan, 1994; Peters et al., 1998).
- Word sense disambiguation (WSD, see Navigli (2009b) for a survey) is the task of, given the context where a word occurs, selecting the most adequate of its senses from a sense inventory. However, the previous points confirm that WSD is an ill-defined task and is very dependent on the purpose (Wilks, 2000).
- Dictionaries do not provide the sense corresponding to a word occurring in a definition. After the first version of PAPEL was released, Navigli (2009a) actually presented a method for disambiguating words in dictionary definitions. Still, given the aforementioned problems on WSD, the term-based structure of PAPEL was kept.
- Finally, in natural language, the study of vagueness is as, or even more, important that studying ambiguity (see e.g. Santos (1997)).

When we started to extract relations from other dictionaries (and thesauri), we confirmed that the senses of words occurring in more than one resource did not match for different resources. Moreover, not all definitions in Wiktionary.PT have a sense number and synonymy lists do not always indicate the corresponding synonymous sense. Since we are extracting information from more than one lexical resource, an alternative would be to align the word senses in different resources (represented as definitions in dictionaries or synsets in thesauri), as others did (e.g. Vossen et al. (2008); Henrich et al. (2012)). Still, given the aforementioned utility of a lexical resource as PAPEL, we decided to keep CARTÃO as a term-based resource.

In the following chapters, we explain how the structure of CARTÃO can evolve to a resource that handles word senses. After the additional steps of the ECO approach, the result is Onto.PT, a resource structured in synsets. We recall that this approach is flexible in a way that it enables the construction (and further augmentation) of a wordnet, based on the integration of knowledge from multiple heterogeneous sources and, from this point, it does not require an additional analysis of the extraction context. The only requirement is that the initial information is represented as tb-triples, which is kind of a standard representation.

Chapter 5

Synset Discovery

As referred in the previous chapter, a LKB structured in words, instead of concepts, does not handle lexical ambiguity and might lead to serious inconsistencies. To deal with that issue, wordnets are structured in synsets, which are groups of words sharing a common meaning and thus representing a concept. This chapter is about the discovery of synsets from a term-based LKB, which is the first step for moving towards a sense-aware resource.

Since a synset groups words according to their synonymy, in this step, we only use the network established by the synonymy triples extracted from dictionaries. On the one hand, co-occurrence graphs extracted from corpora have shown to be useful for identifying not only synonymous words, but also word senses (Dorow, 2006). It should be mentioned that, in opposition to other kinds of relation, synonymous words share similar neighbourhoods, but may not co-occur frequently in corpora text (Dorow, 2006), which leads to few textual patterns connecting this kind of words. So, as referred in section 3.2.2, 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)). On the other hand, in synonymy networks extracted from dictionaries, clusters tend to express concepts (Gfeller et al., 2005) and can therefore be exploited for the establishement of synsets. Methods for improving the organisation of synonymy graphs, extracted from different resources, are presented by Navarro et al. (2009).

As other authors noticed for PAPEL (Prestes et al., 2011), we confirmed that synonymy networks extracted from dictionaries connect more than half of the words by, at least, one path. Therefore, as others did for discovering new concepts from text (e.g. Lin and Pantel (2002)), we used a (graph) clustering algorithm on our synonymy networks. This kind of work is related to WSD. More specifically, it can be seen as word sense induction (WSI, Navigli (2012)) as it discovers possible concepts of a word, without exploiting an existing sense inventory.

As discussed in section 4.3, from a linguistic point of view, word senses are not discrete, so their representation as crisp objects does not reflect the human language. A more realistic approach for coping with this fact is to represent synsets as models of uncertainty, such as fuzzy sets, to handle word senses and natural language concepts. Our clustering algorithm can be used for the discovery of fuzzy synsets. 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.

There is work on fuzzy concept discovery, as Velldal (2005), who describes a similar work to Lin and Pantel (2002), but 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 term-based synonymy pairs. The fuzzy membership values are based on human judgements of the synonymy pairs.

This chapter starts by defining synonymy networks, which are the target of clustering, and then describes the clustering algorithm we have used. Before a final discussion, we present our work towards the automatic creation and the evaluation of a simple thesaurus and a thesaurus with fuzzy memberships for Portuguese. This work was originally reported in Gonçalo Oliveira and Gomes (2011a).

5.1 Synonymy networks

Synonymy networks are a particular kind of term-based lexical networks. Formally, they are graph structures $N = (V, E)$, with $|V|$ nodes and $|E|$ edges, $E \subset V^2$. Each node $v_a \in V$ represents a word and each edge connecting v_a and v_b, $E(v_a, v_b)$, indicates that, in some context, words a and b have the same meaning and are thus synonymous. In other words, it indicates that $t = \{a \text{ synonym-of } b\}$ holds.

Synonymy networks may be established, for instance, by term-based synonymy triples extracted from dictionaries (e.g. *bravo* synonym-of *corajoso*). Each of those constitute a synonymy pair (hereafter, synpair), $p = \{a, b\}$, which describes an edge of the synonymy network.

When extracted from dictionaries, these networks tend to have a clustered structure (Gfeller et al., 2005; Navarro et al., 2009). Therefore, we exploit them in order to identify clusters, which may be used as the synsets of a thesaurus/wordnet.

We represent each node $v_a \in V$ as a (adjacency) vector $\vec{v}_a$, where each dimension is a word in the network. If nodes v_i and v_j are connected, their connection is weighted, so $w_{ij} > 0$. Otherwise, $w_{ij} = 0$. The network may therefore be seen as an adjacency matrix M with $|N|$ columns and $|N|$ rows, where each column i is the vector of the word in v_i:

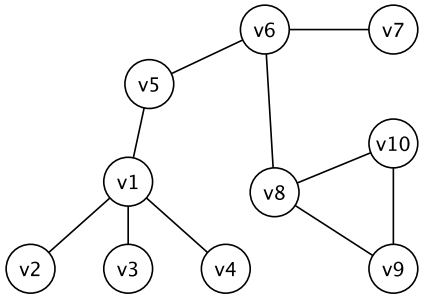
$$M_{ij} = \begin{cases} w_{ij} \in \mathbb{N} & \text{, if } E(v_i, v_j) \text{ exists} \\ 0 & \text{, otherwise} \end{cases}$$

Figure 5.1 shows a synonymy network and its adjacency matrix, considering that all edges weight 1.

More than including words that are connected in the network, a cluster should include very similar words. Given that similar words have similar neighbourhoods, in the following sections, the similarity between two words, a and b, is given by the similarity between their adjacency vectors, $sim(a, b) = sim(\vec{v}_a, \vec{v}_b)$.

5.2 The (fuzzy) clustering algorithm

In order to identify clusters in the synonymy network N, we apply an algorithm for graph clustering (Schaeffer, 2007). Furthermore, we take advantage of the different



$\vec{v}_1$	$\vec{v}_2$	$\vec{v}_3$	$\vec{v}_4$	$\vec{v}_5$	$\vec{v}_6$	$\vec{v}_7$	$\vec{v}_8$	$\vec{v}_9$	$\vec{v}_{10}$
1	1	1	1	1	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
1	0	1	0	0	0	0	0	0	0
1	0	0	1	0	0	0	0	0	0
1	0	0	0	1	1	0	0	0	0
0	0	0	0	1	1	1	1	0	0
0	0	0	0	0	1	1	0	0	0
0	0	0	0	0	1	0	1	1	1
0	0	0	0	0	0	0	1	1	1
0	0	0	0	0	0	0	1	1	1

Figure 5.1: A graph and its corresponding representation as an adjacency matrix.

computed similarities between the nodes for constituting fuzzy clusters, which are sets whose elements have degrees of membership μ. Fuzzy clusters lead to fuzzy synsets, where words have a membership degree. We can say that the fuzzy synset representation is between the "*there are no word senses*" view and the discrete sense division in dictionaries and wordnets.

While most graph clustering algorithms, such as fuzzy c-means (Bezdek, 1981) or Markov Clustering (MCL, van Dongen (2000)), would suit this purpose, we decided to use a simpler algorithm, because we wanted to cluster based only on the similarity of the adjacencies. In opposition to fuzzy c-means, in our algorithm, there is no need to keep two matrixes, one with the memberships and another with the centroids, which is important because synonymy graphs can be very large and memory is sometimes not enough. Moreover, there is no need to specify the number of clusters – words are organised into m clusters, where m is never higher than the number of unique words, $|N|$. As for MCL, we have made some experiments (Gonçalo Oliveira and Gomes, 2010a). However, the higher complexity of the algorithm did not result in clearly better results. For instance, it often discovered clusters with more than 25 words, usually impractical. Not to speak about the longer processing time.

5.2.1 Basic algorithm

The basic idea of our algorithm is that each node and its neighbourhood define a potential cluster. Also, clusters might be overlapping, as one word might have more than one sense and thus be included in more than one synset. Clusters with fuzzy membership are identified after running the following procedure on N:

1. Create an empty sparse matrix C, $|N| \times |N|$, which is the clustering similarity matrix.
2. Fill each cell C_{ij} with the similarity between words in $v_i \in V$ and $v_j \in V$, represented as vectors $\vec{v_i}$ and $\vec{v_j}$.
3. Normalise the rows of C, so that the values in each column, C_j, sum up to 1.
4. Extract a fuzzy cluster F_i from each column C_i, consisting of the words in nodes v_j where $sim(\vec{v_i}, \vec{v_j}) > 0$ and thus $C_{ij} > 0$. The value in C_{ij} is used as

the membership degree of the word in v_j to F_i, $\mu_{F_i}(v_j)$.

5. For each cluster F_i with all elements included in a larger 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, $\mu_{F_k}(v_j) = \mu_{F_i}(v_j) + \mu_{F_j}(v_j)$.

Figure 5.2 is the normalised clustering matrix C for the network in figure 5.1, where we present the resulting fuzzy clusters as well. Similarities are computed with the cosine similarity measure, as follows:

$$sim(a,b) = cos(\vec{v_a}, \vec{v_b}) = \frac{\vec{v_a}.\vec{v_b}}{|\vec{v_a}||\vec{v_b}|} = \frac{\sum_{i=0}^{|V|} v_{ai} \times v_b i}{\sqrt{\sum_{i=0}^{|V|} v_{ai}^2 \times \sum_{i=0}^{|V|} v_{bi}^2}} \tag{5.1}$$

$\vec{v}_1$	$\vec{v}_2$	$\vec{v}_3$	$\vec{v}_4$	$\vec{v}_5$	$\vec{v}_6$	$\vec{v}_7$	$\vec{v}_8$	$\vec{v}_9$	$\vec{v}_{10}$
0.27	0.17	0.17	0.17	0.14	0.06	0.00	0.00	0.00	0.00
0.21	0.33	0.16	0.16	0.13	0.00	0.00	0.00	0.00	0.00
0.21	0.16	0.33	0.16	0.13	0.00	0.00	0.00	0.00	0.00
0.21	0.16	0.16	0.33	0.13	0.00	0.00	0.00	0.00	0.00
0.13	0.10	0.10	0.10	0.25	0.14	0.10	0.07	0.00	0.00
0.07	0.00	0.00	0.00	0.11	0.32	0.23	0.08	0.09	0.09
0.00	0.00	0.00	0.00	0.17	0.29	0.41	0.14	0.00	0.00
0.00	0.00	0.00	0.00	0.08	0.07	0.10	0.28	0.24	0.24
0.00	0.00	0.00	0.00	0.00	0.09	0.00	0.27	0.32	0.32
0.00	0.00	0.00	0.00	0.00	0.09	0.00	0.27	0.32	0.32

- $F_{1,2,3,4,5} = \{v_1(0.94), v_2(1), v_3(1), v_4(1), v_5(0.68), v_6(0.19), v_7(0.17), v_8(0.08)\}$
- $F_{1,5,6,7,8,9,10} = \{v_1(0.06), v_5(0.32), v_6(0.81), v_7(0.83), v_8(0.92), v_9(1), v_{10}(1)\}$

Figure 5.2: Clustering matrix C after normalisation and resulting fuzzy sets.

If $\mu_{F_i}(v_a) > 0$, the word v_a has a sense with a common meaning to the other words in F_i. The membership degree $\mu_{F_i}(v_a)$ may be seen as the confidence on the usage of the word v_a with the meaning of the synset F_i.

Also, step 3 of the algorithm is optional. In a normalised C, all membership degrees of the same word sum up to 1, $\sum \mu_{F_i}(v_j) = 1$. Therefore, membership degrees of a can also be interpreted as the possible senses of the word a and the likelihood of the word a conveying their meanings. However, normalising C will make highly connected words to have low memberships.

In order to obtain simple synsets from fuzzy synsets F_i, one has just to apply a threshold θ to the membership degrees, so that all words a with membership lower than θ, $\mu_{F_i}(v_a) > \theta$, are excluded from the synset. In this case, attention should be paid when C is normalised, as using the same θ for all fuzzy synsets might prevent that highly connected words are included in any synset.

Finally, any measure for computing the similarity of two vectors, as their cosine, can be used in step 2 of the algorithm. If M is a binary matrix, as the one in figure 5.1, measures typically used for computing the similarity between sets, such as the Jaccard coefficient, are a suitable alternative.

5.2.2 Redundancy and weighted edges

It might be useful to take advantage of redundancy and weight the connections according to the number of times a synpair is extracted. In this case, M will not be a binary matrix. Each edge of the graph becomes a triplet $E(v_a, v_b, w_{ab})$, where a synpair $\{a, b\}$ has an associated weight w_{ab}, relative to the number of times it was extracted. Even though it is not common to extract the same synpair more than once from the same dictionary, it is if more than one dictionary is used. Furthermore, 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.

If a synpair might be extracted more than once, the problem of discovering synsets becomes similar to the problem of discovering concepts from corpora text, as described by Lin and Pantel (2002). Inspired by their work, instead of using the plain similarity value between vectors, we compute the association of the connected words using the pointwise mutual information (pmi). For this purpose, each dimension of the vectors $\vec{v_i} \in C$ will have the pmi between the word in v_i and each other word, computed using expression 5.2. However, as the pmi is biased towards infrequent words, it should be multiplied by the discounting factor in expression 5.3, also suggested by Lin and Pantel (2002). The similarity of two words is finally given, for instance, by the cosine between their vectors (expression 5.4).

$$pmi(a,b) = \frac{\frac{M_{ab}}{S}}{\frac{\sum_{j=0}^{|V|} M_{aj}}{S} \times \frac{\sum_{i=0}^{|V|} M_{ib}}{S}}, S = \sum_{i=0}^{|V|}\sum_{j=0}^{|V|} M_{ij} \tag{5.2}$$

$$df(a,b) = \frac{M_{ab}}{M_{ab}+1} \times \frac{min\left(\sum_{j=0}^{|V|} M_{aj}, \sum_{i=0}^{|V|} M_{ib}\right)}{min\left(\sum_{j=0}^{|V|} M_{aj}, \sum_{i=0}^{|V|} M_{ib}\right)+1} \tag{5.3}$$

$$sim(a,b) = cos(\vec{v_a}, \vec{v_b}) = \frac{\vec{v_a}.\vec{v_b}}{|\vec{v_a}||\vec{v_b}|} = \frac{\sum_{i=0}^{|V|} pmi(a,v_i) \times pmi(b,v_i)}{\sqrt{\sum_{i=0}^{|V|} pmi(a,v_i)^2 \times \sum_{i=0}^{|V|} pmi(b,v_i)^2}} \tag{5.4}$$

5.3 A Portuguese thesaurus from dictionaries

The clustering procedure described in the previous section was used to create CLIP (formerly known as Padawik), a Portuguese thesaurus with information extracted from three dictionaries. For the creation of CLIP, we used the synonymy

tb-triples (synpairs) of CARTÃO (introduced in section 4), which establish a synonymy network. However, this work was done before the last version of CARTÃO was available. The presented results were obtained with relations extracted using the grammars of PAPEL 2.0 in DLP, DA (in a previous modernisation stage) and the 25th October 2010 Wiktionary.PT dump.

5.3.1 Synonymy network data

Before running the clustering procedure, we examined some properties of the synonymy network established by synpairs collected from the three dictionaries. Table 5.1 shows the following properties, typically used to analyse graphs:

- Number of nodes $|V|$, which corresponds to the number of unique lexical items in the synpair arguments.
- Number of edges $|E|$, which corresponds to the number of unique synpairs.
- Average degree ($\overline{deg}(N)$) of the network (see expression 5.5), which is the average number of edges per node.
- Number of nodes of the largest connected sub-network $|V_{lcs}|$, which is the largest group of nodes connected directly or indirectly in N.
- Average clustering coefficient $\overline{CC_{lcs}}$ of the largest connected sub-network, which measures the degree to which nodes tend to cluster together as a value in [0-1] (see expression 5.7). In random graphs, this coefficient is close to 0. The local clustering coefficient $CC(v_i)$ (see expression 5.8) of a node v_i quantifies how connected its neighbours are.

$$\overline{deg}(N) = \frac{1}{|V|} \times \sum_{i=1}^{|V|} deg(v_i) : v_i \in V \quad (5.5) \qquad deg(v_i) = |E(v_i, v_k)| : v_k \in V \quad (5.6)$$

$$\overline{CC} = \frac{1}{|V|} \times \sum_{i=1}^{|V|} CC(v_i) \quad (5.7)$$

$$CC(v_i) = \frac{2 \times |E(v_j, v_k)|}{K_i \times (K_i - 1)} : v_j, v_k \in neighbours(v_i) \wedge K_i = |neighbours(v_i)| \quad (5.8)$$

Weights were not considered in the construction of table 5.1. Since we have extracted four different types of synonymy, considering the POS of the connected items (nouns, verbs, adjectives and adverbs, see more details in section 4.2.4), in the same table, we present the properties of the four synonymy networks independently.

POS	$\|V\|$	$\|E\|$	$\overline{deg}(N)$	$\|V_{lcs}\|$	$\overline{CC_{lcs}}$
Nouns	39,355	57,813	2.94	25,828	0.14
Verbs	11,502	28,282	4.92	10,631	0.17
Adjectives	15,260	27,040	3.54	11,006	0.16
Adverbs	2,028	2,206	2.52	1,437	0.10

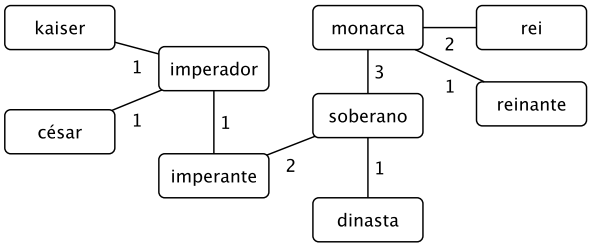
Table 5.1: Properties of the synonymy networks.

Table 5.1 shows that the nouns network is the largest, the adverbs is the smaller, but the verbs and adjectives networks are more connected. Still, all the networks are quite sparse, which stands out by the low ratio between the number of edges and nodes. It is thus possible to represent them as sparse matrixes and minimise memory consumption. Clustering coefficients are slightly lower than those of graphs extracted from Wiktionaries, between 0.2 and 0.28 (Navarro et al., 2009), but still higher than 0, which confirms that our networks have a clustered structured.

An interesting fact is that, for all POS, the networks contain a core subgraph, the largest connected sub-network *lcs*, which contains always more than half of the total nodes. If there was no ambiguity, this would mean that all the words in *lcs* were synonymys of each other, which is, of course, not true. For PAPEL, this fact is also noticed by Prestes et al. (2011). This shows that we cannot apply blind transitivity to the synonymy represented by the synpairs and points out the need of additional organisation of synonymy automatically extracted from dictionaries.

5.3.2 Clustering examples

Before reporting on the quantitative results for the whole synonymy network, we discuss real examples of the synsets obtained using the cosine similarity and *pmi*. Figure 5.3 presents one weighted subgraph and the fuzzy synsets obtained after clustering. The subgraph connects words denoting a person who rules, and divides them in two slightly different concepts. A caeser/emperor, which is someone who rules an empire, and a king, which rules a kingdom. Several words can denote both concepts, but with different membership degrees.



$synset_1$	μ	$synset_2$	μ
kaiser (*kaiser*)	1.0	reinante (*regnant*)	1.0
césar (*caeser*)	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 5.3: Weighted synonymy network and resulting fuzzy synsets.

Following, to have and idea on how ambiguity was handled in the establishment of fuzzy synsets, we selected several polysemic Portuguese nouns, looked at some of the synsets containing them, and divided the synsets manually into the possible senses of the words. Table 5.2 shows the result of this exercise for the words: *pasta*,

cota and *planta*. Besides an italian dish (not included), the word *pasta* might have the popular meaning of money, the figurative meaning of a mixture of things, or it might denote a file or a briefcase. As for the word *cota*, besides height (not included), it can be a quota or portion, or refer to an old and respectable person, and informally denote a father or a mother. The word *planta* might either denote a plan or some guidelines, or it might denote a vegetable. Besides some synonyms of plant/vegetable (e.g. *planta*, *vegetal*), the synset with the vegetable meaning contains many actual plants or vegetables (e.g. *maruge*, *camélia*). After analysing this problem, we noticed that the dictionary DA contains several definitions of plants where the first sentence is just *planta*, without any *differentia*. Therefore, even though the correct relation to extract would be hypernymy, our grammars see those definitions as denoting synonymy. Another limitation shown by these examples is that, sometimes, the fuzzy synsets contain words which are not synonyms, but have similar neighbourhoods.

5.3.3 Thesaurus data for different cut points

After analysing the fuzzy synsets, we inspected the impact of applying different cut-points (θ) in the transformation of the fuzzy thesaurus into a simple thesaurus. Tables 5.3 and 5.4 present the properties of the different thesauri obtained with different values for θ. Considering just the words of the thesauri, table 5.3 includes the number of words, how many of those are ambiguous, the average number of senses per word, and the number of senses of the most ambiguous word. As for synsets, table 5.4 shows the total 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), as well as the largest synset. Both tables do not consider synsets of size 1.

Before collecting the data in Tables 5.3 and 5.4, we followed one of the clustering methods for word senses proposed for EuroWordNet, which suggests that synsets with three members in common can be merged (Peters et al., 1998). However, the design of our clustering algorithm and the configuration of our synonymy networks 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, computed as follows, where $|Synset|$ denotes the number of words of a synset:

$$Overlap(Synset_a, Synset_b) = \frac{Synset_a \cap Synset_b}{min(|Synset_a|, |Synset_b|)} \tag{5.9}$$

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. For instance, with $\theta = 0.5$, despite the establishment of 8,000 clusters, each word has only one sense, which means there is no ambiguity. Out of curiosity, the largest synset in CLIP, with $\theta = 0.075$, denotes the concept of *money*. It contains the following 58 words:

- *jimbo, pastel, bagarote, guines, baguines, parrolo, marcaureles, ouro, grana, arame, massaroca, tutu, metal, bagalho, níquel, bilhestres, milho, jan-da-cruz, china, cumquibus, mussuruco, cobre, numerário, pilim, bagaço, pasta, zerzulho, painço, finanças, chelpa, calique, posses, bagalhoça, pecuniária, boro, dieiro, pila, gaita,*

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), ...
		poupanças(0.5202), *pé-de-meia*(0.289), *coscorrinho*(0.0799), *maquia*(0.0698), *moenda*(0.0648), ... , *economia*(0.0394), *rendor*(0.0285), *rédito*(0.0236), ... ,*ganhança*(0.0182), *lucro*(0.015), *gimbo*(0.0135), *chelpa*(0.0135), *pecúnia*(0.0135), *patacaria*(0.0135), *provento*(0.0134), *arame*(0.0133), *zerzulho*(0.0133), *metal*(0.0133), *carcanhol*(0.0133), *pecunia*(0.0133), *cunques*(0.0133), *pecuniária*(0.0133), *bagarote*(0.0133), *matambira*(0.0133), *jan-da-cruz*(0.0133), *bagalho*(0.0133), *cacau*(0.0133), *boro*(0.0133), *calique*(0.0133), *marcaureles*(0.0133), *teca*(0.0133), *níquel*(0.0133), *mussuruco*(0.0133), *massaroca*(0.0133), *baguines*(0.0133), *bilhestres*(0.0133), *parrolo*(0.0133), *pastel*(0.0133), *cum-quibus*(0.0133), *dieiro*(0.0133), *pilim*(0.0133), *pataco*(0.0128), *bagalhoça*(0.0125), *bago*(0.0125), *china*(0.0125), *cobre*(0.0125), *numo*(0.0125), *gage*(0.0123), *maco*(0.0121), *jimbo*(0.012), ... , *guines*(0.0119), *pasta*(0.0114), ...
	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)
	file	*diretório*(1.0), *dossier*(0.9176), *pasta*(0.1118), ...
	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-família*(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), *papai*(0.6183), *papá*(0.6169), *tatá*(0.4934), *pai*(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),...
planta	guidelines	*prospectiva*(0.5166), *prospecto*(0.0805), *prospeto*(0.0663), *calendarização*(0.0595), *prisma*(0.055), *óptica*(0.055), *programa*(0.0452), *planos*(0.0354), *intuitos*(0.0354), *traçado*(0.034), *traçamento*(0.0295), *olhar*(0.0284), *perspectiva*(0.0271), *gizamento*(0.0261), *alçado*(0.0258), *horizonte*(0.0228), *planificação*(0.0228), *visualidade*(0.0227), *gázeo*(0.0227), *panorama*(0.0213), *calendário*(0.0206), *aspeito*(0.0176), ... , *conspecção*(0.017), *programação*(0.0168), *desenho*(0.0164), *terrapleno*(0.0157), *diagrama*(0.0152), *fácies*(0.0149), *ângulo*(0.0145), *estampa*(0.0141), *esquema*(0.0141), *contenença*(0.0134), *duaire*(0.0133), *duairo*(0.0133), *arquitectura*(0.0128), *probabilidade*(0.0127), *vista*(0.0126), *viseira*(0.0124), *design*(0.0116), *faceta*(0.0113), *janela*(0.0112), *alinhamento*(0.0112), *abordagem*(0.011), *desígnio*(0.0107), *planta*(0.0107), *painel*(0.0105), *projecto*(0.0104)
		mapam(0.4176), *gráfico*(0.4176), *organigrama*(0.2037), *mapa*(0.0512), *tábua*(0.0417), *carta*(0.0359), *catálogo*(0.0339), *planta*(0.0227), *procedência*(0.0225)
		gizamento(0.1106), *planificação*(0.0483), *programação*(0.0357), *terrapleno*(0.0332), *diagrama*(0.0322), *traçamento*(0.0313), *esquema*(0.0298), *arquitectura*(0.0271), *design*(0.0247), *programa*(0.0187), *desenho*(0.0174), *rascunho*(0.0169), *idéia*(0.0161), *plana*(0.0142), *traçado*(0.0141), *prospecto*(0.0129), *planta*(0.0114)
	vegetable	*plantas*(0.65226775), *marrugem*(0.53500473), *caruru-guaçu*(0.4826975), *planta*(0.325316), *caruru*(0.21026045), *marugem*(0.19776152), *vinagreira*(0.15554681), *vegetal*(0.14959434), *murugem*(0.0829055), *bananeirinha-do-brejo*(0.06475778), *pranta*(0.038034387), *camélia*(0.030711966), *alçado*(0.02910556), *traçado*(0.027653778), *cordão-de-san-francisco*(0.024771051), *maruge*(0.024124833), *rosa-do-japão*(0.023929605), *japoneira*(0.023929605), *cordão*(0.021231819), *melancia*(0.015556527), *presentação*(0.011791914), *condicionamento*(0.011791912), *erva-ferro*(0.011321816)

Table 5.2: Fuzzy synsets of polysemic words.

θ	Words			
	Total	**Ambiguous**	**Avg(senses)**	**Most ambiguous**
0.025	39,350	21,730	3.18	18
0.05	39,288	17,585	1.86	9
0.075	38,899	12,505	1.44	7
0.1	38,129	8,447	1.26	6
0.15	35,772	4,198	1.12	4
0.25	30,266	1,343	1.04	3
0.5	22,203	0	1.0	1

Table 5.3: Noun thesauri words data, using different cut points θ.

θ	Synsets				
	Total	**Avg(size)**	$size = 2$	$size > 25$	$max(size)$
0.025	13,344	9.39	3,921	576	80
0.05	12,416	5.89	4,224	119	62
0.075	12,086	4.64	4,878	47	58
0.1	11,748	4.10	5,201	34	58
0.15	11,044	3.64	5,248	16	58
0.25	9,830	3.22	5,095	10	58
0.5	8,004	2.77	5,011	3	47

Table 5.4: Noun thesauri synsets data, using different cut points θ.

pataco, cacau, matambira, gimbo, cunques, caroço, fanfa, maco, pecúnia, jibungo, maquia, dinheiro, bago, numo, teca, pecunia, quantia, guita, patacaria, carcanhol

Every word in this synset may be used to denote money, which indicates there are many ways of referring to this concept in Portuguese, namely: 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 CLIP refers to alcoholic intoxication:

- *torcida, zurca, zuca, trapisonda, torta, piela, perua, raposada, bicancra, gateira, carraspana, samatra, gata, pizorga, tachada, caroça, pifão, ardina, carrega, rasca, zerenamora, touca, zola, parrascana, gardunho, gardinhola, tropecina, ema, cardina, tiorga, pisorga, berzundela, dosa, chiba, bebedeira, perunca, canjica, raposa, taçada, raposeira, cartola, ganso, tortelia, turca, cabrita, borracheira, piteira, pifo, bebedice, marta, zangurriana, bezana, vinhaça, bêbeda*

5.3.4 Comparison with handcrafted Portuguese thesauri

As referred in the previous chapters, there are two public handcrafted thesauri for Portuguese, TeP and OpenThesaurus.PT (OT.PT). Therefore, in order to evaluate CLIP, our first option was to compare this thesaurus, created automatically, with the handcrafted alternatives. Besides the "fuzzy" CLIP ($\theta = 0.01$), we also included a version of this thesaurus, obtained with $\theta = 0.075$. This value 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. Finally, we validated the clustering algorithm, which was used to reconstruct the handcrafted thesauri from its synpairs.

Thesauri properties

This comparison is focused on the size and ambiguity numbers of the thesauri. For nouns, verbs, and adjectives, tables 5.5 and 5.6 put CLIP side-by-side to the handcrafted thesauri. Table 5.5 is relative to words, while table 5.6 is relative to synsets. The presented properties are the same as in tables 5.3 and 5.4.

It is clear that both automatically created thesauri are larger than the handcrafted ones. For nouns, the former have two times more words than TeP. Furthermore, as expected, the fuzzy thesaurus is more ambiguous, both because its words have, on average, more senses, and because its synset sizes are much higher.

Thesaurus	**POS**	**Words**			
		Quantity	**Ambiguous**	**Avg(senses)**	**Most ambiguous**
OT.PT	Nouns	6,110	485	1.09	4
	Verbs	2,856	337	1.13	5
	Adjectives	3,747	311	1.09	4
TeP 2.0	Nouns	17,158	5,805	1.71	20
	Verbs	10,827	4,905	2.08	41
	Adjectives	14,586	3,735	1.46	19
CLIP-0.01	Nouns	39,354	24,343	7.78	46
	Verbs	11,502	10,411	14.31	42
	Adjectives	15,260	10,636	10.36	43
CLIP-0.075	Nouns	38,899	12,505	1.44	7
	Verbs	11,070	5,717	1.76	7
	Adjectives	14,964	6,644	1.69	6

Table 5.5: Thesaurus words comparison.

Thesaurus	**POS**	**Synsets**				
		Quantity	**Avg(size)**	$size = 2$	$size > 25$	$max(size)$
OT.PT	Nouns	1,969	3.38	778	0	14
	Verbs	831	3.90	226	0	15
	Adjectives	1,078	3.80	335	0	17
TeP 2.0	Nouns	8,254	3.56	3,079	0	21
	Verbs	3,978	5.67	939	48	53
	Adjectives	6,066	3.50	3,033	19	43
CLIP-0.01	Nouns	20,102	15.23	3,885	3,756	109
	Verbs	7,775	21.17	307	2,411	89
	Adjectives	8,896	17.77	1,326	2,157	109
CLIP-0.075	Nouns	12,086	4.64	4,878	47	58
	Verbs	4,198	4.63	1,189	14	49
	Adjectives	5,666	4.45	1,980	11	46

Table 5.6: Thesaurus synsets comparison.

Thesauri overlaps

The second comparison inspected how common the contents of the thesauri are. We measure the word overlap and the synset overlap of a thesaurus $Thes$ with a reference thesaurus Ref using the following expressions, where $|Thes|$ is the number of synsets of $Thes$, S_{Ti} and S_{Rj} are synsets in $Thes$ and Ref respectively, and $Overlap$ measures the overlap between two synsets, computed according to expression 5.9:

$$WordOverlap(Thes, Ref) = \frac{commonWords(Thes, Ref)}{|Thes|} \tag{5.10}$$

$$SynsetOverlap(Thes, Ref) = \frac{\sum_{i=1,j=1}^{|Thes|,|Ref|} max(Overlap(S_{Ti}, S_{Rj}))}{|Thes|} \tag{5.11}$$

Tables 5.7, 5.8 and 5.9 present the measured overlaps for nouns, verbs and adjectives, respectively. These tables show that the overlaps between the words of the automatically generated and the handcrafted thesauri are low, especially for nouns, and for the handcrafted thesaurus OT, which is smaller. Furthermore, as a consequence of the low word overlap, even though higher, the overlap between automatically generated and handcrafted synsets is below 0.57 (for nouns), 0.67 (for verbs), and 0.70 (for adjectives).

	Reference							
	TeP		**OT.PT**		**CLIP-0.01**		**CLIP-0.075**	
Thesaurus	**Words**	**Synsets**	**Words**	**Synsets**	**Words**	**Synsets**	**Words**	**Synsets**
TeP	1.0	1.0	0.28	0.35	0.74	0.66	0.73	0.57
OT.PT	0.79	0.68	1.0	1.0	0.92	0.82	0.90	0.67
CLIP-fuzzy	0.33	0.57	0.14	0.45	1.0	1.0	0.99	0.97
CLIP-0.075	0.32	0.35	0.14	0.19	1.0	1.0	1.0	1.0

Table 5.7: Noun thesauri overlaps.

	Reference							
	TeP		**OT.PT**		**CLIP-0.01**		**CLIP-0.075**	
Thesaurus	**Words**	**Synsets**	**Words**	**Synsets**	**Words**	**Synsets**	**Words**	**Synsets**
TeP	1.0	1.0	0.25	0.36	0.70	0.58	0.66	0.54
OT.PT	0.95	0.80	1.0	1.0	0.97	0.83	0.89	0.62
CLIP-fuzzy	0.66	0.67	0.24	0.59	1.0	1.0	0.96	0.97
CLIP-0.075	0.65	0.53	0.23	0.25	1.0	1.0	1.0	1.0

Table 5.8: Verb thesauri overlaps.

	Reference							
	TeP		**OT.PT**		**CLIP-0.01**		**CLIP-0.075**	
Thesaurus	**Words**	**Synsets**	**Words**	**Synsets**	**Words**	**Synsets**	**Words**	**Synsets**
TeP	1.0	1.0	0.21	0.26	0.57	0.55	0.56	0.49
OT.PT	0.86	0.74	1.0	1.0	0.87	0.76	0.83	0.62
CLIP-fuzzy	0.56	0.70	0.21	0.51	1.0	1.0	0.98	0.98
CLIP-0.075	0.56	0.52	0.21	0.23	1.0	0.62	1.0	1.0

Table 5.9: Adjective thesauri overlaps.

The obtained numbers confirm that these resources are more complementary than overlapping, as it had been noticed when comparing PAPEL, TeP and other resources (Santos et al., 2010), or when comparing the verbs in PAPEL, TeP, OT and Wiktionary.PT (Teixeira et al., 2010). Therefore, once again, we conclude that, instead of using them as gold standards for evaluation, it would be more fruitful to integrate TeP and OT in a broad-coverage Portuguese thesaurus.

Reconstruction of the handcrafted thesauri

The goal of the third automatic evaluation step was to validate the clustering algorithm. This validation consisted of running the clustering algorithm on a network established by the synpairs of the thesaurus, and then observing how far the generated thesaurus was from the original. For this purpose, the thesauri were first converted into synpairs, such that each pair of words in a synset gave rise to one synpair. For instance, the synset $S = \{a, b, c\}$ resulted in the synpairs $\{a, b\}$, $\{a, c\}$, $\{b, a\}$, $\{b, c\}$, $\{c, a\}$ and $\{c, b\}$.

Tables 5.10, 5.11, and 5.12 show the overlaps between the result of clustering TeP (CleP) and the original TeP, for nouns, verbs and adjectives, respectively. For OT, the equivalent overlaps were always between 0.99 and 1.0 and are omitted.

These numbers show that applying our algorithm on synonymy networks is an adequate approach for discovering synsets. Using this algorithm, we could obtain a very similar thesaurus from the synpairs, as the synset overlap is always higher than 0.9. The only exception is the overlap of the verbs in TeP on CleP-0.075 (0.82). This is probably explained by the high ambiguity of the verbs in TeP (see tables 5.5 and 5.6). It should be added that, although all thesauri are based on the same set of words, there are word overlaps below 1.0. These situations indicate that some words were not included in any synset and were thus left out of the thesaurus.

	Reference					
	TeP		**CleP-0.01**		**CleP-0.075**	
Thesaurus	**Words**	**Synsets**	**Words**	**Synsets**	**Words**	**Synsets**
TeP	1.0	1.0	0.99	0.99	0.99	0.92
CleP-0.01	1.0	1.0	1.0	1.0	1.0	0.98
CleP-0.075	1.0	0.97	1.0	1.0	1.0	1.0

Table 5.10: Reconstruction of TeP with the clustering algorithm (nouns).

	Reference					
	TeP		**CleP-0.01**		**CleP-0.075**	
Thesaurus	**Words**	**Synsets**	**Words**	**Synsets**	**Words**	**Synsets**
TeP	1.0	1.0	0.99	0.94	0.98	0.82
CleP-0.01	1.0	0.99	1.0	1.0	0.99	0.98
CleP-0.075	1.0	0.93	1.0	1.0	1.0	1.0

Table 5.11: Reconstruction of TeP with the clustering algorithm (verbs).

	Reference					
	TeP		**CleP-0.01**		**CleP-0.075**	
Thesaurus	**Words**	**Synsets**	**Words**	**Synsets**	**Words**	**Synsets**
TeP	1.0	1.0	0.97	1.0	0.97	0.96
CleP-0.01	1.0	1.0	1.0	1.0	1.0	0.99
CleP-0.075	1.0	0.99	1.0	1.0	1.0	1.0

Table 5.12: Reconstruction of TeP with the clustering algorithm (adjectives).

5.3.5 Manual evaluation

In order to complement the previous validation, we performed the manual evaluation of part of the noun synsets of CLIP. Given that the manual evaluation of fuzzy memberships is a difficult task, for this evaluation, we used CLIP with $\theta = 0.075$ (CLIP-0.075), and did not use the membership values.

Moreover, in order to make manual evaluation faster and less tedious, we selected a subset of the noun synsets in CLIP-0.075. First, we removed all the words without occurrences in the frequency lists of AC/DC[1] (Santos and Bick, 2000), which compile

[1]Available through `http://www.linguateca.pt/ACDC/` (September 2012)

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 on average 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/synpair as either:

- Correct (1): if, in some context, all the words could have the same meaning;
- Incorrect (0): if at least one word could never 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 5.13, show that the average accuracy of CLIP synsets is higher than 73%. Assuming that the samples were completely random, which were not due to the frequency contraints, the margin of error would be 4.1% in a confidence interval of 95%.

Furthermore, the average inter-annotator agreement ($\overline{IAA}$), given by the number of matches between the classifications of two different judges, is higher than 82%. However, the average Kappa coefficient ($\overline{\kappa}$) is 0.42 for synsets and 0.43 for synpairs[2] which, according to Landis and Koch (1977) and Green (1997), shows fair/moderate agreement. The agreement values point out the subjectivity of evaluating synonymy, as it is quite dependent on the judge intuition.

Classification	**Synsets** sample = 440 × 2 synsets		**Synpairs** sample = 440 × 2 synpairs	
Correct	646	(73.4%)	660	(75.0%)
Incorrect	231	(26.3%)	218	(24.8%)
N/A	3	(0.3%)	2	(0.2%)
$\overline{IAA}$	82.7%		83.2%	
$\overline{\kappa}$	0.42		0.43	

Table 5.13: Results of manual evaluation of synsets and synpairs.

When we decided to evaluate our data as synsets and also as synpairs, we intended to have two different perspectives on the thesaurus quality. However, both kinds of evaluation yielded similar results, as the correction of synpairs is 75%. On the one hand, as opposing to a complete synset, it should be easier to select a correct pair and also to decide on its correction. On the other hand, when faced upon a synset, all the words make up a context, which sometimes guides the annotator towards a better decision, while spending more time for longer synsets.

5.4 Discussion

We presented a clustering procedure that was used for the automatic discovery of Portuguese synsets from synonymy networks, acquired from dictionaries. The

[2]Given that we had multiple judges, the average Kappa coefficient is the average of the Kappa coefficients for each pair of samples.

discovered synsets result in CLIP, a Portuguese thesaurus, larger than public domain Portuguese handcrafted thesauri. CLIP was compared with those thesauri, which lead us to the conclusion that we can obtain an even larger thesaurus if we integrate all thesauri. Given that OT.PT, the smaller thesaurus used in our experimentation, is currently used for suggesting synonyms in OpenOffice[3] writer, CLIP can be seen as a larger alternative for the same purpose. Still, since size is not the only important property of a thesaurus, it is always possible to create a smaller thesauri, after filtering less common words.

The proposed algorithm may be used for the creation of a fuzzy thesaurus, where words have membership degrees to each synset. Having in mind that word senses are not discrete, representing natural language concepts as fuzzy synsets is closer to reality than using simple synsets. Moreover, a fuzzy thesaurus is a useful resource for NLP. For instance, in WSD, choosing the synset where the target word has higher membership might be used as a baseline. As far as we know, the fuzzy version of our thesaurus is the first ever Portuguese thesaurus with fuzzy memberships.

The presented approach has however shown some limitations. For instance, a fixed cut-point is probably not the best option while moving from a fuzzy thesaurus to a thesaurus without fuzzy memberships. Therefore, we have recently added the possibility of having a variable cut-point, relative to the highest membership in the set. Possibly the main limitation of our approach is that synsets are not created when word senses do not have dictionary entries with synonyms. This is something we will have to deal in the future. Finally, the manual evaluation showed interesting but not optimal results (75% accuracy), which indicates that there is still room for improvement.

In any case, as we will discuss in the following chapters, CLIP could be used as the synset-base for a future wordnet-like resource. But TeP is a similar public alternative. And as it is created manually by experts, we have high confidence on its contents, so, we decided to use it as the starting point of our synset-base. The next chapter describes how TeP can be enriched with synonymy information extracted from dictionaries, in order to have a broader thesaurus. In order to discover new synsets, only the synpairs not added to TeP are the target of a clustering algorithm, similar to the one presented here.

[3]See `http://www.openoffice.org/` (September 2012)

Chapter 6

Thesaurus Enrichment

General language dictionaries and language thesauri cover the same kind of knowledge, but represent it differently. While the former consist of lists of word senses and respective natural language sense descriptions, the latter group synonymous words together, so that they can be seen as possible lexicalisations of concepts. WordNet (Fellbaum, 1998) can actually be seen as a resource that bridges the gap between both kinds of resources, because each synset contains a textual gloss.

However, in previous chapters, we have shown that, even though they intend to cover the same kind of knowledge, most of the information in public handcrafted Portuguese thesaurus is complementary to the information extracted from dictionaries. Therefore, it should be more fruitful to integrate their information in Onto.PT instead of using them merely as a reference for comparison. Another aspect in favour of this option is that, besides its size, TeP was manually created by experts. This means that, more than integrating the information in TeP, we can take advantage of its structure to have more reliable synsets and more controlled sense granularity.

The work presented in this chapter can be seen both as an alternative or a complement of the previous chapter, as we use the synsets of TeP as a starting point for the construction of a broader thesaurus. To this end, we follow a four-step approach for enriching an existing electronic thesaurus, structured in synsets, with information extracted from electronic dictionaries, represented as synonymy pairs (synpairs)[1]:

1. Extraction of synpairs from dictionary definitions;
2. Assignment of synpairs to suitable synsets of the thesaurus;
3. Discovery of new synsets after clustering the remaining synpairs;
4. Integration of the new synsets in the thesaurus.

In step 1, any approach for the automatic acquisition of synpairs from dictionaries, such as the one described in chapter 4, may be followed. Therefore, we will not go further on this step. We start this chapter by presenting its main contribution, which is the algorithm for the automatic assignment of synpairs to synsets. Then, we evaluate the algorithm against a gold standard and select the most adequate settings for using it in the enrichment of TeP. Any graph clustering procedure suits step 3 of our approach. We chose to follow an approach similar to the one introduced

[1]Synpairs are synonymy tb-triples. They can be extracted from several sources, however, as we are dealing with general language knowledge, dictionaries are the obvious targets.

in chapter 5, described together with step 4. Finally, before a final discussion, we report our work towards the construction of TRIP, a large thesaurus for Portuguese, using the aforementioned four-step approach. Among other information, we discuss on the coverage of the synpairs by TeP, we examine the properties of the networks to be clustered, we evaluate the results of clustering, and compare the evolution of the thesaurus after each enrichment step. Earlier stages of this work were reported in Gonçalo Oliveira and Gomes (2011b). A more recent version was recently submitted to a journal (Gonçalo Oliveira and Gomes, 2012b).

6.1 Automatic Assignment of synpairs to synsets

The goal of this procedure is to select the most suitable synsets of a thesaurus for attaching each synpair, extracted previously. It is assumed that, if a thesaurus contains a word, it contains all its possible senses. We recall that this might not be true, especially because words do not have a finite number of senses (Kilgarriff, 1996). However, broad coverage LKBs typically limit the number of senses of each word. By creating artificial boundaries on meanings, the former knowledge bases become more practical for computational applications.

6.1.1 Goal

By assigning a synpair $p_{xy} = \{v_x, v_y\}$ to a synset $S_a = \{v_1, v_2, ..., v_n\}$, we mean that words v_x and v_y are added to S_a, which becomes $S_a = \{v_1, v_2, ..., v_n, v_x, v_y\}$. Synpair p_{xy} is already in a thesaurus T if there is a synset $S_{xy} \in T$ containing both v_x and v_y. In this case, no assignment is needed. Otherwise, if the thesaurus does not contain the synpair, the adequate synset(s) to assign the synpair is selected. This synset would be the most similar to the synpair.

Before running the assignment algorithm, a synonymy network N is obtained using all the extracted synpairs, such that $p_{xy} = \{v_x, v_y\}$ establishes an edge between nodes v_x and v_y (see more on synonymy networks in section 5.1). Synonymy networks can be represented, for instance, as a binary sparse matrix, the adjacency matrix $M(|V| \times |V|)$, where $|V|$ is the number of edges and $M_j = \vec{v_j}$ is the adjacency vector of the word in v_j. For example, for the node v_1 in the network of figure 6.1:

$$\vec{v_1} = [1, 1, 0, 0, 1, 0, 1, 0, 0, 0]$$

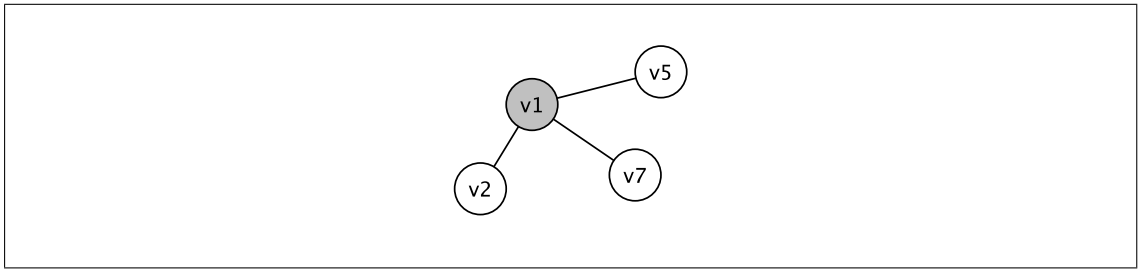


Figure 6.1: Illustrative synonymy network.

6.1.2 Algorithm

The proposed assignment algorithm may be described in the following 5 steps.

For each synpair $p_{xy} = \{v_x, v_y\}, p \in N$:

1. If there is a synset $S_{xy} \in T$ containing both elements of p_{xy}, $v_x \in S_{xy} \wedge v_y \in S_{xy}$, the synpair is already represented in T, so nothing is done. End.
2. Otherwise, select all the candidate synsets $S_j \in C : C \subset T$, $C = \{S_1, S_2, ..., S_n\}$ containing one of the elements of p_{xy}, $\forall(S_j \in C) : v_x \in S_j \vee v_y \in S_j$.
3. Represent the synpair and the candidate synsets as vectors:
   - Synpair vector ($\vec{p_{xy}}$): if $v_x \in S_j$, $\vec{p_{xy}} = \vec{v_y}$, otherwise, $\vec{p_{xy}} = \vec{v_x}$
   - (set of) Synset vectors ($\vec{S_j}$): $S_j = \{v_1, v_2, ..., v_n\}$, $\vec{S_j} = \{\vec{v_1}, \vec{v_2}, ..., \vec{v_n}\}$
4. Compute the average similarity between $\vec{p_{xy}}$ and each candidate synset $S_j \in C$:

$$\overline{sim}(\vec{p_{xy}}, \vec{S_j}) = \frac{1}{1 + \log_2 |S_j|} \sum_{i=1}^{|S_j|} sim(\vec{p_{xy}}, \vec{v_i}), \vec{v_i} \in \vec{S_j}$$

   We do not use the arithmetic mean in order to minimise the bias towards the selection of smaller synsets. This way, the weight of the size is not directly proportional to smaller synsets.
5. Finally, assign p_{xy} to all candidate synsets with similarity higher than a predefined threshold, $C' : \forall(S_k \in C'), sim(\vec{p_{xy}}, \vec{S_k}) \geq \sigma$.

In step 4, any measure for computing the similarity between two binary vectors may be used, including the following, typically used to compute the similarity between sets:

$$Jaccard(v_a, v_b) = \frac{\vec{v_a}.\vec{v_b}}{|\vec{v_a}| + |\vec{v_b}| - \vec{v_a}.\vec{v_b}} \qquad Overlap(v_a, v_b) = \frac{\vec{v_a}.\vec{v_b}}{min(|\vec{v_a}|, |\vec{v_b}|)}$$

$$Dice(v_a, v_b) = 2 \times \frac{\vec{v_a}.\vec{v_b}}{|\vec{v_a}| + |\vec{v_b}|} \qquad Cosine(v_a, v_b) = \frac{\vec{v_a}.\vec{v_b}}{|\vec{v_a}|.|\vec{v_b}|}$$

Moreover, as this kind of similarity is measured according to the information given by the network, larger and more complete graphs will provide better results. So, if the goal is to assign just a few triples, other sources of information, such as occurrences in a corpus, should be considered as alternatives to generate the synonymy network.

In the end, the assignment algorithm may run for another iteration, using the resulting synsets and the remaining synpairs. But the new synsets will be larger and less reliable than the original, which were created manually. Therefore, the threshold σ should be increased in the new iteration.

6.2 Evaluation of the assignment procedure

The evaluation of the assignment procedure has two main goals. First, it quantifies the performance of the assignment algorithm. Second, it enables the selection of the most adequate settings, including the similarity measure and the best threshold σ to use in the integration of the synpairs of PAPEL/CARTÃO in TeP.

6.2.1 The gold resource

To compare the performance of the assignment algorithm using different settings, we randomly selected 355 noun synpairs of PAPEL 2.0 (Gonçalo Oliveira et al., 2010b) and had them assigned, by two human annotators, to the synsets of TeP 2.0 (Maziero et al., 2008). Before their selection, we made sure that all 355 synpairs had at least one candidate synset in TeP. The manually assigned synpairs constitute a small gold collection, used to evaluate the procedure with different settings. Even though the creation of this resource was a time-consuming task, we now have a reference that helps us understand the behavior of the algorithm. Furthermore, it is now possible to repeat this kind of evaluation as many times as needed.

Lexical-semantic knowledge is typically subjective and thus hard to evaluate. Besides depending heavily on the vocabulary range and intuition of the human annotator, when it comes to the division of words into senses, even for expert lexicographers, there is not a consensus because word senses are most of the time fuzzy and also because language evolves everyday (see section 4.3).

In order to minimise this problem, both annotators manually selected the assigments for the same 355 synpairs. On average, there were 4.31 candidate synsets for each synpair with a standard deviation of 3.27. Also on average, the first annotator assigned each synpair to 2.03±1.37 synsets, while, for the second, this number was 2.64±2.30. Their matching assignments were 70% and their kappa agreement 0.43, which means fair/moderate agreement (Landis and Koch, 1977; Green, 1997) and shows, once again, how subjective it is to evaluate this kind of knowledge.

6.2.2 Scoring the assignments

In order to select the best assignment settings, we performed an extensive comparison of the assignment performance, using different similarity measures (introduced in section 6.1.2) and different thresholds σ. In all the experimentation runs, we used all the noun synpairs in CARTÃO, which includes PAPEL 3.0 and the synpairs extracted from Wiktionary.PT and DA, to establish the synonymy network for computing similarities. More about the size of this network and on its coverage by TeP can be found in section 6.4.1.

The evaluation score of each setting was obtained using typical information retrieval measures, namely precision, recall and F-score. For a synpair in the set of assigned synpairs, $p_i \in P$, these measures are computed as follows:

$$Precision_i = \frac{|Selected_i \cap Correct_i|}{|Selected_i|} \qquad Precision = \frac{1}{|P|}\sum_{i=1}^{|P|} Precision_i$$

$$Recall_i = \frac{|Selected_i \cap Correct_i|}{|Correct_i|} \qquad Recall = \frac{1}{|P|}\sum_{i=1}^{|P|} Recall_i$$

$$F_\beta = (1+\beta^2) \times \left(\frac{Precision \times Recall}{(\beta^2 \times Precision) + Recall} \right)$$

Besides the typical F_1-score, we computed $F_{0.5}$, which favours precision over recall. We prefer to have a more reliable resource, rather than a larger resource with lower correction. Furthermore, the synpairs not assigned to synsets will have a second chance of being integrated in the thesaurus, during the clustering step.

Since there could be more than one possible adequate synset for a synpair, in addition the aforementioned measures, we computed a relaxed recall (*RelRecall*). For a single synpair, *RelRecall* is 1 if at least one correct synset is selected:

$$RelRecall_i = \begin{cases} 1, & \text{if } |Selected \cap Correct|_i > 0 \\ 0, & \text{otherwise} \end{cases} \qquad RelRecall = \frac{1}{|P|}\sum_{i=1}^{|P|} RelRecall_i$$

Using *RelRecall*, we may as well compute the relaxed F_β, $RelF_\beta$.

$$RelF_\beta = (1+\beta^2) \times \left(\frac{Precision \times RelRecall}{(\beta^2 \times Precision) + RelRecall} \right)$$

6.2.3 Comparing different assignment settings

Tables 6.1, 6.2, and 6.3 present the evaluation scores of assignments using different settings, respectively against the references of annotator 1, annotator 2, and the intersection between annotator 1 and annotator 2. For each synpair, the intersection reference includes just the synsets selected by both annotators, and has consequently lower scores. Also, although we ran this evaluation in a wider range of values for σ, for the sake of simplicity, we only present those more relevant for understanding the behaviour of the algorithm. In the tables, we have also included the scores if all the candidates were selected (*All*), which can be seen as a baseline. Another baseline for precision is the random chance of selecting a correct candidate, which is 59.4%, 67.8% and 48.8%, respectively for annotator 1, annotator 2 and for the intersection.

The similarity measures are an indicator for the synset assignment, and they are applied in two different modes:

- *Best*: only the best candidate synset with similarity equal or higher than σ is selected. More than one synset may be selected, but only if there is a tie.
- *All*: all the synsets with similarity equal or higher than σ are selected.

As expected, better precisions are obtained with higher values of σ. The best precision (around 82% and 92%) is consistently obtained with the cosine measure, mode *All*, and $\sigma = 0.35$. There is also no surprise on the best recall, which is, of course, 100% for the baseline using all candidate synsets.

Table 6.1: Evaluation against annotator 1

Measure	Mode	σ	Sets/Pair	P	R	RR	F_1	RF_1	$F_{0.5}$	$RF_{0.5}$
All	–	–	4.31	0.47	**1.00**	**1.00**	0.64	0.64	0.53	0.53
Jaccard	*Best*	0.00	1.07	0.73	0.35	0.79	0.47	**0.76**	0.60	0.74
		0.15	0.52	0.84	0.18	0.45	0.30	0.59	0.49	0.72
	All	0.05	3.14	0.55	0.81	0.96	0.65	0.70	0.59	0.60
		0.1	1.52	0.70	0.45	0.69	0.55	0.69	**0.63**	0.70
		0.15	0.77	0.81	0.28	0.48	0.41	0.60	0.59	0.71
		0.2	0.41	0.87	0.17	0.32	0.29	0.47	0.48	0.65
Overlap	*Best*	0.00	1.06	0.73	0.35	0.78	0.47	0.75	0.60	0.74
		0.15	1.04	0.74	0.34	0.77	0.47	0.75	0.60	0.74
	All	0.1	3.98	0.49	0.95	0.98	0.65	0.66	0.55	0.55
		0.4	1.28	0.72	0.37	0.61	0.49	0.66	0.60	0.70
		0.45	1.05	0.75	0.31	0.53	0.44	0.62	0.58	0.69
		0.5	0.88	0.79	0.28	0.48	0.41	0.60	0.58	0.70
		0.55	0.65	0.81	0.19	0.37	0.31	0.51	0.50	0.66
		0.6	0.50	0.86	0.15	0.31	0.26	0.45	0.45	0.63
Dice	*Best*	0.00	1.06	0.73	0.35	0.79	0.47	**0.76**	0.60	0.74
		0.15	0.88	0.77	0.30	0.69	0.43	0.73	0.58	**0.75**
	All	0.1	2.97	0.56	0.79	0.95	**0.66**	0.71	0.60	0.61
		0.15	2.00	0.64	0.56	0.79	0.60	0.71	**0.63**	0.67
		0.2	1.26	0.73	0.39	0.63	0.51	0.67	0.62	0.71
		0.25	0.81	0.82	0.29	0.50	0.43	0.62	0.60	0.73
		0.3	0.55	0.86	0.20	0.38	0.33	0.53	0.52	0.69
		0.35	0.35	**0.88**	0.14	0.27	0.23	0.41	0.42	0.61
Cosine	*Best*	0.00	1.05	0.73	0.34	0.78	0.46	0.75	0.59	0.74
		0.15	0.94	0.76	0.31	0.73	0.44	0.75	0.59	**0.75**
	All	0.1	3.34	0.54	0.85	0.97	**0.66**	0.69	0.58	0.59
		0.15	2.40	0.62	0.67	0.88	0.65	0.73	**0.63**	0.66
		0.2	1.58	0.69	0.46	0.70	0.55	0.70	**0.63**	0.69
		0.25	1.08	0.76	0.35	0.58	0.48	0.66	0.62	0.72
		0.3	0.74	0.84	0.26	0.47	0.39	0.60	0.58	0.73
		0.35	0.48	**0.88**	0.18	0.35	0.30	0.50	0.50	0.67
		0.4	0.32	0.86	0.12	0.24	0.22	0.38	0.39	0.57

When it comes to the other scores, the choice of the best setting is not completely clear, as several scores are very close. For instance, the best F_1, using the second annotator as a reference, is obtained by the baseline of all candidates. This is quite surprising, but is explained by the fact that this annotator selected, on average, more than half of the candidates as correct (see section 6.2.1). For the other references, the best F_1 is obtained by all candidates with cosine above a low σ (0.1 and 0.15).

Nevertheless, the settings with the best F_1 scores have low precision (below 60%) which, for the sake of the reliability, and consequent usability, of the resource, is more important than recall. Also, we should remind that the words in unassigned synpairs will be integrated later in the thesaurus, after clustering. Looking at $F_{0.5}$ and $RF_{0.5}$ scores, the best settings are still not clear but, using the cosine similarity it is possible to obtain the best values or very close to the best. Using that measure and annotators 1 and 2 as reference, $\sigma = 0.15$ with mode *All* achieves the best $F_{0.5}$, while the mode *Best* is better for the intersection. On the other hand, $RF_{0.5}$ is more consistent across references. Its best result is always obtained with the cosine similarity measure, $\sigma = 0.15$ and mode *Best*.

In the experimentation described in section 6.4, we decided to support our choice in $RF_{0.5}$, and thus used the cosine measure, mode *Best* and $\sigma = 0.15$. With these settings, precision is still just 66% for the intersection of annotators, but it is 76% using the first annotator as reference and 81% using the second. We admit that there could be arguments for selecting different settings.

Measure	Mode	σ	Sets/Pair	P	R	RR	F_1	RF_1	$F_{0.5}$	$RF_{0.5}$
All	–	–	4.31	0.61	**1.00**	**1.00**	**0.76**	0.76	0.66	0.66
Jaccard	*Best*	0.00	1.07	0.78	0.37	0.92	0.51	0.84	0.64	0.81
		0.15	0.52	0.83	0.2	0.49	0.32	0.61	0.51	0.73
	All	0.05	3.13	0.69	0.75	0.97	0.72	0.80	0.70	0.73
		0.1	1.52	0.79	0.42	0.73	0.55	0.76	0.67	0.77
		0.15	0.77	0.83	0.26	0.49	0.40	0.62	0.58	0.73
		0.2	0.41	0.88	0.16	0.34	0.28	0.49	0.47	0.67
Overlap	*Best*	0.00	1.06	0.78	0.37	0.90	0.50	0.84	0.63	0.80
		0.15	1.04	0.78	0.36	0.9	0.49	0.83	0.63	0.80
	All	0.1	3.97	0.63	0.93	1.00	0.75	0.77	0.68	0.68
		0.4	1.28	0.81	0.32	0.64	0.45	0.71	0.62	0.77
		0.45	1.05	0.82	0.25	0.56	0.39	0.67	0.57	0.75
		0.5	0.88	0.84	0.23	0.51	0.36	0.64	0.55	0.74
		0.55	0.65	0.87	0.18	0.39	0.29	0.54	0.49	0.70
		0.6	0.50	0.89	0.13	0.32	0.23	0.47	0.41	0.66
Dice	*Best*	0.00	1.06	0.78	0.37	0.92	0.51	0.84	0.64	0.81
		0.15	0.88	0.81	0.32	0.81	0.46	0.81	0.62	0.81
	All	0.1	2.97	0.69	0.71	0.96	0.70	0.80	0.69	0.73
		0.15	2.00	0.75	0.52	0.85	0.61	0.80	0.69	0.77
		0.2	1.26	0.80	0.34	0.66	0.48	0.72	0.63	0.77
		0.25	0.81	0.84	0.27	0.52	0.41	0.64	0.59	0.75
		0.3	0.55	0.88	0.19	0.41	0.31	0.56	0.51	0.71
		0.35	0.35	**0.92**	0.12	0.29	0.21	0.44	0.39	0.64
Cosine	*Best*	0.00	1.05	0.79	0.37	0.92	0.51	**0.85**	0.65	0.81
		0.15	0.94	0.81	0.34	0.85	0.48	0.83	0.64	**0.82**
	All	0.1	3.34	0.67	0.78	0.97	0.72	0.80	0.69	0.72
		0.15	2.40	0.74	0.61	0.91	0.67	0.82	**0.71**	0.77
		0.2	1.58	0.77	0.40	0.74	0.53	0.76	0.65	0.77
		0.25	1.08	0.82	0.30	0.60	0.44	0.69	0.61	0.76
		0.3	0.74	0.85	0.24	0.50	0.38	0.63	0.57	0.75
		0.35	0.48	**0.92**	0.16	0.37	0.27	0.53	0.47	0.71
		0.4	0.32	**0.92**	0.12	0.27	0.21	0.42	0.39	0.62

Table 6.2: Evaluation against annotator 2.

6.3 Clustering and integrating new synsets

After the assignment stage, there are synpairs that have not been assigned to a synset, either because:

- The thesaurus does not contain any of the synpair elements, and there are thus no assignment candidates.
- All similarities between the synpair and the candidate synsets are lower than σ.

During the clustering stage, new synsets are discovered from the network established by the those synpairs, $N' = (V, E)$. The main difference between N' and N is that N' will have both less nodes and edges per node. Furthermore, depending on the σ used, N' might have similar properties to N or, for lower σ, N' tends to be constituted by several small isolated subgraphs, where all words have a common meaning. Either way, the goal of this stage is to identify new meanings, not covered by the thesaurus. Given that typical synonymy networks extracted from dictionaries tend to have a clustered structure, some authors (Gfeller et al., 2005; Navarro et al., 2009; Gonçalo Oliveira and Gomes, 2011a) propose a clustering algorithm for discovering clusters, which we believe that can be seen as synsets.

This stage is inspired by the aforementioned works, and involves the application of a clustering algorithm to N'. At this point, the majority of the ambiguous words, which are those with more connections, are expected to be already included in synsets of the thesaurus. Therefore, a simplified version of the clustering procedure introduced in chapter 5 is suitable for discovering synsets in the network N':

Measure	Mode	σ	Sets/Pair	P	R	RR	F_1	RF_1	$F_{0.5}$	$RF_{0.5}$
All	–	–	4.31	0.37	**1.00**	**1.00**	0.54	0.54	0.42	0.42
Jaccard	*Best*	0.00	1.07	0.63	0.49	0.81	0.55	**0.71**	0.59	0.66
		0.15	0.52	0.72	0.25	0.45	0.37	0.56	0.53	0.65
	All	0.05	3.14	0.45	0.86	0.96	0.59	0.61	0.49	0.50
		0.1	1.52	0.59	0.52	0.69	0.55	0.63	0.57	0.61
		0.15	0.77	0.70	0.33	0.48	0.45	0.57	0.57	0.64
		0.2	0.41	0.78	0.20	0.32	0.32	0.46	0.50	0.61
Overlap	*Best*	0.00	1.06	0.63	0.48	0.81	0.55	**0.71**	**0.60**	0.66
		0.15	1.04	0.64	0.47	0.80	0.54	**0.71**	**0.60**	0.67
	All	0.1	3.98	0.39	0.96	0.98	0.55	0.56	0.44	0.44
		0.4	1.28	0.63	0.44	0.61	0.52	0.62	0.58	0.63
		0.45	1.05	0.67	0.37	0.53	0.47	0.59	0.57	0.63
		0.5	0.88	0.70	0.32	0.48	0.44	0.57	0.57	0.64
		0.55	0.65	0.74	0.25	0.37	0.37	0.50	0.53	0.62
		0.6	0.50	0.79	0.19	0.31	0.30	0.45	0.48	0.60
Dice	*Best*	0.00	1.06	0.63	0.49	0.81	0.55	**0.71**	**0.60**	0.66
		0.15	0.87	0.64	0.45	0.7	0.53	0.67	0.59	0.65
	All	0.1	2.97	0.46	0.85	0.95	0.60	0.62	0.51	0.51
		0.15	2.00	0.54	0.65	0.80	0.59	0.64	0.56	0.58
		0.2	1.26	0.62	0.45	0.63	0.52	0.62	0.57	0.62
		0.25	0.81	0.71	0.35	0.50	0.47	0.59	0.59	0.66
		0.3	0.55	0.77	0.25	0.38	0.38	0.51	0.54	0.64
		0.35	0.35	0.81	0.16	0.27	0.26	0.41	0.44	0.58
Cosine	*Best*	0.00	1.05	0.64	0.48	0.81	0.55	**0.71**	**0.60**	0.66
		0.15	0.94	0.66	0.45	0.75	0.53	0.70	**0.60**	**0.68**
	All	0.1	3.34	0.44	0.89	0.97	0.59	0.60	0.49	0.49
		0.15	2.40	0.52	0.75	0.88	**0.61**	0.66	0.55	0.57
		0.2	1.58	0.58	0.53	0.70	0.55	0.64	0.57	0.60
		0.25	1.08	0.66	0.41	0.58	0.51	0.61	0.59	0.64
		0.3	0.74	0.74	0.32	0.48	0.45	0.58	0.59	0.67
		0.35	0.48	**0.82**	0.21	0.35	0.34	0.49	0.52	0.64
		0.4	0.32	0.80	0.15	0.25	0.25	0.37	0.43	0.55

Table 6.3: Evaluation against intersection of annotators 1 and 2.

1. Create a new sparse matrix $M(|V| \times |V|)$.
2. In each cell M_{ij}, put the similarity between the adjacency vectors of the word in v_i with the adjacency vectors of v_j, $M_{ij} = sim(\vec{v}_i, \vec{v}_j)$;
3. Extract a cluster C_i from each row M_i, consisting of the words v_j where $M_{ij} > \theta$, a selected threshold. A lower θ leads to larger synsets and higher ambiguity, while a larger θ will result on less and smaller synsets or no synsets at all.
4. For each cluster C_i with all elements included in a larger cluster C_j ($C_i \cup C_j = C_j$ and $C_i \cap C_j = C_i$), C_i and C_j are merged, giving rise to a new cluster C_k with the same elements of C_j.

After clustering, we will have a thesaurus T with synsets S_i and a set of discovered clusters C. A simple thing to do would be to handle the clusters as synsets and add then to the thesaurus. However, some of the clusters might be already included or partly included in existing synsets. Therefore, before adding the clusters to T, we compute the similarity between the words in each synset S_i and the words in each discovered cluster C_j. For this purpose, we measure the overlap between the former sets, using the overlap coefficient:

$$Overlap(S_i, C_j) = \frac{|S_i \cap C_||}{min(|S_i|, |C_j|)} \tag{6.1}$$

For each cluster, we only select the synset with the highest overlap. Then, if the overlap is higher than a threshold μ, we merge the cluster with the synset. Otherwise, we add the cluster to T.

6.4 A large thesaurus for Portuguese

In this section, we describe the steps performed towards the automatic enrichment of TeP 2.0 with lexical items in PAPEL 3.0, DA, Wiktionary.PT and also OpenThesaurus.PT. The latter was added to our synonymy network after its conversion to synpairs, given that each pair of words in the same synset establishes a synpair. The result is TRIP, a larger and broader Portuguese thesaurus that, so far, integrates synonymy information of five public domain lexical-semantic resources.

In order to understand the amount of new synonymy information, we first observed the coverage of the synpairs from the other resources by TeP. These numbers are presented in this section. Then, we report on the assignment step, where the settings were selected after the observation of the numbers in section 6.2.3. On the clustering step, we describe the properties of our synonymy network and report on the evaluation of the obtained clusters, together with clustering examples. Finally, we present the evolution of the original thesaurus, after each stage, until it gets to its final form.

6.4.1 Coverage of the synpairs

After removing symmetric synpairs, we collected 67,401 noun, 28,895 verb, and 34,844 adjective synpairs from the four resources used. According to their coverage by TeP, there are different kinds of synpairs. Some of them are already represented, which means that TeP has at least one synset containing both elements of the synpair. For other synpairs, there is only one synset with one of its elements ($|C| = 1$), or several synsets containing one of the synpair elements ($|C| > 1$). Finally, there are synpairs without candidate synsets in TeP ($|C| = 0$).

Table 6.4 summarises the former numbers according to the POS of the synpairs, together with the average number of candidates for each synpair not represented in TeP ($|\bar{C}|$). Depending on the POS, the synpair proportions are different. For instance, almost half of the verb synpairs is already represented in TeP, while, for nouns, this proportion is 22.5%. The proportion of synpairs without candidate synsets in TeP is lower for adjectives (7.6%) and verbs (2.1%), but is 22.1% for nouns. At the same time, for each POS, more than 40% of the synpairs have more than one candidate synset in TeP.

Synpairs with at least one candidate in TeP are those exploited in the assignment stage, where they might be assigned to a synset. On the other hand, synpairs without synset candidates have 100% new vocabulary. Their only chance to be added to the thesaurus is by being included in a cluster, in the clustering stage.

6.4.2 Assignment of synpairs to synsets

After observing the evaluation of the assignment procedure, we decided to use the cosine similarity, mode *Best*, with $\sigma = 0.15$, which, against all three human references, obtained the best $RF_{0.5}$. Though, we added a second assignment iteration,

	POS		
	Noun	**Verb**	**Adjective**
Synpairs	61,025	28,895	34,844
In TeP	15,183 (22.5%)	13,891 (48.1%)	11,930 (34.2%)
$\|C\| = 0$	14,902 (22.1%)	615 (2.1%)	2,659 (7.6%)
$\|C\| = 1$	8,902 (13.2%)	960 (3.3%)	3,365 (9.7%)
$\|C\| > 1$	28,414 (42.2%)	13,429 (46.6%)	16,890 (48.5%)
$\|\overline{C}\|$	4.30	8.49	4.34

Table 6.4: Coverage of the synpairs by TeP.

where synpairs have a second chance of being assigned to a synset, this time using the same similarity measure, but with a higher threshold, $\sigma = 0.35$. The previous value obtained the best precision in the mode *All* and, once again, against all human references. The second iteration intends to integrate unassigned synpairs, in which, after the first iteration, there is high confidence on the assignment to a synset.

After the assignment stage, 37,767 noun, 14,459 verb and 20,310 adjective synpairs were assigned to, at least, one TeP synset. Of those, respectively 35,247, 14,246 and 19,595 were assigned during the first iteration and 2,520, 213 and 715 during the second. Table 6.5 presents examples of real assignments and the iteration where they were accomplished[2].

It.	**Synpair**	**Synset**
1^{st}	{***alimentação***, *mantença*}	{*sustento, alimento, mantimento, alimentação*}
1^{st}	{***escravizar***, *servilizar*}	{*oprimir, tiranizar, escravizar, esmagar*}
1^{st}	{***permanente***, *inextinguível*}	{*durador, duradoiro, duradouro, durável, permanente, perdurável*}
2^{nd}	{***cortadura***, *cortadela*}	{*golpe, cisão, cortadela, rasgue, corte, incisura, rasgo, cortadura, incisão*}
2^{nd}	{***reificar***, *substancializar*}	{*realizar, coisificar, efetivar, efeituar, consumar, efectivar, efetuar, concretizar, reificar, hipostasiar, substantificar*}
2^{nd}	{*encorajante*, ***entusiasmante***}	{*empolgante, entusiasmante, galvanizante, galvanizador*}

Table 6.5: Examples of assignments.

6.4.3 Clustering for new synsets

In order to discover new synsets, the clustering procedure in section 6.3 was applied to the remaining synpairs, with $\theta = 0.5$. Before clustering, we analysed some properties of the synonymy networks they form. After clustering, some of the obtained results were evaluated manually.

The discovered clusters were integrated in the enriched TeP, following the integration procedure described in section 6.3, using a threshold $\mu = 0.5$, empirically defined. Not many synsets were however merged. More precisely 81 noun, 16 verb and 29 adjective clusters were merged to existing synsets. The rest of the clusters were added as new synsets.

[2]Intentionally, no translations are provided because, if translated, most of the provided examples would not capture the essence of this task.

Properties of the synonymy networks

In a similar fashion to what was done for the complete network (table 5.1), table 6.6 contains the total number of nodes ($|V|$) and edges ($|E|$), and the average network degree ($\overline{deg}(N)$, computed according to expression 5.5). It contains as well the number of sub-networks (Sub-nets), which are group of nodes connected directly or indirectly in N; the number of nodes of the largest and second largest sub-networks ($|V_{lcs}|$ and $|V_{lcs2}|$); and the average clustering coefficient of the largest sub-network ($\overline{CC_{lcs}}$, computed according to expression 5.7).

From table 6.6, we notice that these synonymy networks are significantly different from the original. First, they are smaller, as they only contain about half of the nouns, one sixth of the verbs and one third of the adjectives. Second, they have substantially lower degrees, and clustering coefficients close to 0, which means they are less connected and do not tend to form clusters. Nevertheless, they still have one large core sub-network and several smaller.

This confirms that a simpler clustering algorithm is suitable for our purpose, especially because ambiguity is much lower and several clusters are already defined by complete small sub-networks. The noun network contains 4,470 sub-networks of size 2 and 1,127 of size 3. These numbers are respectively 437 and 97 for verbs, and 1,303 and 262 for adjectives.

POS	$\|V\|$	$\|E\|$	$\overline{deg}(N)$	Sub-nets	$\|V_{lcs}\|$	$\overline{CC_{lcs}}$	$\|V_{lcs2}\|$
Noun	21,272	15,294	1.44	6,556	2,816	0.03	66
Verb	1,807	1,197	1.32	614	153	0.00	29
Adjective	4,695	3,050	1.30	1,743	169	0.02	50

Table 6.6: Properties of the synonymy networks remaining after assignment.

Clustering Examples

Figures 6.2, 6.3 and 6.4 illustrate the result of clustering in three sub-networks. The first sub-network results in only one cluster, with several synonyms for someone who speaks Greek. The second and the third are divided into different clusters, represented by different shades of grey.

In figure 6.3, the sub-network is divided in two different meanings of the verb 'splash', one of them more abstract (*esparrinhar*), and the other done with the feet or hands (*bachicar*), but three words may be used with both meanings. The meanings covered by the four clusters in figure 6.4 are, respectively: a person who gives moral qualities; a person who evangelises; a person who spreads ideas; and a person who is an active member of a cause.

Evaluation of the clustering results

In order to check if the algorithm described in section 6.3 is efficient, and to have an idea on the quality of the discovered clusters, their manual evaluation was performed. Once again, we had two judges classifying pairs of words, collected from the same synset, as synonymous or not. This kind of evaluation is easier and slightly less subjective than the evaluation of complete synsets. Furthermore, in section 5.3.5 we reported similar results using both kinds of evaluation.

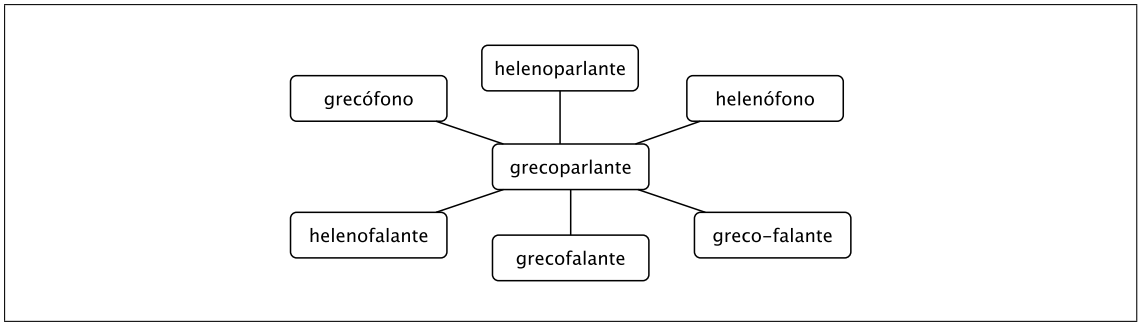


Figure 6.2: Sub-network that results in one adjective cluster – Greek speaker.

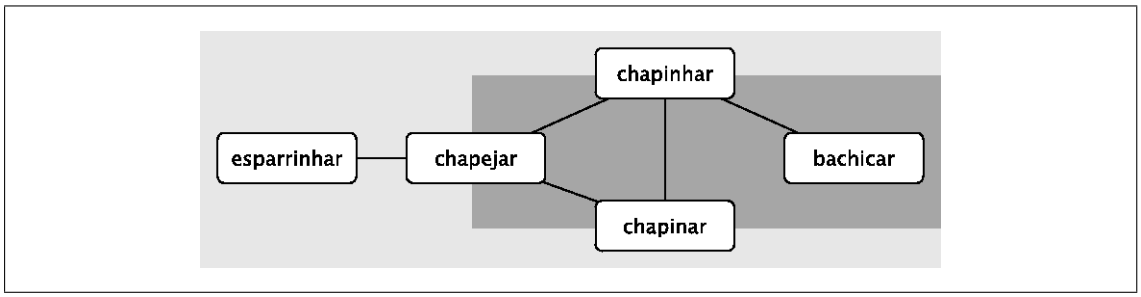


Figure 6.3: Sub-network and resulting verb clusters – two meanings of 'splash'.

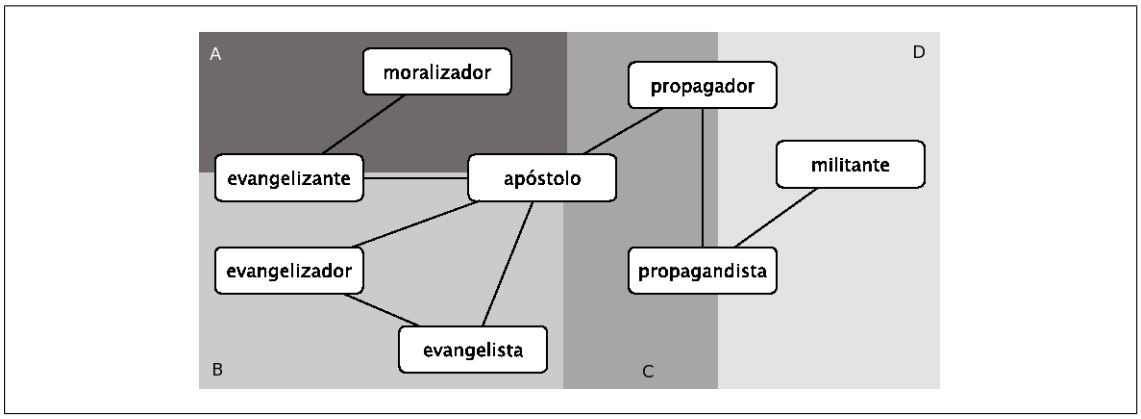


Figure 6.4: Sub-network and resulting noun clusters – a person who: (A) gives moral qualities; (B) evangelises; (C) spreads ideas; (D) is an active member of a cause.

Table 6.7 presents the results of the classification of 330 pairs, 110 from each POS, according to the POS and the judge. For each judge, we present the margin of error at 95% confidence level (ME), where the population is established by all the possible synpairs obtained from the synsets discovered after clustering. The judge agreement is presented by the number of matching classifications between both judges (Matches(1,2)) and their Kappa agreement ($\kappa(1,2)$).

This evaluation confirmed that, as the networks are simpler, the quality of clustering is higher than when the whole network is used (see section 5.3.5). The highest accuracy is that of the adjective synsets, higher than 93% for both judges. Nouns are about 89% accurate for both judges, while verbs are the POS where there is a

POS	Pairs	Judge 1		Judge 2		Matches(1,2)	$\kappa(1,2)$
		Correct	ME	Correct	ME		
Nouns	110	98 (89.1%)	5.8%	98 (89.1%)	5.8%	104 (94.5%)	0.72
Verbs	110	91 (82.7%)	6.8%	98 (89.1%)	5.6%	97 (88.2%)	0.52
Adjs.	110	105 (95.5%)	3.6%	103 (93.6%)	4.4%	106 (96.4%)	0.65

Table 6.7: Evaluation of clustering: correct pairs

higher discrepancy on the accuracy, according to the judge – it is 82% for the first judge and 89% for the second.

One problem when extracting verb synonyms from dictionaries, which was not significant in the evaluation of the synonymy tb-triples, stood out in this evaluation. There are definitions of verbs with the same structure of enumerations, but the last term is a verb on the gerund, referring to an action to specify the previous verb(s). In these cases, the extracted relations should not be synonymy but, perhaps, entailment. This problem, illustrated by the following examples, is partly corrected in further versions of CARTÃO. This was done by discarding synonymy relations where the first argument is a verb on the gerund.

- `circum-navegar - rodear, navegando` (to surround, while sailing)
  → *rodear* synonym-of *circum-navegar*
  → ~~*navegando* synonym-of *circum-navegar*~~
- `palhetear - conversar, mofando` (to talk, while making fun)
  → *conversar* synonym-of *palhetar*
  → ~~*mofando* synonym-of *palhetar*~~

Although not perfect, judge agreement is higher than for other manual evaluations presented in this thesis. It is substantial for nouns and adjectives and moderate for verbs (Landis and Koch, 1977). Once again, the agreement is lower for the less accurate parametters.

6.4.4 Resulting thesaurus

To summarise the evolution of TeP through all the enrichment steps, in tables 6.8 and 6.9 we present the properties of the thesauri after each step, while characterising the thesauri in terms of words and synsets respectively. The thesauri are presented according to the POS of their words, and in the following order:

- TeP 2.0: the original thesaurus;
- 1^{st} iteration: refers to TeP after the first assignment iteration;
- 2^{nd} iteration: refers to the thesaurus after the second assignment iteration;
- Clusters: is the thesaurus consisting only of the synsets discovered after clustering;
- TRIP: the final thesaurus, obtained after the integration of the previous clusters in the thesaurus that resulted from the second iteration.

On the words of each thesaurus (table 6.8) we present the quantity of unique words (Total), the number of words with more than one sense (Ambiguous), the number of average senses per word (Avg(senses)) and the number of senses of the

most ambiguous word (Max(senses)). On the synsets (table 6.9), we present their quantity (Total), their average size in terms of words (Avg(size)), the number of synsets of size 2 ($size = 2$) and size greater than 25 ($size > 25$) and, also, the size of the largest synset ($max(size)$).

Thesaurus	POS	Words			
		Total	Ambiguous	Avg(senses)	Max(senses)
TeP 2.0	Noun	17,149	5,802	1.71	20
	Verb	8,280	4,680	2.69	50
	Adjective	14,568	3,730	1.46	19
	Adverb	1,095	227	1.30	11
1^{st} iteration	Noun	28,693	11,794	1.98	22
	Verb	11,272	6,357	2.85	50
	Adjective	19,148	7,149	1.85	21
	Adverb	1,865	499	1.40	12
2^{nd} iteration	Noun	29,223	11,988	1.99	22
	Verb	11,301	6,374	2.86	50
	Adjective	19,291	7,213	1.85	21
	Adverb	1,914	513	1.40	12
Clusters	Noun	21,126	2,196	1.14	5
	Verb	1,801	177	1.13	4
	Adjective	4,687	359	1.10	5
	Adverb	743	89	1.15	3
TRIP	Noun	45,457	15,392	1.80	22
	Verb	11,924	6,607	2.87	52
	Adjective	22,316	7,782	1.83	22
	Adverb	2,488	694	1.42	12

Table 6.8: Thesauri comparison in terms of words.

After the assignments, the number of words grows and the number of synsets becomes slightly lower. This might seem strange, but as some synsets in TeP are very similar to each other, after the assignments, they become the same synset, and one of them is discarded. Furthermore, as expected, ambiguity becomes higher at this stage. As there is the same number of synsets, but more words, some words are added to more than one synset. And the synsets also become larger, as they are augmented.

The thesaurus obtained after clustering is smaller and much less ambiguous than the others. Besides the high threshold ($\theta = 0.5$), this happens because the words not covered by TeP tend to be less frequent, which are typically more specific and thus less ambiguous. Nevertheless, for nouns, there is still a synset with 31 words.

The words in TRIP are slightly more ambiguous than in TeP and the synsets of TRIP are also larger than TeP's. It is clear that TRIP is much larger than TeP. It contains about two and a half times more noun and adverb lexical items, about 3,500 more verbs and 8,000 more adjectives. The highest number of synsets means that the new thesaurus is broader also in terms of covered natural language concepts. On the other hand, the new thesaurus is more ambiguous and has larger synsets. For instance, it has almost 600 synsets with more than 25 words, which can be seen as too large for being practical (Borin and Forsberg, 2010). TeP has just 66 of those synsets. Nevertheless, we have looked to the largest synsets of TRIP and noticed that most of them are well-formed as they only contain synonymous words.

Thesaurus	POS	Synsets				
		Total	**Avg(size)**	$size = 2$	$size > 25$	$max(size)$
TeP 2.0	Noun	8,254	3.56	3,083	0	21
	Verb	3,899	5.71	907	48	53
	Adjective	6,062	3.5	3,032	18	43
	Adverb	497	2.87	258	0	9
1^{st} iteration	Noun	8,126	7.00	1,227	203	125
	Verb	3,639	8.84	406	189	131
	Adjective	5,945	5.04	1,923	89	87
	Adverb	494	5.28	103	1	27
2^{nd} iteration	Noun	8,126	7.15	1,227	225	129
	Verb	3,639	8.87	406	193	132
	Adjective	5,914	6.05	1,806	161	117
	Adverb	494	5.41	103	1	27
Clusters	Noun	8,879	2.70	4,765	1	31
	Verb	801	2.54	467	0	5
	Adjective	2,063	2.50	1,325	0	8
	Adverb	319	2.68	167	0	7
TRIP	Noun	16,936	4.84	5,986	226	131
	Verb	4,424	7.75	873	193	132
	Adjective	7,948	5.14	3,127	161	117
	Adverb	813	4.34	270	1	27

Table 6.9: Thesauri comparison in terms of synsets.

Largest synsets

Out of curiosity, the largest noun synsets of TRIP refer to concepts that have several figurative and (most of the times) slang synonyms, typically used as insults. For instance, the following are the three largest noun synsets, which denote, respectively, disorder/confusion, alcoholic intoxication, and an imbecile/stupid person:

- *furdúncio, aldrabice, matalotagem, fuzuê,* ***rondão****, desfeita, vergonha, sobresalto, salada_russa, borogodó, latomia, trapizarga, tranquibérnia, alarma, debandada, atabalhoação, siricutico, desorganização, miscelânea, turvação, sarapatel, valverde, equívoco, recacau, canvanza, caravançarai, bafafá,* ***atarantação****, baderna, baralha,* ***baralhada****, cancaburra, rebuliço, salgalhada,* ***barafunda****, abstrusidade,* ***mistifório****, assarapantamento, rebúmbio,* ***trapalhice****, brenha,* ***roldão****,* ***sarrabulhada****, caos, dédalo, estrilho, revolvimento, enovelamento,* ***trapalhada****, barulho, kanvuanza, javardice, embrolho, desordem, desmanho, vasqueiro, forrobodó, garabulha, timaca, pastelada, zona, anarquia,* ***confusão****, rodilhão, floresta, bolo, complicação, feijoada, remexida, amalgamação, sarilho, saricoté,* ***atrapalhação****, feira, foguete, marafunda,* ***salsada****, cambulha, sarrabulho, desarranjo,* ***pipoco****, atropelamento, mixórdia, arranca-rabo, babel, inferno, pessegada, imbróglio, marmelada, choldraboldra, ensalsada, vuvu, bambá, caldeirada, mastigada, maka, ataranto, encrequilha, baixaria, sururu, cegarrega, zorra,* ***salada****,* ***atabalhoamento****, mexida, badanal, escangalho, precipitação, chirinola, enredo, vira-teimão,* ***rolo****, cu-de-boi, desarrumação, embrulhada, indistincção, estricote, envolta, salseiro, enredia, mexedura, atropelo, bagunça, fula-fula, misturada, desconcerto, labirinto, cambulhada, cafarnaum*
- *torcida,* ***embriagamento****, veneno, mona, zurca, trapisonda, lontra, rosca, perua, raposada, rola, tertúlia, carraspana, peleira, pizorga, cabra, chuva, tachada, caroça, ardina, girgolina, égua, carrega, zerenamora, rasca, touca, venena, gardunho, ema, porre, ebriez, carapanta, chiba,* ***ebriedade****, bico,* ***inebriamento****,* ***bebedeira****, carrapata, penca, taçada, canja, garça, ganso, tortelia, turca, cabrita, mela, resina, senisga,* ***bebedice****, bezana, vinhaça,* ***zangurrina****, bêbeda, bibra,* ***borrachice****, zuca, coca, torta, doninha, piela, graxa, trabuzana, água, cegonha, gateira, bicancra, samatra, galinhola, gata, pala, ganza, pifão, bode, cobra,* ***prego****, zola, nêspera, narda, parrascana, vinho, gardinhola, tropecina,* ***embriaguez****, cardina, tiorga, temulência, narceja, pisorga, grossura, dosa, trovoada, carneira, perunca, bruega,*

canjica, raposa, garrana, ***raposeira****, cartola, cachorra, entusiasmo, carpanta, piteira,* ***borracheira****, cabeleira, carrocha, pifo, camoeca, marta, cachaceira,* ***zangurriana****, verniz, carrada*

- *patamaz, boca-aberta, imbecil, lucas, malhadeiro, orate, zé-cuecas, lerdaço, tantã, boleima, babão, jato, zambana,* ***badó****, ânsar, bolônio, chapetão,* ***parvalhão****, haule, papa-moscas, lerdo,* ***patau****, sànona, perturbado, possidónio,* ***babaquara****,* ***tolo****, galafura, babuíno,* ***zângano****, inepto, badana, cabaça, andor, pax-vóbis, idiota, pascoal-bailão, sandeu, asneirão, zé, capadócio, calino, doudivanas, pasguate, parreco, babanca, palerma, molusco, parrana, moco, ansarinho, bajoujo, burro, truão, estulto, pexote, maninelo, lérias, banana, banazola, patego, bobo, estúpido, asno, sonso, ignorante,* ***troixa****,* ***otário****,* ***simplório****, pancrácio, patola, songo-mongo,* ***toleirão****, totó, burgesso, morcão, microcéfalo, patinho, bacoco, babancas, inhenha, pàteta, néscio, matias, parvoinho,* ***mané****, anastácio, manembro, tatamba, bobalhão, bertoldo, patavina, tonto, apedeuto, pachocho, ingênuo, bocoió, simplacheirão, jerico, zote, sebastião, lorpa, atónito, patacão, pato,* ***parvoeirão****, ingénuo, papalvo, pateta, tanso, cretino, bolónio, basbaque, mentecapto, pachola, apaixonado, pasmão, pascácio, tarola,* ***trouxa****, parvo, jumento, geta, arara, gato-bravo, pedaço-de-asno,* ***parvajola****,* ***pacóvio****, laparoto, crendeiro, loura*

In the previous synsets, the words of the original TeP synsets are presented in bold. Other large synsets cover the concepts of a strong critic (100 words, including *ralho, ensinadela, descasca, raspanete, descompostura*), trickery (95 words, including *peta, embuste, manha, barrete, tramóia*), prostitute (73 words, including *pega, menina, mulher-da-vida, meretriz, quenga, rameira, ...*), a rascal/mischievous person (72 words, including, *pulha, traste, gandulo, salafrário, patife, tratante, ...*), and money (60 words, including *pastel, massa, grana, guita, carcanhol*). Also, on clustering, the only noun synset that includes more than 25 words refers to the concept of 'backside' or 'butt', and contains words such as *bufante*, *padaria* or *peida*. In TeP 2.0, the largest noun synset refers to a strike or aggression with some tool, and includes words as *paulada*, *bastonada*, *marretada* and *pancada*.

Furthermore, the largest verb synset in the final thesaurus means to mislead and contains words as *embromar*, *ludibriar*, *embaciar*, *enrolar*, *vigarizar*, or *intrujar*. The largest adjective synset denotes the quality of being shifty or deceitful and contains words as *artificioso*, *matreiro*, *ardiloso*, *traiçoeiro*, and *sagaz*.

6.5 Discussion

We have presented our work towards the enrichment of a thesaurus, structured in synsets, with synonymy information automatically acquired from general language dictionaries. The four-step enrichment approach resulted in TRIP, a large Portuguese thesaurus, obtained after enriching TeP, a Brazilian Portuguese thesaurus, with information extracted from three Portuguese dictionaries and a smaller Portuguese thesaurus. There are some similarities between the work presented here and the work of Tokunaga et al. (2001), for Japanese. However, our thesaurus is simpler, as it does not contain taxonomic information. Furthermore, although it was used for Portuguese, the proposed approach might be adapted to other languages.

Given that it is created using a handcrafted thesaurus as a starting point, the resulting thesaurus is more reliable than the thesaurus obtained in the previous chapter. The evaluation of the assignment procedure and of the obtained clusters also point that out, as they have shown higher precisions. Therefore, in the construction of Onto.PT, the four-step approach, in this chapter, was used instead of that described in the previous chapter, where synsets are discovered from scratch.

Another contribution of this part of the work is that TeP, originally made for Brazilian Portuguese, is enriched with words from dictionaries whose entries contain, mainly[3], words from European Portuguese. Therefore, besides being larger, the new thesaurus has a better coverage of European Portuguese than TeP. Also, once again due to its public domain character, the resulting thesaurus is another suitable alternative to replace OpenThesaurus.PT as the thesaurus of the OpenOffice word processor.

One limitation of the work presented here is the amount of observation labour required to select the best assignment settings. An alternative would be to develop a procedure to learn automatically the best measures and thresholds for associating a synpair to a synset. Given that we already have a small gold resource, a supervised learning approach, would suit this purpose. A simple linear classifier, such as a perceptron (Rosenblatt, 1958) would probably be enough to, given a set of labelled correct and incorrect examples for each assignment, learn the best threshold. This will be devised as future work. Also, in order to get more reliable results, the gold resource should as well be augmented. As it currently contains only nouns, in the future, especially special attention should be given to the inclusion of verbs and adjectives.

[3]Wiktionary.PT covers all variants of Portuguese, and PAPEL contains a minority of words in other variants of Portuguese, including Brazilian, Angolan and Mozambican.

Chapter 7

Moving from term-based to synset-based relations

Typical information extraction (IE) systems are capable of acquiring concept instances and information about these concepts from large collections of text. Whether these systems aim for the automatic acquisition of lexical-semantic relations (e.g. Chodorow et al. (1985); Hearst (1992); Pantel and Pennacchiotti (2006)), of knowledge on specific domains (e.g. Pustejovsky et al. (2002); Wiegand et al. (2012)), or the extraction of open-domain facts (e.g. Agichtein and Gravano (2000); Banko et al. (2007); Etzioni et al. (2011)) they typically represent concepts as terms, which are lexical items identified by their lemma. This is also how CARTÃO is structured. There, semantic relations are denoted by relational triples $t = \{a\ R\ b\}$, where the arguments (a and b) are terms whose meaning is connected by a relation described by R. As we have done throughout this thesis, we refer to the previous representation as term-based triples (tb-triples).

The problem is that a simple term is usually not enough to unambiguously refer to a concept, because the same word might have different meanings and different words might have the same meaning. On the one hand, this problem is not severe in the extraction of domain knowledge, where, based on the "one sense per discourse" assumption (Gale et al., 1992), ambiguity is low. On the other hand, when dealing with broad-coverage knowledge, if ambiguities are not handled, it becomes impractical to formalise the extracted information and to accomplish tasks such as inference for discovering new knowledge.

Therefore, to make IE systems more useful, a new step, which can be seen as a kind of WSD, is needed. Originally baptised as ontologising (Pantel, 2005), this step aims at moving from knowledge structured in terms, identified by their orthographical form, towards an ontological structure, organised in concepts, which is done by associating the terms to a representation of their meaning.

After the steps presented in the previous chapters, we are left with a lexical network, CARTÃO, with tb-triples extracted from text (chapter 4), and with a thesaurus, with synsets (chapter 5 and 6). While the synsets can be seen as concepts and their possible lexicalisations, the identification of the correct sense(s) of the arguments of a tb-triple for which the relation is valid is not straightforward. However, whereas most WSD techniques rely on the context where the words to be disambiguated occur to find their most adequate sense, the tb-triples do not provide their extraction context. While we could recover the context for some of

the tb-triples, DLP is proprietary, which means we cannot use the context of the tb-triples of PAPEL. Not to refer that there are several small definitions that do not provide enough context. Given this limitation, together with the need to map often un-matching (Dolan, 1994; Peters et al., 1998) word sense definitions in different resources, and to define extraction contexts for different heterogeneous resources, we decided to ontologise without using the extraction context. This enables the creation of IE systems with two completely independent modules: (i) one responsible for extracting tb-triples; and (ii) another for ontologising them. In other words, the second module attaches each term in a triple to a concept, represented, for instance, as a synset in a broad-coverage lexical ontology. We believe that this approach is an interesting way of coping with information sparsity, since it allows for the extraction of knowledge from different heterogeneous sources (e.g. dictionaries, encyclopedias, corpora), and provides a way to harmoniously integrate all the extracted information in a common knowledge base.

In this chapter, we propose several algorithms for moving from tb-triples to synset-based relational triples (hereafter, sb-triples), taking advantage of nothing but the existing synsets and a lexical network with tb-triples. We start by presenting the algorithms and then we describe how they were evaluated and compared. The performance results supported the choice of this kind of algorithm in the creation of Onto.PT. Also, given that the ontologising algorithms result in a set of synsets related among themselves by semantic relations, they are suitable for the last step of the ECO approach for creating wordnets. The core of this part of the work was originally reported in Gonçalo Oliveira and Gomes (2012a). Its earlier stages had been reported in Gonçalo Oliveira and Gomes (2011c).

7.1 Ontologising algorithms

Our work on ontologising semantic relations is similar to that presented by Pennacchiotti and Pantel (2006). The main difference is that the previous authors ontologise the semantic relations into WordNet, and exploit its structure, including synsets and existing synset-relations. We, on the other hand, had in mind to ontologise in a synset-base without synset-relations (TeP), so we had to find alternatives, such as exploring all the extracted information.

The goal of the proposed algorithms is to ontologise tb-triples, $\{a\ R\ b\}$, in the synsets of a thesaurus T. Instead of considering the context where the triples were extracted from, or the synset glosses, they exploit the information in a given lexical network N to select the best candidate synsets. A lexical network is established by a set of tb-triples, and is defined as a graph, $N = (V, E)$, with $|V|$ nodes and $|E|$ edges. Each node $v_i \in V$ represents a term, and each edge connecting v_i and v_j, $E(v_i, v_j)$, indicates that one of the meanings of the term in v_i is related to one meaning of the term in v_j. Furthermore, edges may be labelled according to the type of relationship held, $E(v_i, v_j, R)$.

By default, when a lexical network is needed, it is created from the tb-triples given as input. So, the proposed algorithms are better suited to ontologise large amounts of knowledge at once. Still, when there are few input tb-triples, they can exploit an external and larger lexical network or, eventually, the ontology where the triples are being attached to, if the former contains already ontologised triples.

Each algorithm can be seen as a different strategy for attaching terms a and b, in $\{a\ R\ b\}$, to suitable synsets $A_i \in T$ and $B_j \in T$, $A_i = \{a_{i0}, a_{i1}, ..., a_{in}\}$, $B_j = \{b_{j0}, b_{j1}, ..., b_{jm}\}$, where $n = |A_i|$ and $m = |B_j|$. This results in a sb-triple $\{A_i\ R\ B_j\}$. All algorithms, presented below, start by getting all the candidate synsets from the thesaurus, which are those containing term a, $A \in T : \forall(A_i \in A) \rightarrow a \in A_i$, and all with term b, $B \in T : \forall(B_j \in B) \rightarrow b \in B_j$. Also, for all of the proposed algorithms, if T does not contain the term argument of a tb-triple (e.g. a), a new synset containing only this term is created (e.g. $S_a = \{a\}$).

Before presenting the algorithms, we introduce figure 7.1, which contains candidate synsets for attaching terms a and b, as well as a made up lexical network N. There, nodes, identified by letters, can be seen as lexical items (terms), while the connections represent tb-triples of a labelled type (R1, R2 and R3). Note that N intentionally does not contain some lexical items in the synsets (k to p), which happens if they do not occur in any tb-triple. Both the synsets and the network of figure 7.1 will be used in the illustration of some of the algorithms. We intentionally created an example where, depending on the used algorithm, the resulting sb-triple is different.

Sample candidate synsets:

$A_1 = \{a, c, d, k\}$	$B_1 = \{b, g, h\}$
$A_2 = \{a, e, l\}$	$B_2 = \{b, f, o\}$
$A_3 = \{a, m, n\}$	$B_3 = \{b, p\}$
$A_4 = \{a, c, d, e, i, j\}$	

Sample network N:

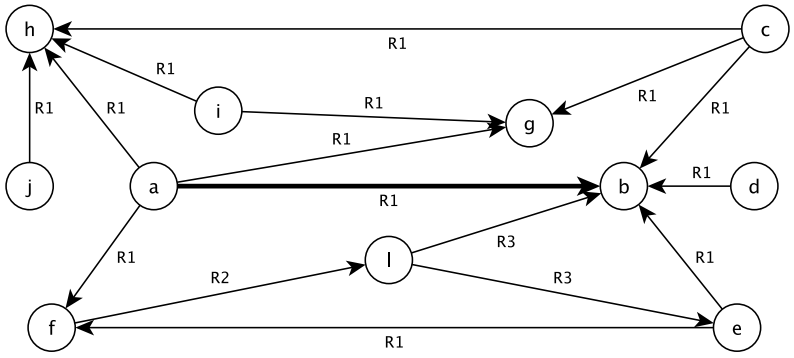


Figure 7.1: Candidate synsets and lexical network for the ontologising examples.

Related Proportion (RP): This algorithm is based on a similar assumption to Pennacchiotti and Pantel (2006)'s anchor approach. First, to attach term a, term b is fixed. For each synset $A_i \in A$, n_i is the number of terms $a_{ik} \in A_i$ such that the triple $\{a_{ik}\ R\ b\}$ holds. The related proportion rp is computed as follows:

$$rp(A_i, \{a, R, b\}) = \frac{n_i}{1 + log_2(|A_i|)} \tag{7.1}$$

All the candidate synsets with the highest rp are added to a new set, C. If $rp < \theta$, a predefined threshold, a is not attached. Otherwise, a is attached to the synset(s) of C with the highest n_i. Term b is attached using the same procedure, but fixing a.

The RP algorithm is illustrated in figure 7.2, where it is used to ontologise the tb-triple $\{a$ R1 $b\}$, given the candidate synsets and the network in figure 7.1[1].

$$\begin{array}{l|l} rp_{A1} = \mathbf{3/4^*} & rp_{B1} = \mathbf{3/3^*} \\ rp_{A2} = 2/3 & rp_{B2} = 2/3 \\ rp_{A3} = 1/3 & rp_{B3} = 1/2 \\ rp_{A4} = 4/6 & \end{array}$$

$$max(rp(A_i, \{a, R1, b\})) = 3/4 \rightarrow A_1$$
$$max(rp(B_i, \{a, R1, b\})) = 3/3 \rightarrow B_1$$
$$\textbf{resulting sb-triple} = \{A_1 \text{ R1 } B_1\}$$

Figure 7.2: Using RP to select the suitable synsets for ontologising $\{a$ R1 $b\}$, given the candidate synsets and the network N in figure 7.1.

Average Cosine (AC): Assuming that related concepts are described by words related to the same concepts, this algorithm exploits all the relations in N to select the most similar pair of candidate synsets. A term adjacency matrix $M(|V| \times |V|)$ is first created based on N, where $|V|$ is the number of nodes (terms). If the terms in indexes i and j are connected (related), $M_{ij} = 1$, otherwise, $M_{ij} = 0$.

In order to ontologise a and b, the most similar pair of synsets, $A_i \in A$ and $B_j \in B$, is selected according to the adjacencies of the terms they include. The similarity between A_i and B_j, represented by the adjacency vectors of their terms, $\vec{A}_i = \{\vec{a}_{i0}, ..., \vec{a}_{in}\}$, $n = |A_i|$ and $\vec{B}_j = \{\vec{b}_{j0}, ..., \vec{b}_{jm}\}$, $m = |B_j|$, is given by the average similarity of each term a_{ik} with each term b_{jl}, in N:

$$sim(A_i, B_j) = \frac{\sum_{k=1}^{|A_i|} \sum_{l=1}^{|B_j|} \cos(\vec{a}_{ik}, \vec{b}_{jl})}{|A_i||B_j|} \tag{7.2}$$

While this expression has been used to find similar nouns in a corpus (Caraballo, 1999), we adapted it to measure the similarity of two synsets, represented as the adjacency vectors of their terms.

The AC ontologising algorithm is illustrated in figure 7.3, where it is used to ontologise the tb-triple $\{a$ R1 $b\}$, given the sample candidate synsets and the sample network in figure 7.1. The example shows that, in opposition to the RP algorithm, AC uses all the relations in the network (R1, R2 and R3), and not just those of the same type of the tb-triple to ontologise (R1).

[1]For the sake of simplicity, we ignored the $1+log_2(|A_i|)$ in the denominator of the rp expression, and considered it to be just the size of the synset, $|A_i|$.

Adjacency matrix M, for the network in figure 7.1:

	$\vec{a}$	$\vec{b}$	$\vec{c}$	$\vec{d}$	$\vec{e}$	$\vec{f}$	$\vec{g}$	$\vec{h}$	$\vec{i}$	$\vec{j}$	$\vec{l}$
$\vec{a}$	1	1	0	0	0	1	1	1	0	0	0
$\vec{b}$	1	1	1	1	1	0	0	0	0	0	1
$\vec{c}$	0	1	1	0	0	0	1	1	0	0	0
$\vec{d}$	0	1	0	1	0	0	0	0	0	0	0
$\vec{e}$	0	1	0	0	1	1	0	0	0	0	1
$\vec{f}$	1	0	0	0	1	1	0	0	0	0	1
$\vec{g}$	1	0	1	0	0	0	1	0	1	0	0
$\vec{h}$	1	0	1	0	0	0	0	1	1	1	0
$\vec{i}$	0	0	0	0	0	0	1	1	1	0	0
$\vec{j}$	0	0	0	0	0	0	0	1	0	1	0
$\vec{l}$	0	1	0	0	1	1	0	0	0	0	1

Cosine values for each pair:

	B_1	B_2	B_3
A_1	$3.15/12 \approx 0.26$	$1.80/12 \approx 0.15$	$1.35/8 \approx 0.17$
A_2	$2.44/9 \approx 0.27$	$\mathbf{3.54/9 \approx 0.39}$ *	$1.59/6 \approx 0.26$
A_3	$1.21/9 \approx 0.13$	$0.81/9 \approx 0.09$	$0.37/6 \approx 0.06$
A_4	$5.48/18 \approx 0.30$	$3.16/18 \approx 0.18$	$1.96/12 \approx 0.16$

$max(sim(A_i, B_j)) \approx 0.39 \rightarrow$ **resulting sb-triple** $= \{A_2$ R1 $B_2\}$

Figure 7.3: Using AC to select the suitable synsets for ontologising $\{a$ R1 $b\}$, given the candidate synsets and the network N in figure 7.1.

Related Proportion + Average Cosine (RP+AC): This algorithm combines RP and AC. If RP cannot select a suitable synset for a or b, because one, or both, the selected synsets have $rp < \theta$, a selected threshold, AC is used.

Number of Triples (NT): Pairs of candidate synsets, $A_i \in A$ and $B_j \in B$, are scored according to the number of tb-triples of type R, present in N, between any of their terms. In other words, the pair that maximises $nt(A_i, B_j)$ is selected:

$$nt(A_i, B_j) = \frac{\sum_{k=1}^{|A_i|} \sum_{l=1}^{|B_j|} E(a_{ik}, b_{jl}, R) \in E}{log_2(|A_i||B_j|)} \tag{7.3}$$

As it is easier to find tb-triples between terms in larger synsets, this expression considers the size of synsets. However, in order to minimise the negative impact of very large synsets, a logarithm is applied to the multiplication of the synsets' size.

The NT ontologising algorithm is illustrated in figure 7.4, where it is used to ontologise the tb-triple $\{a$ R1 $b\}$, given the sample candidate synsets in figure 7.1 and the sample network in the same figure[2].

[2]For the sake of the clarity, we ignored the log_2 in the denominator of the $nt(A_i, B_j)$ expression, and considered it to be just $|A_i||B_j|$.

NT values for each pair:

	B_1	B_2	B_3
A_1	$7/12 \approx 0.58$	$4/12 \approx 0.33$	$3/8 \approx 0.38$
A_2	$4/9 \approx 0.44$	$4/9 \approx 0.44$	$2/6 \approx 0.44$
A_3	$3/9 \approx 0.33$	$2/9 \approx 0.22$	$1/6 \approx 0.17$
A_4	$\mathbf{11/18 \approx 0.61*}$	$6/18 \approx 0.33$	$4/12 \approx 0.33$

$max(nt(A_i, B_j)) \approx 0.61 \rightarrow$ **resulting sb-triple** $= \{A_4$ R1 $B_1\}$

Figure 7.4: Using NT to select the suitable synsets for ontologising $\{a$ R1 $b\}$, given the candidate synsets and the network N in figure 7.1.

Number of Triples + Average Cosine (NT+AC): This algorithm combines NT and AC. If NT cannot select a suitable pair $\{A_i, B_j\}$, or if the pair that maximises NT has $nt_{max}(A_i, B_j) < \theta$, where θ is a predefined threshold, AC is used instead.

PageRank (PR): The PageRank algorithm (Brin and Page, 1998) ranks the nodes of a graph according to their structural importance. Traditionally, the initial weights are uniformly distributed across all the nodes in the graph:

$$PR(v_i; t = 0) = \frac{1}{N} \tag{7.4}$$

At each iteration, PageRank is computed according to expression 7.5, where α is the so called dampening factor (typically 0.85), $In(v_k)$ is the number of edges to node v_k (in-edges) and $Out(v_i)$ is the number of edges from v_i (out-edges).

$$PR(v_k; t+1) = (1-\alpha) + \alpha \sum_{v_l \in In(v_k)} \frac{PR(v_l; t)}{Out(v_l)} \tag{7.5}$$

PageRank may biased according to the desired purpose, as in the Personalized PageRank WSD algorithm (Agirre and Soroa, 2009), for selecting the adequate wordnet synset for the occurrence of a word. In the previous algorithm, only the synsets with context words have initial weights.

As ontologising can be seen as a kind of WSD, the idea of this method is also to personalise PageRank for selecting the most adequate synsets for a and b, the arguments of the tb-triple. However, there are two main differences. First, we do not have a wordnet with relations between synsets. Second, our only context consists of a and b. Therefore, for ontologising a tb-triple, instead of synsets, we PageRank the terms in N, giving initial weights, of 0.5, only to a and b. Each synset in the thesaurus T is then scored with the average PageRank ($\overline{PR}$) of the terms it includes:

$$\overline{PR}(A_i) = \frac{\sum_{k=1}^{|A_i|} PR(a_{ik})}{1 + log_2(|A_i|)} \tag{7.6}$$

Finally, the pair of synsets (A_i, B_j), such that A_i and B_j maximise $\overline{PR}(A_i)$ and $\overline{PR}(B_j)$ respectively, is selected.

Minimum Distance (MD): This algorithm assumes that related synsets contain terms that are close in N. For this purpose, it selects the closest pair of synsets, given the average (edge-based) distance of their terms:

$$\overline{dist}(A_i, B_j) = \frac{\sum_{k=1}^{|A_i|} \sum_{l=1}^{|B_j|} dist(a_{ik}, b_{jl})}{|A_i||B_j|} \tag{7.7}$$

The minimum distance between two nodes is the number of edges in the shortest path between them, computed using Dijkstra's algorithm (Dijkstra, 1959). If a term in a synset (a_{ik} or b_{jl}) is not in N, they are removed from A_i and B_j before this calculation. If this algorithm was applied for attaching ontologise the tb-triple $\{a$ R1 $b\}$, given the situation of figure 7.1, there would be several ties for the best pair of synsets, because this network is simpler than most real networks.

7.2 Ontologising performance

For Portuguese, TeP is the only freely available lexical resource with synset-relations. However, these relations are limited to antonymy, which is not a very prototypical semantic relation. Therefore, in order to quantify the performance of the algorithms presented in the previous section, and to compare them for ontologising different relations, we have created a gold reference, manually, with a set of tb-triples extracted from dictionaries and their plausible attachments to the synsets of two handcrafted thesauri. Only after this, we used TeP as a gold resource and the algorithms for ontologising antonymy tb-triples.

This section starts by describing the resources involved in the creation of our handcrafted gold reference and reports on the results using each algorithm for ontologising hypernymy, part-of and purpose-of tb-triples. Then, we present the results of ontologising antonymy tb-triples in TeP.

7.2.1 Gold reference

The gold reference for this evaluation consisted of the synsets of TeP 2.0 and OpenThesaurus.PT, where samples of tb-triples from PAPEL 2.0 were attached.

Synsets

In order to eliminate the noise from automatic procedures, we decided to include only synsets from handcrafted thesauri in our gold reference. As referred in the previous chapters, for Portuguese, there are currently two free handcrafted broad-coverage thesauri: TeP 2.0 (Maziero et al., 2008) and OpenThesarus.PT (OT.PT). TeP is the largest by far (see section 3.1.2) and is created by experts, so its synsets were the best alternative for our gold reference. However, TeP was created for

Brazilian Portuguese and thus contains some unusual words or meanings for European Portuguese. On the other hand, OT.PT is smaller, but made for European Portuguese, and contains words and meanings not covered by TeP.

Therefore, we used TeP as a starting point for the creation of a new noun thesaurus[3], TePOT, with the noun synsets from both TeP and OT.PT. The thesauri are merged according to the following automatic procedure:

1. The overlap between each synset in OT.PT, O_i, and each synset of TeP, T_j, is measured. For each $O_i \in$ OT.PT a first set of candidates, $C_i = \{C_{i1}, C_{i2}, ..., C_{in}\} \subset$ TeP, will contain the TeP synsets that maximise the Overlap measure, $Overlap(O_i, C_{ik}) = max(Overlap(O_i, T_j))$:

$$Overlap(O_i, T_j) = \frac{O_i \cap T_j}{min(|O_i|, |T_j|)}$$

   If $max(Overlap(O_i, T_j)) = 0$, it means that the OT.PT synset contains only words that are not in TeP, and is thus added to TePOT as it is.

2. Otherwise, the candidate(s) in C_i with higher Jaccard coefficient are selected, $C_{il} \in C'_i \rightarrow Jaccard(O_i, C_{il}) = max(Jaccard(O_i, C_{ik}))$:

$$Jaccard(O_i, C_{ik}) = \frac{O_i \cap C_{ik}}{O_i \cup C_{ik}}$$

   Usually, C'_i has just one synset but, if it has more than one, they are merged in the same synset. Then, the new synset is merged with O_i. A new TePOT synset S_i will contain all words in O_i and in the synsets in C'_i. $S_i = \{w_1, w_2, ..., w_m\} : \forall(w_j \in S_i) \rightarrow w_j \in O_i \vee w_j \in C_{il}, C_{il} \in C'_i$.

3. Synsets of TeP which have not been merged with any OT.PT synset are finally added to TePOT without any change.

In the end, TePOT contains 18,501 nouns, organised in 8,293 synsets – 6,237 of the nouns are ambiguous and, on average, one synset has 3.84 terms and one term is in 1.72 synsets.

Tb-triples

The algorithms were evaluated for ontologising tb-triples of three different types: hypernymy, part-of and purpose-of, all held between nouns. The tb-triples used were obtained from PAPEL 2.0, which was, at the time when we started to create the gold reference, the most recent version of PAPEL. As a resource extracted automatically from dictionaries, the reliability of PAPEL is not 100% (see section 4.2.5 for evaluation details), but it was the largest lexical-semantic resource of this kind freely available. In order to minimise the noise of using incorrect tb-triples, we added additional constraints to their selection, namely:

[3]We only used nouns because the reported experimentations only dealt with semantic relations between nouns, namely hypernymy, part-of, and purpose-of.

- Only tb-triples supported by CETEMPúblico (Santos and Rocha, 2001), a newspaper corpus of Portuguese, were used. This was done based on the results of the automatic validation, as reported in section 4.2.5. We thus had some confidence on the quality of the triples, as their arguments co-occurred at least once in the corpus, connected by discriminating textual patterns for their relation.
- Triples with the following frequent but abstract arguments were discarded: *acto* (act), *efeito* (effect), *acção* (action), *estado* (state), *coisa* (thing), *qualidade* (quality) as well as tb-triples with arguments with less than 25 occurrences in CETEMPúblico. Some of the frequent and abstract arguments were actually considered as "empty heads" (see more on section 3.2.1 of this thesis) since PAPEL 3.0. This means that, in the current version of PAPEL, there are not hypernymy tb-triples where these words are the hypernym.

Furthermore, we unified all meronymy relations (part-of, member-of, contained-in, material-of) in a unique type, part-of. This option relied on the fact that the distinction of different meronymy subtypes is sometimes too fine-grained, and because, as it occurs for English (Ittoo and Bouma, 2010), for Portuguese there are textual patterns that might be used to denote more that one subtype.

Attachments

From the previous selection of tb-triples, we chose those held between words included in, at least, one TePOT synset, and whose attachment raised no doubts. It was possible to have tb-triples where all possible attachments were correct, as well as tb-triples without a plausible attachment, because the sense of one of the arguments was not covered by the thesaurus.

For each tb-triple, the gold reference contained all plausible attachments, as in the examples of figure 7.5. In the end, the gold reference consisted of 452 tb-triples and their possible attachments, with those that were plausible marked. Table 7.1 shows the distribution of tb-triples per relation type, the average number of possible attachments, and the average number of plausible attachments. The proportion of plausible attachments per tb-triple can be seen as the random chance of selecting a plausible attachment from the possible alternatives. This number is between 40%, for hypernymy, and 50% for purpose-of.

Relation	**tb-triples**	**Attachments**		
		Avg(possible)	**Avg(plausible)**	
Hypernym-of	210	13.7	5.5	(40.2%)
Part-of	175	11.2	5.5	(49.5%)
Purpose-of	67	13.5	6.8	(50.1%)

Table 7.1: Matching possibilities in the gold resource.

7.2.2 Performance comparison

In order to compare the performance of the algorithms, we used them to ontologise the 452 tb-triples in the gold reference into the candidate synsets. However, instead

tb-triple = (*documento* **hypernym-of** *recibo*) (*document* hypernym-of *receipt*)	
A_1: *documento, declaração* A_2: *escritura, documento*	B_1: *recibo, comprovante, nota, quitação, senha*
plausible sb-triples = $\{A_1, B_1\}$	

tb-triple = (*planta* **part-of** *floresta*) (*plant* part-of *forest*)	
A_1: *relação, quadro, planta, mapa* A_2: *vegetal, planta* A_3: *traçado, desenho, projeto, planta, plano*	B_1: *bosque, floresta, mata, brenha, selva*
plausible sb-triples = $\{A_2, B_1\}$	

tb-triple = (*passageiro* **purpose-of** *carruagem*) (*passenger* purpose-of *carriage*)	
A_1: *passageiro, viajante* A_2: *passageiro, viador* A_3: *passageiro, transeunte*	B_1: *carriagem, carruagem, carraria* B_2: *carruagem, carro, sege, coche* B_3: *carruagem, caleça, caleche* B_4: *actividade, carruagem, operosidade, diligência*
plausible sb-triples = $\{A_1, B_1\}$, $\{A_1, B_2\}$, $\{A_1, B_3\}$, $\{A_2, B_1\}$, $\{A_2, B_2\}$, $\{A_2, B_3\}$	

tb-triple = (*máquina* **hypernym-of** *câmara*) (*machine* hypernym-of *camera*)	
A_1: *motor, máquina*	B_1: *câmara, parlamento, assembleia, assembléia* B_2: *quarto, repartimento, apartamento, câmara, compartimento, aposento, recâmara, alcova*
plausible sb-triples = {}	

Figure 7.5: Example of gold entries.

of using only the 452 tb-triples as a lexical network, we used all the tb-triples in CARTÃO (see section 4). After comparing the automatic attachments with the attachments in the gold reference, we computed typical information retrieval measures, including precision, recall and three variations of the F-score: F_1 is the classic, $F_{0.5}$ favors precision, and RF_1 uses a relaxed recall ($RelRecall$), instead of the classic recall – $RelRecall$ is 1 if at least one correct attachment is selected. For a tb-triple in the set of tb-triples to ontologise, $t_i \in T$, these measures are computed as follows:

$$Precision_i = \frac{|AutomaticAttachments_i \cap GoldAttachments_i|}{|AutomaticAttachments_i|} \qquad Precision = \frac{1}{|T|}\sum_{i=1}^{|T|} Precision_i$$

$$Recall_i = \frac{|AutomaticAttachments_i \cap GoldAttachments_i|}{|GoldAttachments_i|} \qquad Recall = \frac{1}{|T|}\sum_{i=1}^{|T|} Recall_i$$

$$RelRecall_i = \begin{cases} 1, & \text{if } |AutomaticAttachments_i \cap GoldAttachments_i| > 0 \\ 0, & \text{otherwise} \end{cases}$$

$$RelRecall = \frac{1}{|T|}\sum_{i=1}^{|T|} RelRecall_i \qquad F_\beta = (1+\beta^2) \times \left(\frac{Precision \times Recall}{(\beta^2 \times Precision) + Recall}\right)$$

Table 7.2 presents the scores obtained for each measure, according to algorithm and relation type. In RP and RP+AC, the threshold θ was empirically set to 0 and 0.55, respectively. In NT+AC, θ was set to 3.

Algorithm	**Hypernym-of** (210 tb-triples)				
	Precision (%)	**Recall (%)**	$\mathbf{F}_1(\%)$	$\mathbf{F}_{0.5}(\%)$	$\mathbf{RF}_1(\%)$
RP	53.8	12.4	20.2	32.3	50.3
AC	**60.1**	15.7	24.9	38.4	**59.8**
RP+AC	58.5	15.6	24.6	37.7	58.5
NT	57.7	**17.3**	**26.6**	**39.4**	57.7
NT+AC	58.7	15.3	24.3	37.4	58.6
PR	46.2	11.5	18.5	28.9	45.7
MD	58.6	15.8	24.9	38.0	58.6
	Part-of (175 tb-triples)				
	Precision (%)	**Recall (%)**	$\mathbf{F}_1(\%)$	$\mathbf{F}_{0.5}(\%)$	$\mathbf{RF}_1(\%)$
RP	56.9	10.6	17.9	30.4	47.0
AC	58.7	14.9	23.8	37.0	58.7
RP+AC	**64.1**	**16.6**	**26.3**	**40.7**	**64.1**
NT	50.7	15.8	24.1	35.2	50.7
NT+AC	59.2	15.2	24.2	37.5	59.2
PR	50.6	12.6	20.2	31.6	49.9
MD	59.1	15.3	24.4	37.6	59.1
	Purpose-of (67 tb-triples)				
	Precision (%)	**Recall (%)**	$\mathbf{F}_1(\%)$	$\mathbf{F}_{0.5}(\%)$	$\mathbf{RF}_1(\%)$
RP	51.5	5.1	9.3	18.3	32.6
AC	63.2	13.0	21.5	35.6	63.2
RP+AC	**63.4**	13.6	22.3	36.5	**63.4**
NT	48.1	**15.4**	**23.3**	33.7	48.1
NT+AC	62.2	13.9	22.7	**36.6**	62.2
PR	56.3	10.8	18.2	30.6	56.3
MD	60.9	12.7	20.9	34.5	60.9

Table 7.2: Ontologising algorithms performance results.

The comparison shows that the best performing algorithms for hypernymy are AC and NT, which have close F_1 and RF_1. NT is more likely to originate ties for the best attachments than AC, and thus to have higher recall. However, its precision is lower than AC's. For part-of, RP+AC is clearly the best performing algorithm. For purpose-of, RP+AC has also the best precision and RF_1, but its scores are very close to AC. Moreover, it is outperformed by NT and NT+AC in the other measures. Once again, NT has higher recall and thus higher F_1. NT+AC combines good precision and recall in an interesting way and has therefore the best $F_{0.5}$. However, as that the set of purpose-of tb-triples contains only 67 instances, the results for this relation might not be significant enough to take strong conclusions.

These results show as well that PR has the worst performance for hypernymy and part-of tb-triples, which suggests that PageRank is not adequate for this task. For purpose-of, RP is the worst algorithm, especially due to the low recall.

7.2.3 Performance against an existing gold standard

In the second performance evaluation, we used the proposed algorithms to ontologise antonymy relations. For this purpose, the antonymy sb-triples of TeP were converted to tb-triples. This resulted in 46,339 unique antonymy pairs – 7,633 between nouns,

25,951 between verbs, 12,279 between adjectives, and 476 between adverbs. From those, four random samples were created, respectively with 800 noun antonymy tb-triples, 800 verb tb-triples, 800 adjective tb-triples and 476 adverb tb-triples.

The proposed algorithms were used, with the same parameters as above, to ontologise the tb-triples of the random samples into the synsets of TeP, which was our gold resource. Table 7.3 shows the distribution of antonymy tb-triples of the samples per POS of the arguments, the average number of possible attachments, and the average number of attachments in TeP (correct). The proportion of correct attachments per tb-triple can be seen as the random chance of selecting a correct attachment from the possible alternatives. This number is between 41.6%, for nouns, and 78.9% for adverbs. In order to measure the performance, we compared the resulting attachments with the real sb-triples in TeP, and computed the same measures as in the previous comparison: precision, recall, F_1, $F_{0.5}$ and RF_1.

Relation	**tb-triples**	**Attachments**	
		Avg(possible)	**Avg(correct)**
Nouns	800	6.1	2.5 (41.6%)
Verbs	800	5.7	2.5 (43.5%)
Adjectives	800	3.2	1.8 (57.1%)
Adverbs	476	1.6	1.3 (78.9%)

Table 7.3: Matching possibilities in the gold collection for antonymy.

Ontologisation was however done using different lexical networks N:

1. First, for each POS of the arguments, we used all the tb-triples of that kind as lexical network N. For instance, for antonymy between nouns, N contained all the antonymy tb-triples from TeP between nouns and nothing else. These results are presented in table 7.4.
2. The previous scenario can be seen as optimal and highly unlikely, because N is a direct transformation of the information in TeP, which can thus be seen as complete. Therefore, in order to simulate more realistic scenarios, we made additional runs where we used only one half, one fourth and one eighth of the original N, with tb-triples selected randomly. These results are shown in tables 7.5, 7.6, 7.7 and 7.8, respectively for antonymy tb-triples between nouns, verbs, adjectives and adverbs.
3. In the last run, instead of using the tb-triples from TeP, we used the tb-triples of CARTÃO. As this resource only contained antonymy between adjectives, most algorithms would not work. So, we only did this evaluation for antonymy between adjectives. These results are presented in table 7.9.

Table 7.4 shows the effectiveness of RP, AC, NT as NT+AC, when the lexical network is complete. By complete, we mean that the conditions can be seen as optimal, because all the possible tb-triples resulting from the sb-triples are present. In this scenario, MD has always the highest recall but, for adverbs, this leads to a precision below the random chance. All the other algorithms have precisions above the random chance. RP is the best performing algorithm with almost 100% precision and RF_1. AC, RP+AC, NT, and NT+AC also perform well, with precisions and

Algorithm	**Nouns** (800 tb-triples)				
	Precision (%)	**Recall (%)**	**F**$_1$(%)	**F**$_{0.5}$(%)	**RF**$_1$(%)
RP	**99.8**	82.0	**90.0**	**95.6**	**99.8**
AC	95.4	78.7	86.3	91.5	95.4
NT	96.1	76.1	84.9	91.3	96.1
NT+AC	96.5	75.9	85.0	91.5	96.5
PR	56.3	53.4	54.9	55.7	56.3
MD	52.8	**85.2**	65.2	57.1	52.8
	Verbs (800 tb-triples)				
	Precision (%)	**Recall (%)**	**F**$_1$(%)	**F**$_{0.5}$(%)	**RF**$_1$(%)
RP	**99.7**	84.5	**91.5**	**96.2**	**99.7**
AC	92.3	83.0	87.4	90.3	92.3
NT	95.2	80.1	87.0	91.7	95.2
NT+AC	95.2	80.1	87.0	91.7	95.2
PR	52.0	56.9	54.3	52.9	52.0
MD	69.6	**87.1**	77.4	72.5	69.6
	Adjectives (800 tb-triples)				
	Precision (%)	**Recall (%)**	**F**$_1$(%)	**F**$_{0.5}$(%)	**RF**$_1$(%)
RP	**100.0**	94.8	**97.3**	**98.9**	**100.0**
AC	95.2	93.0	94.1	94.7	95.2
NT	96.1	91.3	93.6	95.1	96.1
NT+AC	96.3	91.3	93.7	95.3	96.3
PR	70.6	73.6	72.1	71.2	70.6
MD	61.6	**96.8**	75.3	66.5	61.6
	Adverbs (476 tb-triples)				
	Precision (%)	**Recall (%)**	**F**$_1$(%)	**F**$_{0.5}$(%)	**RF**$_1$(%)
RP	**100.0**	90.2	94.9	**97.9**	**100.0**
AC	99.2	92.5	**95.7**	97.8	99.2
NT	96.8	91.7	94.2	95.8	96.8
NT+AC	98.3	89.5	93.7	96.4	98.3
PR	92.4	91.7	92.1	92.3	92.4
MD	64.8	**95.8**	77.2	69.2	64.8

Table 7.4: Results of ontologising samples of antonymy tb-triples of TeP in TeP, using all TeP's antonymy relations as a lexical network N.

RF_1 always higher than 90%, and F_1 always higher than 84%. This confirms our initial intuition on using these algorithms.

However, it is not expected to extract a complete lexical network from dictionaries, and even less from other kinds of text. Even though dictionaries have an extensive list of words, senses and (implicitly) relations, they can hardly cover all possible tb-triples resulting from a sb-triple, especially for large synsets.

In tables 7.5, 7.6, 7.7 and 7.8 we present the results of a more realistic scenario, because we only use part of TeP's lexical network, respectively 50%, 25% and 12.5%. As expected, performance decreases for smaller lexical networks, but it decreases more significantly for some algorithms than for others. For instance, RP's performance decreases drastically, and it has never the best F_1 nor RF_1, as it had when using all the network. On the other hand, RP+AC is the algorithm that performs better with missing tb-triples. Besides having the best precision in most scenarios, this algorithm has always the best F_1, $F_{0.5}$ and RF_1. There are also some situations where AC, NT and NT+AC have a very close performance to RP+AC.

The main difference between the previous kind of scenario and a real scenario is that the part of the lexical network used was selected randomly, which tends to

% of N	Algorithm	Precision (%)	Recall (%)	$\mathbf{F}_1$(%)	$\mathbf{F}_{0.5}$(%)	$\mathbf{RF}_1$(%)
50%	RP	89.1	61.5	72.8	81.7	85.7
	AC	93.4	78.0	85.0	**89.9**	93.4
	RP+AC	93.1	**78.4**	**85.1**	**89.9**	93.1
	NT	89.2	72.7	80.1	85.4	89.2
	NT+AC	**94.1**	76.3	84.3	**89.9**	**94.1**
	PR	57.3	55.0	56.1	56.9	57.3
	MD	64.6	69.6	67.0	65.5	64.6
25%	RP	**88.5**	40.9	55.9	71.8	71.1
	AC	82.8	70.4	76.1	80.0	82.8
	RP+AC	84.1	72.5	**77.9**	**81.5**	**84.1**
	NT	82.0	65.6	72.9	78.1	82.0
	NT+AC	77.6	**73.2**	75.3	76.7	77.6
	PR	55.7	54.0	54.8	55.3	55.7
	MD	58.5	67.7	62.7	60.1	58.5
12.5%	RP	**87.6**	29.2	43.8	62.6	58.9
	AC	76.8	66.0	71.0	74.4	76.8
	RP+AC	79.0	70.3	**74.4**	**77.1**	**79.0**
	NT	74.8	56.2	64.2	70.2	74.8
	NT+AC	63.4	**73.0**	67.9	65.1	63.4
	PR	53.3	53.4	53.4	53.4	53.3
	MD	49.0	64.6	55.7	51.5	49.9

Table 7.5: Results of ontologising 800 antonymy tb-triples, between nouns, of TeP in TeP, using only part of the TeP's antonymy relations as a lexical network.

% of N	Algorithm	Precision (%)	Recall (%)	$\mathbf{F}_1$(%)	$\mathbf{F}_{0.5}$(%)	$\mathbf{RF}_1$(%)
50%	RP	95.2	76.3	84.7	90.7	95.2
	AC	92.7	82.3	87.2	90.4	92.7
	RP+AC	**96.6**	**82.8**	**89.2**	**93.5**	**96.6**
	NT	93.5	79.2	85.8	90.2	93.5
	NT+AC	94.9	80.1	86.9	91.5	94.9
	PR	51.3	56.5	53.8	52.3	51.3
	MD	79.1	78.0	78.6	78.9	79.1
25%	RP	93.8	61.0	73.9	84.7	89.6
	AC	93.5	**82.3**	87.5	91.0	93.5
	RP+AC	**94.5**	**82.3**	**88.0**	**91.8**	**94.5**
	NT	91.2	77.0	83.5	88.0	91.2
	NT+AC	94.2	80.9	87.0	91.2	94.2
	PR	51.4	56.5	54.8	52.4	51.4
	MD	75.6	76.3	75.6	75.7	75.6
12.5%	RP	**93.0**	47.6	63.0	78.1	82.7
	AC	88.6	78.0	83.0	86.2	88.6
	RP+AC	89.9	79.2	**84.2**	**87.6**	**89.9**
	NT	87.5	71.8	78.8	83.8	87.5
	NT+AC	88.0	79.2	83.4	86.1	88.0
	PR	51.3	55.7	53.4	52.1	51.3
	MD	70.2	**74.9**	72.4	71.1	70.2

Table 7.6: Results of ontologising 800 antonymy tb-triples, between verbs, of TeP in TeP, using only part of the TeP's antonymy relations as a lexical network.

result in uniformly distributed missing tb-triples. What happens when extracting information from text is that some parts of the network might be almost complete, while other parts, possibly those with less frequent words and relations, will be almost incomplete.

% of N	Algorithm	Precision (%)	Recall (%)	$\mathbf{F}_1$(%)	$\mathbf{F}_{0.5}$(%)	$\mathbf{RF}_1$(%)
50%	RP	90.0	82.6	86.1	88.4	90.0
	AC	93.3	90.0	91.6	92.6	93.3
	RP+AC	**96.6**	**92.8**	**94.8**	**95.8**	**96.6**
	NT	94.0	86.1	89.9	92.3	94.0
	NT+AC	95.9	88.1	91.8	94.2	95.9
	PR	70.6	73.4	72.0	71.1	70.6
	MD	73.1	85.3	78.8	75.2	73.1
25%	RP	87.0	61.4	72.0	80.3	82.2
	AC	90.6	84.1	87.2	89.2	90.6
	RP+AC	**93.5**	**89.3**	**91.3**	**92.6**	**93.5**
	NT	91.3	78.4	84.3	88-4	91.3
	NT+AC	92.4	84.3	88.2	90.6	92.4
	PR	70.6	73.6	72.1	71.2	70.6
	MD	70.2	85.6	77.1	72.8	70.2
12.5%	RP	85.1	49.8	62.8	74.5	75.0
	AC	86.0	74.9	80.1	83.5	86.0
	RP+AC	**88.4**	**85.6**	**87.0**	**88.4**	**88.4**
	NT	87.3	66.7	75.6	82.2	85.9
	NT+AC	84.5	82.6	83.5	84.1	84.5
	PR	67.0	71.1	69.0	67.7	67.0
	MD	63.8	80.3	71.1	66.6	63.8

Table 7.7: Results of ontologising 800 antonymy tb-triples, between adjectives, of TeP in TeP, using only part of the TeP's antonymy relations as a lexical network.

% of N	Algorithm	Precision (%)	Recall (%)	$\mathbf{F}_1$(%)	$\mathbf{F}_{0.5}$(%)	$\mathbf{RF}_1$(%)
50%	RP	94.2	85.7	89.8	92.3	94.2
	AC	97.3	80.5	88.1	93.4	96.5
	RP+AC	**97.5**	87.2	**92.1**	**95.2**	**97.5**
	NT	95.1	73.7	83.1	89.9	93.1
	NT+AC	93.9	81.2	87.1	91.1	93.9
	PR	91.0	**91.0**	91.0	91.0	91.0
	MD	70.6	90.2	79.2	73.8	70.6
25%	RP	93.8	78.9	85.7	90.4	93.7
	AC	94.9	69.9	80.5	88.6	90.9
	RP+AC	**95.9**	88.7	**92.2**	**94.4**	**95.9**
	NT	93.2	51.9	66.7	80.4	83.1
	NT+AC	83.1	81.2	82.1	82.7	83.1
	PR	87.5	**89.5**	88.5	87.9	87.5
	MD	71.7	89.5	79.6	74.7	71.7
12.5%	RP	96.0	72.9	82.9	90.3	93.3
	AC	94.3	49.6	65.0	79.9	81.4
	RP+AC	**96.7**	85.7	**90.8**	**94.2**	**96.1**
	NT	90.5	28.6	43.4	63.1	69.9
	NT+AC	73.0	81.2	76.9	74.5	73.0
	PR	85.2	**91.0**	88.0	86.3	85.2
	MD	76.8	87.2	81.7	78.7	76.8

Table 7.8: Results of ontologising 476 antonymy tb-triples, between adverbs, of TeP in TeP, using only part of the TeP's antonymy relations as a lexical network.

Table 7.9 shows the result of what can be seen as a real scenario, because the lexical network used, CARTÃO, was extracted automatically from a different source than the synsets. In this run, RP is the most precise algorithm in a trade-off for lower recall, because it only uses information of relations of the same type of the tb-triple.

Algorithm	**Adjectives** (800 tb-triples)				
	Precision (%)	**Recall (%)**	$\mathbf{F}_1$(%)	$\mathbf{F}_{0.5}$(%)	$\mathbf{RF}_1$(%)
RP	**99.4**	40.8	57.8	**77.2**	**87.0**
AC	62.0	50.2	55.5	59.2	62.0
RP+AC	69.3	69.7	**69.5**	69.4	69.3
NT	70.2	21.1	32.5	48.0	50.9
NT+AC	51.6	77.9	62.0	55.3	51.6
PR	50.5	61.4	55.4	52.4	50.5
MD	60.0	**74.9**	66.6	62.4	60.0

Table 7.9: Results of ontologising 800 antonymy tb-triples, between adjectives, of TeP in TeP, using all CARTÃO as an external lexical network N.

When there is not enough information, the tb-triple is simply not ontologised. The algorithms AC and, especially, RP+AC, have more balanced results in all measures, and are more in agreement with the evaluation of the other types of relation, in section 7.2.2. This happens because, although the network is incomplete, AC uses other relations to compensate the lack of information on relations of the same type. We can say that AC is more tolerant to missing information.

We should finally remind that the antonymy relation is not very prototypical because it connects concepts with an opposite meaning. Antonyms are similar in all contextual properties but one (see more about antonymy in Murphy (2003)). Furthermore, the previous evaluation showed that the best algorithm was dependent on the semantic relation. Therefore, we cannot make a blind extrapolation of these results to a real scenario, where there are other types of relations, and the lexical network is frequently incomplete.

7.3 Discussion

This chapter presented several algorithms for ontologising tb-triples, which attach related term arguments to suitable synsets in a wordnet-like resource. In our case, this resource does not contain any explicit semantic relations between synsets. Therefore, the proposed algorithms use nothing besides the synsets themselves and a set of provided tb-triples, which establish a lexical network. If the lexical network contains all the extracted information, which is what happens in the creation of Onto.PT, the former can be seen as the extraction context.

We have obtained interesting results but, as other authors referred (e.g. Pennacchiotti and Pantel (2006)), we have confirmed that ontologising is a challenging task. Other alternatives should be devised, so we leave some ideas for further experiments:

- To ontologise, manually, a representative set of tb-triples (the gold reference might be a starting point) or, automatically, the tb-triples in which confidence is very high. Then, exploit the resulting sb-triples to ontologise the remaining tb-triples in a new algorithm, or learn a classifier for this task automatically;
- We did not conclude if there was any especially noisy (or useful) type of relation in the application of the cosine similarity, so we have used them all. More experiments should be done to analyse this deeper;
- Investigate if we can take advantage of the occurrence of tb-triples in textual corpora to increase or decrease the confidence of an attachment;

- The gold reference should be enlarged, not just in terms of tb-triples, but also in terms of covered relations, in order to have more significant results.

Nevertheless, when attaching hypernymy, part-of and purpose-of tb-triples, all algorithms precisions outperform the random chance. Given the obtained results, in the creation of Onto.PT, we decided to use AC for ontologising the hypernymy tb-triples (almost one third of CARTÃO) and RP+AC for ontologising the rest of the tb-triples. These choices are mainly supported by the precision and *RF* measures.

Experimentation using the antonymy relations of TeP, a handcrafted gold resource, were also performed. Although this resource only contains sb-triples, these experimentations confirmed that RP+AC is the best performing algorithm. Only when the lexical network is complete, RP seems to perform better. But this is not the typical case, and it is definitely not the case in the creation of Onto.PT, because the lexical network is automatically extracted from dictionaries and thus larger and incomplete.

In the next chapter, we present the current version of Onto.PT. It results from the application of the automatic step described here for ontologising the tb-triples extracted from dictionaries (CARTÃO network, see chapter 4), in the enriched TeP (TRIP thesaurus, see chapter 6).

Chapter 8

Onto.PT: a lexical ontology for Portuguese

In the previous chapters, we have presented individual automatic steps towards the acquisition and integration of lexical-semantic knowledge in a LKB. Each step is implemented by a module and, if combined with the others, as described by the diagram in figure 8.1, results in the three step approach we propose to the automatic construction of a wordnet-like resource. This approach was named ECO, which stands for Extraction, Clustering and Ontologisation. Briefly, the ECO approach starts by extracting instances of semantic relations, represented as tb-triples, from textual sources. Then, synsets are discovered from the extracted synonymy tb-triples (synpairs). If there is a an available synset-based thesaurus, its synsets are first augmented, and new synsets are only discovered from the remaining synpairs. Finally, the term arguments of the non-synonymy tb-triples are ontologised, which means that they are attached to the synsets, in the thesaurus, that transmit suitable meanings and make the tb-triple true. This results in a wordnet, where synsets are connected by semantic relations (sb-triples).

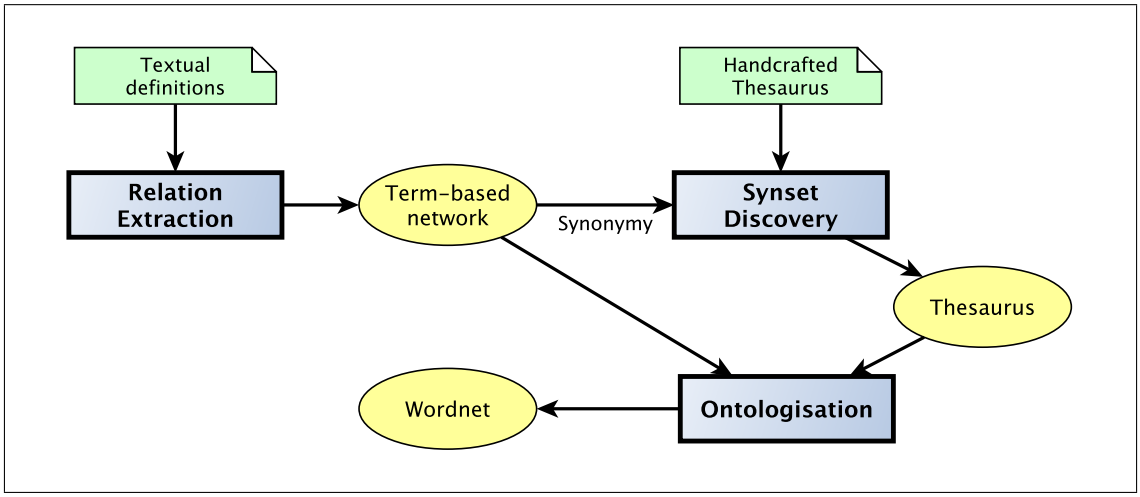


Figure 8.1: Diagram of the ECO approach for creating wordnets from text.

Each module of ECO is completely independent of the others and can be used alone, in order to achieve its specific task. For instance, given a set of synpairs, an existing thesaurus may be enriched automatically and have its original synsets

augmented; or given an existing wordnet, the ontologisation module can be used to integrate new relations, obtained from other sources.

This flexible approach was applied for Portuguese and resulted in a resource named Onto.PT. Therefore, Onto.PT is not a static resource. It is in constant development and may have different instantiations, depending on the resources used in its creation, the version of the modules, and other parameters specific to each module. In this chapter, we present the most recent version of Onto.PT, which was available during the writing of this thesis. This version (v.0.35) integrates five lexical resources of Portuguese, earlier presented, in section 3.1.2:

- The proprietary dictionary Dicionário PRO da Língua Portuguesa (DLP, 2005), indirectly, through PAPEL.
- The public domain dictionary Dicionário Aberto (Simões and Farinha, 2011).
- The collaborative dictionary Wiktionary.PT[1], 19th October 2011 dump.
- The public domain thesaurus TeP 2.0 (Maziero et al., 2008).
- The collaborative thesaurus OpenThesaurus.PT[2].

We first provide an overview of Onto.PT, which presents the number of synsets and semantic relations included. As the underlying lexical network is slightly different from the one presented in section 4, we start the characterisation by updating those numbers. After the overview, we add information on the main alternatives there are for using and exploring Onto.PT. Then, we discuss some aspects on the evaluation and quality of this resource. We end up with a section dedicated to some of the tasks where Onto.PT can be used.

8.1 Overview

This section presents an overview of Onto.PT v.0.35, the current version of this resource, obtained after improvements on the extraction step, as well as the addition of new antonymy relations. Before presenting Onto.PT itself, we refer the most relevant improvements and update the numbers on the underlying lexical network, which we decided to call CARTÃO 3.1.

Then, we present the number of synsets. Besides the (augmented) synsets of TeP, and the newly discovered synsets, Onto.PT contains a majority of single-item synsets, resulting from the arguments of tb-triples not covered by the existing synsets. After this, we present the types of semantic relations in Onto.PT, which are the same as in CARTÃO, together with the quantities of its sb-triples. Finally, we provide an example of a sb-triple of each of the covered relation types.

We should add that, after the three ECO steps, the synsets of Onto.PT are ordered according to the average frequency of their lexical items, using the frequency lists of AC/DC [3]. Inside each synset, lexical items are also ordered according to their frequency. Although quite different, this can be seen as a rough approximation to the order of the synsets in Princeton WordNet. In the previous resource, the lexical

[1] Available from `http://pt.wiktionary.org/` (September 2012)
[2] Available from `http://openthesaurus.caixamagica.pt/` (September 2012)
[3] Available from `http://www.linguateca.pt/ACDC/` → `Frequência` (September 2012)

items inside a synset are ordered according to the frequency each one of them is used to denote the sense corresponding to the meaning of the synset. This information is based on the annotations of SemCor (Miller et al., 1994), a sense annotated corpus.

8.1.1 Underlying lexical network

After the manual evaluation of the semantic relation extraction (see section 4.2.5), we identified a few problems in this step. Besides minor changes in the grammars, some lemmatisation rules were refined and some filters were added to avoid relations between lexical items such as:

- *cf*, used several times in the middle of DA definitions for introducing bibliographic references;
- *transitivo*, *intransitivo*, *reflexivo*, and other verb classifying words, incorrectly extracted from Wiktionary as synonyms of verbs;
- synonymy between verbs in the gerund, often incorrect because the verb in the gerund refers to an action that specifies the previous verb (e.g. in *estender, puxando* and other examples in section 6.4.3).

These corrections resulted in CARTÃO 3.1, a new version of this resource, after augmentation with:

- Antonymy relations from TeP 2.0, which comprise 4,276 sb-triples – 1,407 between nouns, 1,158 between verbs, 1,562 between adjectives and 149 between adverbs. Given that the final Onto.PT synsets are not exactly the same as in TeP, the former antonymy relations were converted to tb-triples. For this purpose, each sb-triple resulted in several antonymy tb-triples, each one connecting one lexical item from the synset in the first argument with an item from the synset in the second argument.
- Synsets from OpenThesaurus.PT, more precisely, those we could identify the POS, which comprise 3,925 synsets – 1,971 nouns, 831 verbs, 1,079 adjectives and 44 adverbs. As TeP was our synset-base, the former relations were converted to tb-triples, whose arguments would later be added to TeP synsets. For this purpose, each synset resulted in several synonymy tb-triples, each one connecting two different lexical items in the synset.

Finally, the tb-triples connecting two lexical items not occurring in CETEMPúblico or in TeP were discarded, unless they were extracted from more than one resource. This can be seen as a first approach to eliminate very unfrequent and probably unuseful words from Onto.PT. Table 8.1 shows the distribution of the tb-triples in the lexical network used to create Onto.PT v.0.35.

8.1.2 Synsets

We recall that the synsets of TeP 2.0 were used as a starting point for creating the Onto.PT synset-base. The assignment algorithm described in section 6.1.2 was used to enrich TeP with the synpairs of CARTÃO, after the second resource was augmented, as referred in the previous section. Following the experimentation described in section 6.2, we decided to use the cosine similarity measure, with mode

Relation	Args.	Quantity
Synonym-of	n,n	84,015
	v,v	37,068
	adj,adj	45,149
	adv,adv	2,626
Hypernym-of	n,n	91,466
Part-of	n,n	3,809
	n,adj	5,627
Member-of	n,n	6,369
	n,adj	114
	adj,n	948
Contained-in	n,n	364
	n,adj	280
Material-of	n,n	873
Causation-of	n,n	1,411
	n,adj	30
	adj,n	706
	n,v	78
	v,n	10,144
Producer-of	n,n	1,721
	n,adj	77
	adj,n	505
Purpose-of	n,n	7,100
	n,adj	85
	v,n	8,713
	v,adj	373
Has-quality	n,n	998
	n,adj	1,258
Has-state	n,n	345
	n,adj	216
Property-of	adj,n	10,617
	adj,v	27,431
Antonym-of	n,n	17,172
	v,v	49,422
	adj,adj	25,321
	adv,ad	683
Place-of	n,n	1,393
Manner-of	adv,n	2,166
	adv,adj	1,800
Manner without	adv,n	249
	adv,v	16
Total		448,738

Table 8.1: Quantities of relations used for the construction of Onto.PT.

Best and $\sigma = 0.15$, to assign the CARTÃO synpairs to the TeP synsets. Unassigned synpairs had a second chance of being assigned to a synset, in a second assignment iteration, using the same similarity measure, but $\sigma = 0.35$. Finally, clusters were discovered on the remaining synpairs, which originated new synsets. Clustering was performed using the algorithm described in section 6.3, with a threshold $\mu = 0.5$.

Table 8.2 shows the distribution of the Onto.PT synsets according to their POS. The current version of Onto.PT contains 108,837 synsets, of which 104,971 are involved in at least one sb-triple. Besides the discovered synsets, Onto.PT contains 78,724 synsets with only one lexical item, resulting from arguments of tb-triples

not covered by the synset-base. These lexical items are generally words without synonyms, or which are unfrequent.

POS	Synsets		
	size > 1	size = 1	Total
Nouns	16,927	44,511	61,438
Verbs	4,114	22,134	26,248
Adjectives	7,666	11,010	18,676
Adverbs	812	1,294	2,106
Total	29,519	78,949	108,468

Table 8.2: Onto.PT v.0.35 synsets.

8.1.3 Relations

The results reported in section 7.2 lead to the choice of different ontologising algorithms for attaching different types of tb-triples to the Onto.PT synsets. We started by ontologising the hypernymy tb-triples using the AC algorithm. Then, we ontologised all the other tb-triples, originally in CARTÃO, using the RP+AC algorithm. Finally, for ontologising the antonymy tb-triples from TeP 2.0, we used the RP algorithm, which guarantees that the attachments are as close as possible to the original attachments.

Table 8.3 shows the distribution of the about 173,000 sb-triples in Onto.PT v.0.35, according to their relation and type of connected synsets. We divide the sb-triples into those connecting: two single-item synsets ($1 \rightarrow 1$); one single-item synset with one with more lexical items ($1 \rightarrow n$, $n \rightarrow 1$); and two synsets with more than one lexical item ($n \rightarrow n$). The table also presents the names of each subtype of semantic relation, regarding the POS of its arguments. Those subtypes are the same as in PAPEL and CARTÃO, and are described in appendix A, together with an example for each, in English, and the name of the inverse relation.

Almost half of the sb-triples in Onto.PT are hypernymy. The second relation with most sb-triples is property-of between adjectives and verbs. The relation with less sb-triples is manner-without. The table also shows that the majority of sb-triples (about 82,000) connect one single-item synset with a synset with more than one lexical item. Only about 20,000 connect two synsets with more than one item.

8.1.4 Relation examples

In order to have an idea on the contents of Onto.PT, table 8.4 presents an example of a sb-triple of each relation type. For the sake of simplicity, we omitted less frequent words from larger synsets.

8.2 Access and Availability

Onto.PT and other resources developed in the scope of this research are available from the project's website, at `http://ontopt.dei.uc.pt`. There, Onto.PT is available for download, in a file where it is represented as a Semantic Web model. Alternatively, it can be queried through a web interface.

Relation	Args	Given name	Sb-triples			
			$1 \rightarrow 1$	$1 \rightarrow n$ $n \rightarrow 1$	$n \rightarrow n$	**Total**
Hypernymy	n,n	*hiperonimoDe*	2,753	40,480	37,105	80,338
Part	n,n	*parteDe*	546	1,621	1,502	3,669
	n,adj	*parteDeAlgoComPropriedade*	687	2,391	1,846	4,924
Member	n,n	*membroDe*	634	3,057	2,169	5,860
	n,adj	*membroDeAlgoComPropriedade*	20	66	24	110
	adj,n	*propriedadeDeAlgoMembroDe*	286	409	217	912
Contained	n,n	*contidoEm*	53	161	136	350
	n,adj	*contidoEmAlgoComPropriedade*	46	110	104	260
Material	n,n	*materialDe*	59	306	458	823
Causation	n,n	*causadorDe*	191	604	569	1,364
	n,adj	*causadorDeAlgoComPropriedade*	4	8	18	30
	adj,n	*propriedadeDeAlgoQueCausa*	114	226	283	623
	n,v	*causadorDaAccao*	9	25	42	76
	v,n	*accaoQueCausa*	529	2,505	4,747	7,781
Producer	n,n	*produtorDe*	263	632	727	1,622
	n,adj	*produtorDeAlgoComPropriedade*	15	31	31	77
	adj,n	*propriedadeDeAlgoProdutorDe*	23	219	203	445
Purpose	n,n	*fazSeCom*	690	2,868	3,181	6,739
	n,adj	*fazSeComAlgoComPropriedade*	12	36	36	84
	v,n	*finalidadeDe*	1,703	3,860	2,594	8,157
	v,adj	*finalidadeDeAlgoComPropriedade*	23	165	137	325
Place	n,n	*localOrigemDe*	614	508	168	1,290
Quality	n,n	*temQualidade*	194	491	271	956
	n,adj	*devidoAQualidade*	54	241	805	1,100
State	n,n	*temEstado*	61	151	113	325
	n,adj	*devidoAEstado*	10	31	153	194
Property	adj,n	*dizSeSobre*	2,350	4,622	2,781	9,753
	adj,v	*dizSeDoQue*	7,949	14,560	2,625	25,134
Antonymy	n,n	*antonimoNDe*	6	111	1,794	1,911
	v,v	*antonimoVDe*	4	69	1,762	1,835
	adj,adj	*antonimoAdjDe*	166	308	1,679	2,153
	adv,adv	*antonimoAdvDe*	12	14	81	107
Manner	adv,n	*maneiraPorMeioDe*	84	814	942	1,840
	adv,adj	*maneiraComPropriedade*	57	833	719	1,609
Manner without	adv,n	*maneiraSem*	0	82	137	219
	adv,v	*maneiraSemAccao*	1	10	6	17
Total			20,249	82,685	70,182	173,116

Table 8.3: Relational sb-triples of Onto.PT

Since the release of the first public version of Onto.PT, on April 2012, until 7th September 2012, the website had 650 visits, and 343 unique visitors. About 44% of the total number of visits are from Portugal, and 43% are from Brazil. The remaining 13% if the visits came from other countries, including Spain (3.5%), USA (1.8%), and France (1.5%).

In this section, we first describe, briefly, the representation of Onto.PT as a Semantic Web model. Then, we introduce OntoBusca, a web interface for Onto.PT, inspired by the WordNet search interface.

Examples of sb-triples
{*chouriça, chinguiço, chouriço, salsicha*} hiperonimoDe {*tabafeia, atabafeia, tabafeira*}
{*centro, núcleo, meio*} parteDe {*corpúsculo, indivisível, átomo*} {*vício, pecado, desvirtude*} parteDeAlgoComPropriedade {*grosseirão, anómalo, vicioso, desprimoroso*}
{*aluno, estudante, académico, educando, leccionando*} membroDe {*escola, colégio, seguidores*} {*coisa, assunto, cousa, ente*} membroDeAlgoComPropriedade {*coletivo*} {*português, lusitano, lusíada, luso*} propriedadeDeAlgoMembroDe {*portugal, lusitânia*}
{*sentido, pensamento, raciocínio, idéia*} contidoEm {*sentença, juízo, julgamento*} {*bílis, bila*} contidoEmAlgoComPropriedade {*biliário, biliar, bilioso*}
{*folha_de_papel*} materialDe {*canhenho, caderno, caderneta, livrete*}
{*escolha, discernimento*} causadorDe {*selecionamento, escolha, triagem, seleção, selecionado*} {*amor, predilecção, afecto, paixão, escatima*} causadorDeAlgoComPropriedade {*passional*} {*reactor, reaccionário, xiconhoca*} propriedadeDeAlgoQueCausa {*reacção, feedback, retroacção*} {*frio, griso, indeferença, briol*} causadorDaAccao {*entrevar, entrevecer, encangar, encarangar, encaranguejar*} {*mover, agitar, inquietar, alvoroçar*} accaoQueCausa {*inquietação, agitação, alvoroço, excitação*}
{*alfarrobeira, pão-de-san-joão, farrobeira*} produtorDe {*alfarroba, farroba, ferroba*} {*excitação, fermentação, levedação*} produtorDeAlgoComPropriedade {*lêvedo, crescido, fermentado, levedado*} {*fonador*} propriedadeDeAlgoProdutorDe {*música, som, sonância, canto, toada*}
{*antipirina, analgesina*} fazSeCom {*anilina*} {*comparação, equiparação*} fazSeComAlgoComPropriedade {*comparativo, confrontante, confrontativo*} {*apurar, calcular, contar*} finalidadeDe {*cálculo, operação, contagem, cômputo, apuração, computação*} {*diluir, fluidificar, humectar*} finalidadeDeAlgoComPropriedade {*humectante, humectativo, humente*}
{*lua, luar*} localOrigemDe {*lunícola, selenita*}
{*mórbido*} temQualidade {*morbidez, morbideza, nocividade, morbosidade*} {*grosseiro, crasso, grassento*} devidoAQualidade {*bronquice, crassície, crassidade, crassidão*}
{*agitação, desvairamento, delírio, exaltação*} temEstado {*devaneio, alucinação, delírio, alienação, desvario*} {*espalhado, disperso, esparramado*} devidoAEstado {*desfazimento, disseminação, espalhamento, dispersão*}
{*libertação, desacorrentamento*} antonimoNDe {*subjugação, agrilhoamento, acorrentamento, escravização*} {*retroceder, regredir, involuir*} antonimoVDe {*evoluir, evolucionar, evolver*} {*afinado, rectificado, ensoado, entoado*} antonimoAdjDe {*desafinado, desentoado, destoado*} {*cuidadosamente, atentamente*} antonimoAdvDe {*descortesmente, desatentamente, descuidadosamente*}
{*daltónico*} dizSeSobre {*daltonismo, discromatopsia*} {*fevroso, nervudo, musculoso, carnudo, musculado*} dizSeDoQue {*ter_músculo*}
{*firmemente, solidamente, fixadamente*} maneiraPorMeioDe {*fundamento, firmeza, consistência, solidez*} {*virtualmente, potencialmente*} maneiraComPropriedade {*possível, potencial, virtual*}
{*contínuo, seguido, seguidamente, ininterruptamente, a_fio*} maneiraSem {*interrupção, aparte, a-propósito*} {*objectivamente, positivamente, concretamente, materialmente*} maneiraSemAccao {*infundir, misturar*}

Table 8.4: Examples of sb-triples in Onto.PT.

8.2.1 Semantic Web model

Onto.PT is freely available and may be downloaded as a RDF/OWL model, typically used in the Semantic Web context. Our choice relied on the fact that RDF (Miller and Manola, 2004) and OWL (McGuinness and van Harmelen, 2004) are standards of the World Wide Web Consortium (W3C) for describing information as triples, consequently ontologies, and they are adequate representations for loading the ontology to a triple store (e.g. Sesame (Broekstra et al., 2002)), which provides useful features, such as indexing, querying and inferencing. Furthermore, as these models are standards, it is easier to find applications developed based on them, which makes them also a suitable representation for sharing Onto.PT with the community.

The structure of the ontology is based on the W3C RDF/OWL representation of Princeton WordNet (van Assem et al., 2006). There are four classes for the existing four kinds of synsets (`NomeSynset`, `VerboSynset`, `AdjectivoSynset`, `AdverbioSynset`) and we have defined all the types of extracted semantic relations, as well as their inverse relations, as `ObjectTypeProperties`. Each synset has two kinds of `DataTypeProperties`: an id (`synsetId`), and one or more lexical forms (`formaLexical`), which are the canonical forms of the lexical items it includes. Figure 8.2 illustrates the schema of the RDF/OWL model. We decided to keep the diagram simple, so it only contains three semantic relations and their

inverses, connected by the `inverseOf` attribute. The same figure also shows the constraints on the type of synset for the first and second arguments of each relation, using respectively the `range` and `domain` attributes.

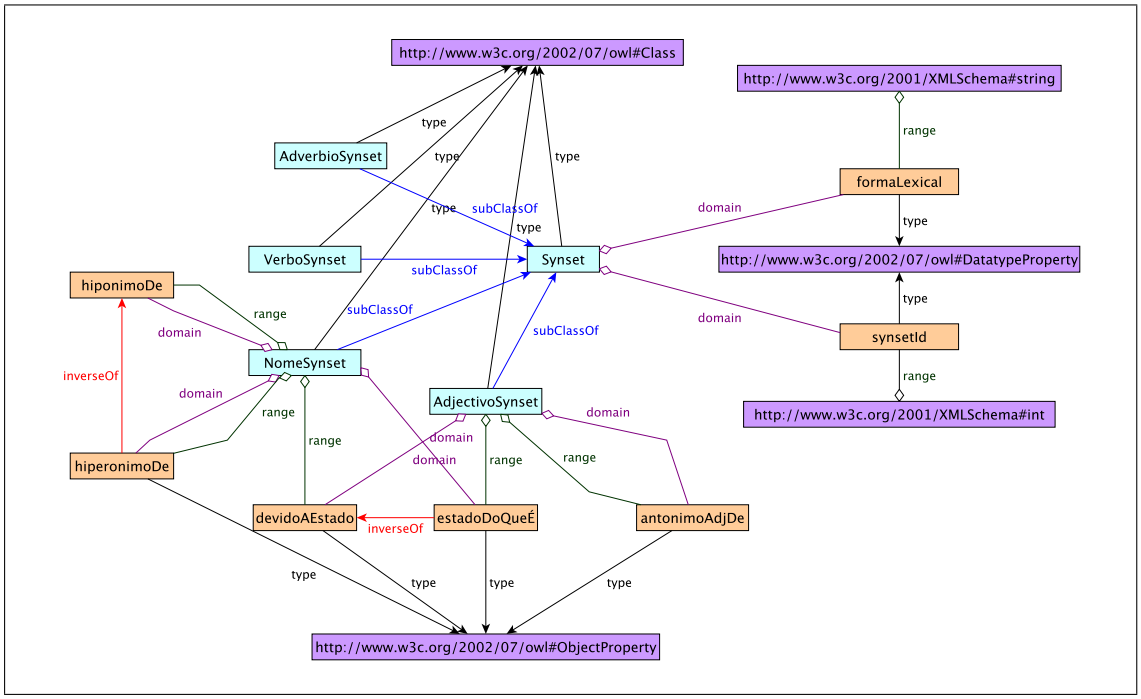


Figure 8.2: Part of the Onto.PT RDF/OWL schema.

The ontology schema was populated with the discovered synsets and sb-triples. Figure 8.3 shows how the synset instances (#id) are represented and how their relations are established. It presents three sb-triples, namely:

- {*malcontente, insatisfeito, desagradado, descontente*} antonimoAdjDe {*ridente, satisfeito, contente, contento*}
  ({unsatisfied, discontent} antonym-of {smiling, satisfied, content})
- {*satisfação, contentamento*} estadoDoQueÉ {*ridente, satisfeito, contente, contento*}
  ({satisfaction, contentment} state-of {smiling, satisfied, content})
- {*sensação, sentir, sentimento*} hiperonimoDe {*satisfação, contentamento*}
  ({sensation, feeling} hypernym-of {satisfaction, contentment})

8.2.2 Web interface

In order to provide a friendlier way to explore the contents of Onto.PT, Onto-Busca was developed. This web interface enables to query a triple store with the RDF/OWL representation of Onto.PT. This interface is very similar to WordNet Search[4] and allows to query for one lexical item, in order to obtain all the synsets containing it. The returned synsets, represented by all their items, may be expanded

[4]The web interface for Princeton WordNet, developed by its authors and available from `http://wordnetweb.princeton.edu/perl/webwn` (September 2012)

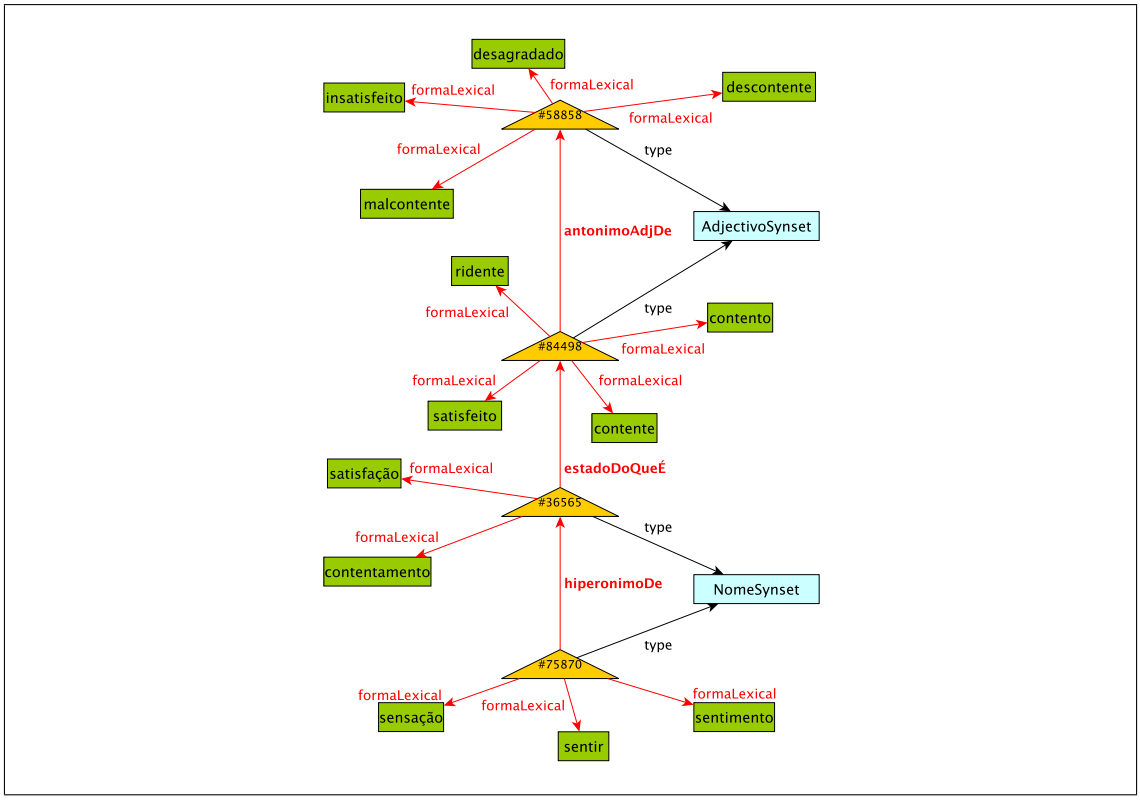


Figure 8.3: Instances in the Onto.PT RDF/OWL model.

in order to access semantic relations. OntoBusca contains also a word cloud with the most frequently searched lexical items.

Figure 8.4 shows OntoBusca, after querying for the word *hospital* and expanding some of the obtained synsets. The available relations include direct and indirect relations. Besides two adjective synsets with meanings close to a person with good intentions, there is only one noun synset with the word *hospital* (hospital, in English). Some of the presented relations indicate, for instance, that this synset is a hyponym (*hiponimoDe*) both of *edifício* (building) and *firma* (firm), and it is a hypernym (*hiperonimoDe*) of *manicômio* (psychiatric hospital) and *gafaria* (leprosy hospital). Furthermore, the property *hospitalar* refers to a hospital (*referidoPorAlgoComPropriedade*), and the *enfermaria* (ward) is part of (*temParte*) a *hospital*. It is as well possible to expand the related synsets, as in the example the synset *enfermaria* is, showing that, for instance, its purpose (*meioPara*) is to be used by *pacientes*/*doentes* (patients/sick people).

8.3 Evaluation

This section is dedicated to the evaluation of Onto.PT v.0.35. In the previous chapters, all the creation steps were validated and the quality of their results was quantified. So, we start this section with a brief summary of the evaluation described there. Then, we complement the evaluation of Onto.PT with the manual evaluation of its sb-triples. We end this section with a coverage evaluation, where we tried to match each of the 164 core concepts of a wordnet, as suggested by the Global

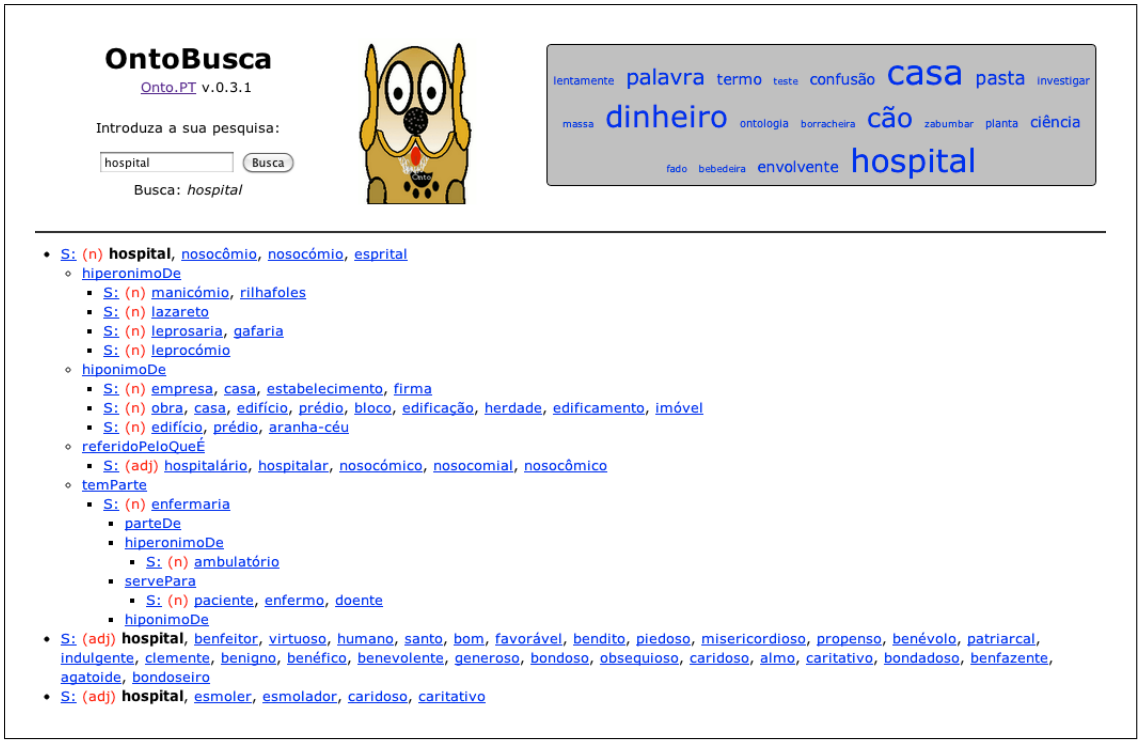


Figure 8.4: OntoBusca, Onto.PT's web interface.

Wordnet Association, with an Onto.PT synset.

8.3.1 Summary of evaluation so far

Among the possible strategies to evaluate an ontology, a survey by Brank et al. (2005) presents four, which are probably the most commonly followed when it comes to domain ontologies:

- Manual evaluation, performed by humans;
- Comparison with an existing gold standard, eventually another ontology;
- Coverage evaluation, based on a dataset on the same domain;
- Task-based evaluation, where the ontology is used by an application to achieve some task.

Even though Onto.PT is not a domain ontology, we can say that, throughout this research, and depending on what we were evaluating, we have followed the first, the second and the third approaches.

First, in the extraction step (chapter 4), before performing the manual classification of some extractions, we evaluated the coverage of the extracted information by handcrafted thesauri, and by a newspaper corpus. Attention should be given to the coverage evaluation because, as Onto.PT is broad-coverage, it is not possible to find something like a corpus of its own domain. A language thesaurus is probably the closest thing. Furthermore, the corpus was used to validate the coverage of the relations, and was based on a limited set of discriminating patterns. When it comes to estimating the accuracy of the extracted relations, manual evaluation is probably

the most reliable. Following this kind of evaluation, we concluded that the accuracy of tb-triples extracted from dictionaries depends on the relation type. Accuracies are between 99%, for synonymy, and slightly more than 70%, for purpose-of and property-of. Hypernymy, which is the relation with more extracted instances, is about 90% accurate.

The enrichment of synsets (chapter 6) was evaluated after the comparison with a reference gold standard, especially created for this task. The agreement on the creation of the reference is moderate, and the accuracy of assigning a synonymy pair to a synset with our procedure and the selected parameters is between 76% and 81%, depending on the judge. We recall that we are using TeP as a starting point for the synset-base. Since the previous resource is created manually, the final quality of synsets should be higher than the aforementioned values. As for the establishment of new clusters from the remaining synonymy pairs, manual evaluation was once again followed, and the accuracy of this step has shown to be around 90%. This is an improvement towards the discovery of clusters from the complete synonymy network, where accuracy for nouns was about 75% (chapter 5).

Finally, for evaluating the attachment of the term arguments of tb-triples to synsets (chapter 7) we used two gold standards – one created manually, for the hypernymy, part-of and purpose-of relations; and TeP, for antonymy relations. Using the best performing algorithms, the precision of this step was measured to be between 99% for antonymy, and 60-64%, for the other relations. This number, however, considers only attachments to synsets with more than one lexical item. For lexical items that originate a single-item synset, attachment is straightforward. Given that more than two thirds of the Onto.PT synsets are single-item, ontologising performance will be higher in the actual creation of Onto.PT.

All combined, we understand that there will be some reliability issues with the contents of Onto.PT, common in an automatic approach. However, it is not possible to speculate on a value or an interval for this reliability, because the ECO approach is not linear. Having this in mind, we decided to complement the previous evaluation hints by classifying a small part of the Onto.PT sb-triples manually, with results provided in the next section. Although manual evaluation has almost the same problems as creating large resources manually – it is tedious, time-consuming, and hard to repeat – it is also the more reliable kind of evaluation.

Besides the manual evaluation, in section 8.4 we present how Onto.PT can be useful in the achievement of some NLP tasks. This can be seen as a task-based evaluation, which is the fourth strategy referred by Brank et al. (2005).

8.3.2 Manual evaluation

As referred earlier, manual evaluation suffers from similar issues as creating large resources manually. It is a tedious, time-consuming job, and hard to repeat. However, one of the few alternatives we found in the literature for evaluating a wordnet automatically is based on dictionaries (Nadig et al., 2008). Since all the available Portuguese dictionaries were exploited in the creation of Onto.PT, a similar evaluation would be biased.

The manual evaluation of Onto.PT considered, first, the synsets alone and, second, the proper sb-triples. More precisely, this evaluation had the following steps:

- Two random samples of 300 sb-triples were collected from Onto.PT: one with

only hypernymy relations and another with other relations (sets A). The only constraint in the selection was that each sb-triple had, at least, one synset argument with more than one lexical item. The option on having a separate set with hypernymy relied both on the fact that hypernymy is one of the most used semantic relations, and its is also the relation with more instances in Onto.PT – almost a half of its sb-triples are hypernymy.

- The synsets in the arguments of the previous sb-triples were reduced, so that they would contain, at most, three lexical items. This was done by keeping only the three lexical items with highest frequency, according to the lists provided by the AC/DC service.
- From both the previous samples, a set with all the (reduced) synsets with more than one lexical item in both was created.
- A group of eight human judges was asked to classify, independently, each synset as correct or incorrect. A correct synset must contain only words that, in some context, have the same meaning. The judges were advised to use online dictionaries, if needed. Each judge classified different quantities of synsets, according to their availability. Still, we made sure that each synset was classified by two different judges.
- From both initial random samples of 300 sb-triples, those connecting one synset classified twice as incorrect were removed, giving rise to a smaller set of sb-triples (sets B).
- Two human judges were asked to classify, independently, each of the remaining sb-triples as correct or incorrect. Once again, the judges were advised to use online dictionaries, if needed. Also, they were provided with a list containing the description and examples of relations, included in appendix A. The arguments of the sb-triples were also shown in their reduced form.

In the 600 sb-triples, there were 774 unique synsets with more than one item. From those, 572 (73.9%) and 58 (7.5%) were respectively classified as correct and incorrect by both judges. For the remaining 144 (18.6%), the judges did not agree.

Table 8.5 presents the results of the manual evaluation of sb-triples. They are separated into a set with the hypernymy and another with the other kinds of relation. On each set and for each judge, we present the proportion of correct sb-triples, both considering that those with incorrect arguments are not correct (A), or considering only the classified sb-triples (B). For a confidence interval of 95%, the margin of error for the results of set A is presented (ME_A). Judge agreement is also shown, by the number of matches (IAA), and by the *Kappa* coefficient.

Relation	**Quantity**		**Judge**	**Correct**		**ME$_A$**	**IAA**	κ
	A	B		A	B			
hiperonimoDe	300	247	J1	65.0%	78.9%	5.4%	82.6%	0.47
			J2	64.7%	78.5%	5.4%		
Others	300	267	J1	78.3%	88.0%	4.7%	90.1%	0.48
			J2	82.0%	92.1%	4.3%		

Table 8.5: Results of the manual evaluation of sb-triples.

The values in the column 'Correct' and sub-column 'A' can be seen as an approx-

imation of the correct sb-triples of Onto.PT. This value is lower for hypernymy than for other relations. A possible explanation for this fact is that hypernymy typically connects more frequent lexical items, which are also those with more senses. The following are among the most frequent hypernyms: *pessoa* (person), *planta* (plant), *árvore* (tree), *indivíduo* (individual), *instrumento* (instrument), *substância* (substance), *lugar* (place), *peça* (piece). On the other hand, the higher correction of the other relations is increased by the antonymy relations from TeP, which are correct, because the resource was created manually. Also, although all relations are very different, there are many connecting items with very specific meaning, hardly attached incorrectly, such as:

- {*atolar, chafurdar, atascar*} causation-of {*atolamento*}
  (to_get_bogged_down_in_the_mud causation-of jam)
- {*trépido, vibrador, vibrante*} has-quality {*vibratilidade*}
  (vibrating has-quality vibratility)
- {*bailariqueiro, bailomaníaco*} property-of {*ter_mania_de_baile*}
  (fond_of_dancing property-of have_the_craze_for_ball)
- {*pegar, apanhar, pescar*} purpose-of {*taloeira*}
  (to_fish purpose-of gaff)
- {*Lisboa*} place-of {*lisbonense, olisiponense, lisboeta*}
  (Lisbon place-of lisbonian)
- {*incalculavelmente*} manner-of {*incalculável, incogitável, inestimável*}
  (incalculably manner-of incalculable)

Having in mind the evaluation of the ontologising algorithms (see chapter 7), the proportion of correct hypernymy sb-triples is still higher than expected. In the aforementioned evaluation, all synsets were correct, which makes that result (60.1% precision) comparable to the proportion of correct hypernymy sb-triples in set B (>78%). Besides a few improvements made in the current version of Onto.PT, our explanation for this discrepancy relies on the following factors:

- In the ontologisation evaluation, only sb-triples connecting two synsets with more than one lexical item were used. Here, the samples contained some sb-triples connecting a single-item synset with a synset with more items.
- Synset arguments were reduced to at most three lexical items, which might hide incorrect lexical items.
- There is some human tolerance when classifying if two sets of lexical items (synsets) effectively hold some relation.

Nevertheless, we believe that the values of this evaluation are closer to the real reliability of Onto.PT, not only because the samples are larger, but also because more than a half of the Onto.PT sb-triples actually connect one or two single-item synsets (see table 8.3). Although possibly less useful, those have higher probability of being correct and thus increasing the reliability of the resource. Another point supporting that this evaluation is closer the real reliability is that it was performed by two judges, who classified about the same proportion of sb-triples as correct, with moderate agreement (Landis and Koch, 1977).

To give an idea on the quality of each relation type, table 8.6 presents the results of the evaluation of the 300 non-hypernymy tb-triples. The number of evaluated sb-triples is not enough to take strong conclusions, but the quantity of *dizSeDoQue* correct tb-triples stands out. It is higher than the extraction accuracy of these relations (*adj* property-of *v*, 71-77% correct in section 4.2.5) from dictionaries. Given that most of the problems about these relations were due to incorrect arguments, we view this improvement as a consequence both of the new lemmatisation rules, and of the removal of tb-triples with arguments not occurring in the corpus, performed before ontologisation. On the other hand, the quality of the *parteDe* sb-triples is quite low. Similarly to what happens with hypernymy, the main problem about this relation seems to be the ambiguity and underspecification of its arguments. This has a negative impact both on the establishment of correct synsets with these words and on the ontologisation of tb-triples.

8.3.3 Global coverage

If compared to the number of sb-triples in Princeton WordNet 3.0 (see section 3.1.1), developed manually between 1985 and 2006, Onto.PT v.0.35 is larger because all of its relations can be inverted. This means that Onto.PT contains about 346,000 sb-triples against the 285,000 of WordNet 3.0. This number, which may soon increase, if more resources are exploited, highlights the potential of an automatic approach.

As this number is insufficient to quantify the coverage of Onto.PT, we evaluated its coverage of base concepts, that should be represented in wordnets. The Global WordNet Association[5] provides several lists with this kind of concepts. One of them contains 164 base concepts, referred to as the "most important" in the wordnets of English, Spanish, Dutch and Italian[6]. The concepts are divided into 98 abstract and 66 concrete, and are represented as Princeton WordNet 1.5 synsets.

In order to evaluate the global coverage of Onto.PT, we tried to make rough matches, manually, between the 164 base concepts and Onto.PT synsets. Given the WordNet synset denoting each of the 164 concepts, we selected the Onto.PT synset closer to its meaning. In the end, we concluded that Onto.PT roughly covers most of the concepts in the list, more precisely 92 abstract and 61 concrete synsets (93%).

All the defined matches are reported in the appendix B of this thesis. More precisely, the concrete concepts are in table B.1, and the abstract in tables B.2. There, we can see that the Onto.PT synsets are, on average, larger than WordNet's, which means, on the one hand, that they are very rich, with many synonyms – most synsets include various levels of language (formal, informal, figurative, older forms...) and variants of Portuguese (Portugal, Brazil, Africa). This can be very useful for tasks from information retrieval (see section 8.4.3) to creative writing (e.g. poetry). On the other hand, the matches show that there are synsets that go beyond including only synonyms – most noisy items are more like near-synonyms and some are closely related words.

Considering just the abstract concepts not covered by Onto.PT (e.g. *change magnitude*, *definite quantity*, *visual property*), they seem to have been created ar-

[5]See website at `http://www.globalwordnet.org/` (September 2012)

[6]See more about this list in `http://www.globalwordnet.org/gwa/ewn_to_bc/corebcs.html` (September 2012)

Relation	Quantity		Judge	Correct		IAA
	A	B		A	B	
parteDe	15	10	J1 J2	33.3% 46.7%	50.0% 70.0%	80.0%
parteDeAlgoComPropriedade	20	18	J1 J2	85.0% 80.0%	94.4% 88.9%	94.4%
membroDe	17	14	J1 J2	82.4% 82.4%	100% 100%	100%
propriedadeDeAlgoMembroDe	4	4	J1 J2	75.0% 100%	75.0% 100%	75.0%
contidoEm	2	2	J1 J2	100% 50.0%	100% 50.0%	50.0%
contidoEmAlgoComPropriedade	1	1	J1 J2	100% 100%	100% 100%	100%
materialDe	4	4	J1 J2	75.0% 100%	75.0% 100%	75.0%
causadorDe	5	5	J1 J2	100% 80.0%	100% 80.0%	80.0%
propriedadeDeAlgoQueCausa	2	2	J1 J2	50.0% 50.0%	50.0% 50.0%	100%
accaoQueCausa	32	30	J1 J2	84.4% 90.6%	90.0% 96.7%	93.3%
produtorDe	7	4	J1 J2	42.9% 28.6%	75.0% 50.0%	75.0%
propriedadeDeAlgoProdutorDe	1	1	J1 J2	100% 100%	100% 100%	100%
fazSeCom	24	22	J1 J2	70.8% 75.0%	77.3% 81.8%	86.4%
fazSeComAlgoComPropriedade	1	1	J1 J2	100% 100%	100% 100%	100%
finalidadeDe	24	21	J1 J2	70.8% 83.3%	81.0% 95.2%	85.7%
localOrigemDe	4	4	J1 J2	50.0% 75.0%	50.0% 75.0%	75.0%
temQualidade	2	2	J1 J2	100% 100%	100% 100%	100%
devidoAQualidade	11	10	J1 J2	81.8% 90.9%	90.0% 100%	90.0%
devidoAEstado	1	1	J1 J2	100% 100%	100% 100%	100%
antonimoNDe	5	5	J1 J2	80.0% 80.0%	80.0% 80.0%	100%
antonimoAdvDe	1	1	J1 J2	100% 100%	100% 100%	100%
antonimoVDe	6	9	J1 J2	66.7% 66.7%	100% 100%	100%
antonimoAdjDe	9	8	J1 J2	88.8% 88.8%	100% 100%	100%
dizSeSobre	29	29	J1 J2	96.6% 93.1%	96.6% 93.1%	96.6%
dizSeDoQue	53	48	J1 J2	81.1% 88.7%	90.0% 97.9%	87.5%
maneiraPorMeioDe	9	8	J1 J2	88.9% 88.9%	100% 100%	100%
maneiraComPropriedade	7	6	J1 J2	85.7% 85.7%	100% 100%	100%

Table 8.6: Results of the manual evaluation of sb-triples per relation type.

tificially, and work as "covert categories" of more specific concepts. All the verb synsets are covered by Onto.PT. If we compare these results with the coverage of the concepts by MWN.PT, Onto.PT covers all the verbs, which are not included in MWN.PT. Furthermore, despite covering the concepts of *human_action* and *magnitude_relation*, their correspondence in MWN.PT are gaps, possibly because its authors did not find a suitable translation for them. Onto.PT does not cover the latter concept too, but we could find a suitable match for the former (*feito*, *obra*,

acto, ...). Another important difference regards the size and the correction of the synsets of MWN.PT and Onto.PT. The former contains small synsets, often with only one word, while the latter, as referred earlier, contains large synsets. On the other hand, given its manual revision, the MWN.PT synsets are supposedly all correct, while Onto.PT, due to its automatic construction, contains incorrections.

8.4 Using Onto.PT

The main goal of creating Onto.PT is its exploitation in the achievement of tasks on the computational processing of Portuguese. As referred earlier in this chapter, this is also a popular approach to validate ontologies (Brank et al., 2005).

In order to illustrate the utility of a resource as Onto.PT, in this section, we provide utilisation scenarios, where this resource can be seen as a valuable contribution. All the scenarios intend to be mere proofs of concept. None of the used techniques are very sophisticated and we did not go further on their evaluation. We start by presenting an exercise on exploring the taxonomy of Onto.PT. Then, we show how Onto.PT can be applied to word sense disambiguation (WSD). After that, we briefly describe how this resource was integrated in an information retrieval (IR) system, in order to enhance query expansion. The IR system was evaluated with the participation in an IR joint task. The last utilisation scenario is about taking advantage of Onto.PT to answer cloze questions automatically.

8.4.1 Exploring the Onto.PT taxonomy

The first usage example is a simple exploration exercise, showing that, besides providing synonyms for lexical items, Onto.PT can be queried to acquire taxonomic information, as well as other semantic information on the organisation of the lexicon. Figure 8.5 shows a four-level taxonomy obtained from Onto.PT, where *cão* (dog) is included. For the sake of simplicity, we omit some synset entries, as well as non-hypernym relations, from this figure.

The taxonomy shows that Onto.PT can be used, for instance, to collect a list of animals, a list of mammals, or a list of dog breeds. Starting with the most general level, with a synset denoting an animal, its is possible to obtain kinds of animals, including birds (*ave*), insects (*insecto*), and mammals (*mamífero*). Mammals can be expanded for obtaining mammal synsets, including cow (*vaca*), whale (*baleia*), cat (*gato*), or dog (*cão*). Finally, if the hypernyms of the dog synset are expanded, several dog breeds are shown, including boxer, mongrel (*rafeiro*) or dalmatian (*dálmata*).

8.4.2 Word sense disambiguation

There is a wide range of knowledge-based WSD algorithms, using a wordnet both as sense inventory and as an additional source of knowledge (e.g. Resnik (1995); Banerjee and Pedersen (2002); Agirre and Soroa (2009)). As Onto.PT is structured in a similar fashion to a wordnet, most of the previous algorithms may be adapted to use Onto.PT for performing Portuguese WSD. We have implemented two algorithms for this task: Bag-of-Words and Personalized PageRank (Agirre and Soroa, 2009).

- S: (n) animal, bicho, balada, piolho, béstia, alimal, minante
- **[hiperonimoDe]**
  - ...
  - S: (n) ave, ribeirinhas, sabacuim, volátil
  - S: (n) micróbio, bactéria, microorganismo, bacilo, microrganismo
  - S: (n) insecto, inseto, xerimbabo
  - ...
  - S: (n) mamífero, mamíferos, mastozoário
  - **[hiperonimoDe]**
    - ...
    - S: (n) vaca, seminarista, vaquinha
    - S: (n) baleia
    - S: (n) gato, grampo, tareco, narro
    - S: (n) rata, rato, ratazana, toupeira
    - S: (n) cão, cachorro, cã, narro, calote, perro, mísula, au-au, adia, bêlfo, jaguara, cátulo
    - **[hiperonimoDe]**
      - ...
      - S: (n) boxer
      - S: (n) rafeiro, vira-lata
      - S: (n) galgo, lebreiro, lebrel
      - S: (n) perdigueiro
      - S: (n) dálmata
      - S: (n) poodle, caniche
      - S: (n) buldogue
      - S: (n) labrador, labradorite
      - S: (n) pastor-alemão
      - S: (n) dobermann
      - S: (n) são-bernardo
      - S: (n) husky
      - S: (n) bigle
      - ...
    - S: (n) castor
    - S: (n) tigre, tigrinho, tigrino
    - S: (n) golfinho, golfim, germão
    - S: (n) raposa, volpe, tamarança
    - S: (n) leopardo, pardo, pantera
    - S: (n) lince
    - S: (n) onça, jaguar, onça-pintada, leopardo-das-neves
    - S: (n) canguru
    - S: (n) gorila, gorilha
    - S: (n) morcego, pacó, guembo
    - S: (n) javali, porco-montês, porco-bravo
    - S: (n) veado, cervo, corço, enho
    - S: (n) gazela
    - S: (n) foca, vítulo, arctocéfalo, boi-marinho
    - S: (n) dromedário
    - S: (n) panda
    - S: (n) lontra, ratão-d'água, nútria
    - S: (n) girafa, camelopárdale
    - ...
  - S: (n) micro-organismo
  - ...

0.9

Figure 8.5: Search example: breeds of dog in Onto.PT.

Given a sentence to disambiguate, both of the algorithms take advantage of a given context $W = \{w_1, w_2, ..., w_n\}$, which includes all the (content) words of the sentence (nouns, verbs and, eventually, adjectives and adverbs). Before applying the algorithms, the sentence is POS-tagged and lemmatised. Then, for each word $w_i \in W$ to be disambiguated, the set of candidate synsets, $C_i = \{S_{i1}, S_{i2}, ..., S_{im}\}$, is retrieved from Onto.PT. Each candidate synset must contain the word to be disambiguated, $S_j \in C \rightarrow w_i \in S_j$. The goal of each algorithm is to select a suitable synset $S_k \in C$, for the occurrence of the word w_i in the context W. The selected synset should transmit the meaning of the word, when in the given context. The selection of the best candidate depends on the used algorithm:

Bag-of-Words: For each candidate S_j, a set $R_j = \{q_{j1}, q_{j2}, ..., q_{jp}\}$ is established with all the words in S_j and in synsets directly related with S_j, in Onto.PT. The selected synset is the one maximising the similarity with the context $S_k : sim(R_k, W) = max(sim(R_i, W))$. Similarities may be computed by measures typically used for comparing the similarity of sets, such as the Jaccard or the Overlap coefficient (both referred in section 6.1.2 and other sections of this thesis).

This algorithm is actually an adaptation of the Lesk algorithm (Lesk, 1986; Banerjee and Pedersen, 2002), with two main differences. First, in the Lesk algorithm adapted for WordNet, the "context" of a sense consists not only of the words in the synset, but also of words in its gloss and in example sentences. As Onto.PT does not contain synset glosses, we use all the words in related synsets. Second, in the Lesk algorithm, the similarity of contexts is given by the number of common terms, while we use a more complex similarity measure. This way, the selection of the most suitable synset is not biased towards synsets with larger "contexts".

Personalized PageRank: As referred in section 7.1, the PageRank algorithm (Brin and Page, 1998) ranks the nodes of a graph according to their structural importance. However, it has been used to solve more specific problems, including WSD with a wordnet (Agirre and Soroa, 2009). Our implementation is based on the later work, and uses all Onto.PT. For such, we consider that Onto.PT is a graph $G = (V, E)$, with $|V|$ nodes, representing the synsets, and $|E|$ undirected edges, for each relation between synsets. For a given context W, only the synsets with words in the context have initial weights, which are uniformly distributed. The rest of the synsets do not have an initial weight. After several iterations, it is expected that more relevant synsets for the given context are ranked higher. Therefore, for each word w_i, this algorithm selects the highest ranked candidate synset.

WSD using Onto.PT is exemplified in the following real sentences, obtained from AC/DC (Santos and Bick, 2000). For each sentence, we used all nouns and verbs as context, and applied the Personalized PageRank algorithm to assign a suitable synset for the occurrence of each noun. Each sentence is presented with the nouns underlined. Then, for each noun, we show, in parenthesis, the number of senses they have in Onto.PT, which is the number of alternative synsets including them, and, of course, we show the selected synset.

(1) *Vai estar, seguramente, colocado num <u>local</u> envergonhado e inacessível que obrigará o pobre <u>cidadão</u> que pretenda reclamar a sujeitar-se à <u>censura</u> de*

todos os funcionários presentes.

(It will, certainly, be placed in some shy and inaccessible place, that will force the poor citizen who wishes to complain submit themself to the censorship of all present workers.)

local (2)	→	{*situação, lado, lugar,* ***local****, sítio, localidade, arrozal, logo, sombral, loco, básis*}
cidadão (1)	→	{*homem, tipo,* ***cidadão****, indivíduo, cara, sujeito, camarada, cabra, gajo, freguês, caramelo, meco, sicrano, nego, tal, zinho, dito-cujo, ...*}
censura (5)	→	{*acusação,* ***censura****, exprobração, increpação, objurgação, objurgatória*}
funcionário (1)	→	{*trabalhador,* ***funcionário****, empregado, contratado*}

(2) *Ambos, na opinião do autor, atingiram um plano muito elevado, pelo que não é o facto de se terem retirado da vida política activa que os deixou igual a toda a gente.*

(Both of them, in the author's opinion, reached a very high level, so it is not the fact that they have withdrawn from active political life that left them as everyone else.)

opinião (12)	→	{***opinião****, voto, conselho, parecer, sugestão, arbítrio, alvitre*}
autor (7)	→	{***autor****, produtor, artífice, perpetrador, fabricador, responsável*}
plano (10)	→	{*nível,* ***plano***}
facto (6)	→	{***facto****, coisa, negócio, realidade, espécie, passo, acto, fenómeno, mistério, cousa*}
vida (16)	→	{***vida****, biografia*}
gente (6)	→	{***gente****, ser_humano*}

(3) *O marketing da convenção prevê a distribuição de 200 outdoors e anúncios comerciais em três emissoras de televisão e oito de rádio.*

(The convention marketing foresees the distribution of 200 billboards and advertisements in three television and eight radio stations.)

marketing (1)	→	{***marketing****, mercadologia*}
convenção (8)	→	{*acordo, negócio, contrato, tratado,* ***convenção****, convénio, concórdia*}
distribuição (7)	→	{***distribuição****, circulação*}
outdoor (1)	→	{*cartaz, ecrã, painel, retábulo,* ***outdoor***}
anúncio (8)	→	{***anúncio****, publicidade, propaganda, comercial, proclamação, cartel, utilitário, pregão, deixa, reclame, reclamo, papeleta, apostolado*}
emissora (2)	→	{***emissora****, transmissora*}
televisão (3)	→	{***televisão****, tv, tevê, televisora*}
rádio (2)	→	{***rádio****, transmissão, radiodifusão, radiocomunicação, radiotransmissão, radiofonia*}

Identifying the synset with the meaning of a word in context is important for handling ambiguities at the semantic level, and is the starting point for sense-aware NLP. Furthermore, it can be used to obtain other related words not referred in the text, useful for several tasks, including IR, where queries can be expanded with related information (see section 8.4.3); writing aids; or text simplification (Woodsend and Lapata, 2011). On the last, synonyms enable to rewrite the sentence with more frequent words, while keeping a very similar meaning. If we replace the nouns of sentence (1) with their synonyms with higher frequency in the AC/DC lists, we obtain the following sentence:

(4) *Vai estar, seguramente, colocado numa situação envergonhada e inacessível que obrigará o pobre homem que pretenda reclamar a sujeitar-se à acusação de todos os trabalhadores presentes.*

8.4.3 Query expansion

In IR, query expansion consists of refining a certain request for information (query) in order to improve the retrieval performance. Expansion may, for instance, replace the terms of the query by their lemmas or stems, give different weights to different terms in the query, or add related terms, such as synonyms, which can be alternatively searched for.

When it comes to adding related terms, LKBs have revealed to be very useful. For instance, Navigli and Velardi (2003) used Princeton WordNet for this task. They made several experiments where they first disambiguate the query terms with respect to WordNet. Then, they expand the query with words in the same synsets, on hypernym synsets, as well as words in the respective synset glosses. For Portuguese, Sarmento et al. (2008) analysed the benefits of using OpentThesaurus.PT and an automatically generated verb thesaurus for query expansion.

A previous version of Onto.PT (v.0.31) was recently used for query expansion in the system Rapportágico (Rodrigues et al., 2012). This system participated in Págico[7] (Mota et al., 2012; Santos et al., 2012), an IR joint task for Portuguese. Págico is briefly described as a task where, given a list of 150 information requests (topics), written in natural language, the goal is to identify pages of the Portuguese Wikipedia which answer each topic. If the answer is not in the answer page, the page supporting the answer should also be given. All the topics were about the culture of Portuguese speaking countries. The following are real examples of Págico topics:

(5) *Grupos indígenas que habitavam o litoral do Brasil quando chegaram os europeus.*

(Indigenous groups who inhabited the coast of Brazil when the Europeans arrived.)

(6) *Viajantes ou exploradores que escreveram sobre o Brasil do século XVI.*

(Travelers or explorers who wrote about the sixteenth-century Brazil.)

(7) *Sambistas negros que abordam o racismo em suas letras.*

(Black samba musicians that addressed racism in their lyrics.)

Rapportágico is based on a shallow analysis of the topic, which it converts into a query for retrieving relevant documents, indexed by the Apache Lucene search engine[8]. The baseline approach of Rapportágico uses the lemmas of the nouns and the verbs in the topic as search keywords. In all the runs submitted to Págico, the baseline approach adds two refinements:

- All occurrences of words related with Portuguese speaking countries (e.g. `lusófono`) were expanded to all the effective names of these countries (in Portuguese, `Portugal`, `Brasil`, `Angola`, `Moçambique`, `Guiné Bissau`, `Cabo Verde`, `São Tomé e Príncipe`, `Timor`);
- The first noun of the topic was considered to be the category of the topic, and was always searched for appended to a very common hypernymy pattern in Wikipedia pages – `é um`, in Portuguese, *is a*, in English

[7] See `http://www.linguateca.pt/Pagico/` (September 2012)

[8] Freely available from `http://lucene.apache.org/` (September 2012)

In addition to the previous baseline, we had two official runs where Onto.PT was used to perform an additional expansion on the 67 topics containing verb phrases (VPs) with only one verb. There, the verbs were disambiguated, and their synonyms were used as search alternatives. The main idea behind this expansion was the improvement of the system's recall. Still, only alternatives with more than 20 occurrences in the the corpora provided by AC/DC were used. The only difference between these two runs was that, in number 2, disambiguation was performed using the Bag-of-Words algorithm, while run number 3 used the Personalized PageRank. Moreover, after the official evaluation, we sent additional unofficial runs, where, besides other experiments, we had similar runs to 2 and 3, but this time, the category of all the topics was disambiguated and expanded as well.

In order to illustrate how expansion worked, figure 8.6 presents the expansions of the category and the VP of the previously shown topics, obtained with the Personalized PageRank. For the sake of simplicity, we omitted the hypernymy pattern from the category expansion.

Topic	**Category**		**VP**	
	Original	Expanded	Original	Expanded
5	*tribo*	`grupo OR tribo`	*habitavam*	`habitar OR colonizar OR povoar OR ocupar`
6	*viajantes ou exploradores*	`viajante OR peregrino OR viageiro OR passageiro OR caminhante OR viandante OR explorador`	*escreveram*	`redigir OR escrever OR grafar`
7	*sambistas*	`sambador OR sambista`	*abordam*	`tratar OR apalavrar OR abordar OR versar`

Figure 8.6: Category and VP expansions in Rapportágico, using Onto.PT.

Given the simplistic approach followed by Rapportágico and the high complexity of Págico, we can say that the obtained results were interesting. Rapportágico's performance was below most of the human participants, but it was better than RENOIR (Cardoso, 2012), the other automatic participant. Nevertheless, RENOIR also followed a simplistic approach, and was heavily penalised by the large number of given answers per topic (100). The most relevant conclusions for our research was that the runs where VPs were expanded into their synonyms performed better than the baseline approach. Among these two runs, Personalized PageRank performed better than the Bag-of-Words method.

The results of the official participation of Rapportágico in Págico are shown in table 8.7, for each run. In the same table, we present the results of the best human participation (actually, a groups of participants), ludIT (Veiga et al., 2012), which show that we are still very far from a human approach to this task, and we show the results of the best run of RENOIR. Performance is given by the following measures:

- Answered topics: number of topics with at least one given answer.
- Given answers: total number of given answers.

- Precision: proportion of correct answers, where, if needed, the support of each answer is considered. This was not addressed by Rapportágico.
- Tolerant precision: proportion of correct answers, where the support is ignored.
- Pseudo-recall: proportion of given answers regarding all the relevant answers, using as gold standard the collection of all the answers given by the topic creators plus those given correctly, at least, by one participant.
- Pseudo-F: harmonic mean of precision and pseudo-recall.
- Originality: correct answers given only by this run, and not by any other run/participant nor by the topic creator.
- Score: combines precision with the number of given answers.

Measure	**Rapportágico Runs**			**ludIT**	**RENOIR-1**
	Baseline	Bag-of-words	Pers. PageRank		
Answered topics	116	115	114	150	150
Given answers	1,718	1,736	1,730	1,387	15,000
Precision	10.64%	11.69%	12.02%	76.78%	2.91%
Tolerant precision	11.18%	12.44%	12.77%	79.24%	3.16%
Pseudo-recall	8.05%	9.03%	9.25%	47.35%	19.39%
Pseudo-F	9.13%	10.19%	10.45%	58.58%	5.06%
Originality	22	5	29	3,442	126
Score	19.07	23.74	25.00	817.75	12.67

Table 8.7: Performance of Rapportágico in Págico

As for the unofficial runs with category expansion, the results were between the baseline and the other official runs. But these should not be seen as final, because they were evaluated against the official gold collection, created after the official participation, and thus not considering a few (possibly) correct new answers that were not in the gold collection. Therefore, we do not include them in figure 8.7.

One interesting point on the evaluation of Rapportágico is that, besides our initial intention of improving the recall, the query expansion with Onto.PT also improved precision. This might occur because, more than finding (real) synonyms, Onto.PT provides intimately-related words, which increase the Lucene score of the correct answers.

We would like to, in a near future, run Rapportágico again, with the most recent version of Onto.PT, in order to see if there was any progress. Furthermore, besides expanding VPs and categories into synonyms, it would be interesting to test Rapportágico with the expansion of categories into their hyponyms. For instance, if the request of a topic was "musicians", this word could be expanded into words as `composer`, `singer`, `guitarrist`, `pianist`, `drummer` and other kinds of musicians.

8.4.4 Answering cloze questions

For English, in the late 1990s and early 2000s, attention has been given to the development of automatic approaches for answering synonymy questions from the Test of English as a Foreign Language (TOEFL)[9]. The proposed methods included corpus-

[9]The state of the art for this task is presented in `http://aclweb.org/aclwiki/index.php?title=TOEFL_Synonym_Questions_(State_of_the_art)` (September 2012)

based approaches (e.g. Landauer and Dumais (1997); Turney (2001)), lexicon-based approaches (e.g. Jarmasz and Szpakowicz (2003)) and the combination of both (e.g. Turney et al. (2003)). After presenting the results of a combined approach (97.5% accuracy), Turney et al. (2003) claimed that this problem was solved for English, but there were other interesting problems to solve, such as analogies.

This kind of tasks are interesting scenarios for testing the usability of lexical-semantic resources, such as Onto.PT or CARTÃO. However, for Portuguese, the most similar exercises we could find with enough examples and ready for being computationally processed, were cloze questions, also known as fill-in-the-blank questions. These questions, illustrated below, consist of: (i) a sentence where one word is missing (stem); (ii) a shuffled list of words including the missing word and a set of distractors. The goal is to select the correct alternative from the list.

Houve influência da _____ oriental sobre a grega, porém não se pode superestimar a importância dessa influência.

(a) cultura (c) praticante
(b) exibição (d) inteligência

REAP.PT (Silva et al., 2012a) is a computer assisted language learning tutoring system that aims at teaching vocabulary to learners of European Portuguese. Cloze questions, created in the scope of the aforementioned project, were kindly provided by its developers. These questions were generated from sentences of the CETEMPúblico corpus (Rocha and Santos, 2000; Santos and Rocha, 2001), with candidate stems selected from the Portuguese Academic Word List (P-AWL) (Baptista et al., 2010). Both the selection of stems (Correia et al., 2010) and the selection of distractors (Correia et al., 2012) were automatically refined to be in accordance. For instance, lexical resources as PAPEL were used to find (and replace) distractors that could be syononyms of the correct answer. The cloze question shown earlier is one of the 3,900 cloze questions we have used. Its correct answer is in bold.

In order to answer the cloze questions automatically, we implemented several algorithms that take advantage either of a lexical network, or of a wordnet. The algorithms that resulted in the best results use the LKB as a graph and are based on the PageRank algorithm (Brin and Page, 1998). The first algorithm exploits a lexical network, where the nodes are lexical items. It works as follows:

1. POS-tag the original sentence, with the correct answer in the blank, so that it is coherent. After tagging, remove the answer from the sentence.
2. PageRank the network, with initial weights uniformly distributed to the context words. The rest of the nodes have initial weights = 0.
3. Select the alternative answer with the highest rank.

The second algorithm uses a wordnet, where the nodes are synsets:

1. POS-tag the original sentence, with the correct answer in the blank, so that it is coherent. After tagging, remove the answer from the sentence.
2. Run the Personalized PageRank WSD algorithm (see section 8.4.2) using all content words of the sentence as context. This means that the initial weights are uniformly distributed to all synsets with, at least, one context word. The rest of the nodes have initial weights = 0.
3. For each alternative answer, retrieve all candidate synsets.
4. Select the alternative in the highest ranked synset.

The first algorithm was run for answering the 3,900 questions with the help of CARTÃO, and also using, independently, the lexical networks that this resource merges – those extracted from DLP (PAPEL), Dicionário Aberto (DA), and Wiktionary.PT. The second algorithm was used for answering the questions with the help of Onto.PT. Table 8.8 shows the accuracy values obtained.

Resource	Nodes	Accuracy
CARTÃO	lexical items	**41.8%**
PAPEL 3.0	lexical items	39.8%
DA	lexical items	36.6%
Wiktionary.PT	lexical items	35.8%
Onto.PT	synsets	37.6%

Table 8.8: Accuracy on answering cloze questions.

These results are clearly higher than the random selection, which is 25%, because there are four possibilities for each question. We should add that our approach is not very sophisticated, and takes advantage only of the question's context and of the LKBs. But there are a few questions with a short context – 355 have less than 8 context words – as well as several questions with named entities – 1,752 contain at least one capitalised word – which are not expected to be found in a LKB. Figure 8.7 is an example of a question with a named entity (with three tokens) and context of five words, which was not answered correctly using any of the resources. On the other hand, figure 8.8 shows a question that was answered correctly using each of the resources.

Mercedes Classe C _____ *carácter desportivo.*

(a) reforça (c) desloca
(b) defende (d) implica

Figure 8.7: Cloze question not answered correctly, using any of the resources.

Despite the poor accuracy of the results obtained, this exercise showed that the organisation of the used LKBs makes sense, and that these resources can be exploited

Uma população (grupo com o mesmo tipo de organismo) pode adaptar-se através da evolução (desenvolvimento gradual) durante muitas _____.

(a) gerações (c) adequações
(b) tradições (d) maiorias

Figure 8.8: Cloze question answered correctly using each resource.

for improving the performance of NLP tasks where it is important to compute the similarity between words or concepts. Moreover, the results are better using CARTÃO and PAPEL than using Onto.PT, which suggests that, for this specific task, a lexical network is more adequate than a wordnet. No strong conclusions can be taken because, even though very similar, the algorithms used are also different. For instance, the algorithm using Onto.PT is more complex and relies on WSD, which can be an additional source of noise.

Another conclusion that emerges from this exercise is that merging knowledge from different resources originates better results. This explains why the accuracy using CARTÃO is higher than using only parts of it, more precisely, PAPEL, DA, or Wiktionary.PT.

Finally, we should add that our results are much lower than the best accuracy for answering TOEFL synonymy questions automatically with the help of a lexicon (78.75%, by Jarmasz and Szpakowicz (2003)). However, the task we have performed is certainly more complex, and cannot be blindly compared. In the future, it would be interesting to try both Onto.PT and CARTÃO in the synonymy problem, analogy problems, and others, as solving cross-words. So far, we did not find usable datasets with such resources for Portuguese, but we will keep looking for them.

Chapter 9

Final discussion

The research described in this thesis is an answer towards our initial goals. It is mainly focused on the creation of a lexical ontology for Portuguese that would minimise the main limitations of existing similar resources. This means that the resulting resource would be public, constructed automatically, created from scratch for Portuguese and structured in word senses.

Having this in mind, this thesis presented the work on the automatic acquisition of lexical-semantic knowledge, and its integration in a unique wordnet-like lexical ontology for Portuguese, dubbed Onto.PT, which can be seen as the materialisation of our goal. Onto.PT can be freely downloaded as a RDF/OWL model or queried through a web interface, both available from `http://ontopt.dei.uc.pt`, together with other resources created in the scope of this research. Besides Onto.PT, each chapter of the thesis described a step towards our final goal, and originates a resource that may be seen as an added value to the range of lexical-semantic resources for Portuguese. The described steps can be combined in ECO, the approach we propose for creating wordnets automatically from text.

We believe that solid steps have been taken, but there is still a long way to go. Onto.PT has already shown to be useful in several NLP tasks and is larger than similar resources, but evaluation showed that there are still reliability issues, and thus room for improvement.

Section 9.1 summarises the main contributions of this research, which include abstract procedures as well as public resources. In the same section, we add information on publications written in the scope of this work, presented in national and international scientific events. Before concluding, we discuss ideas for further work, aiming at the improvement of Onto.PT's reliability and its enrichment with information from additional sources.

9.1 Contributions

The work presented in this thesis resulted in several contributions, especially for the field of the automatic creation of wordnets, and for the state-of-the-art of Portuguese LKBs. We start by enumerating the automatic procedures we have developed for:

1. **Enriching a thesaurus with new synonymy relations** (chapter 6).
2. **Discovering synsets (or fuzzy synsets) from dictionary definitions** (chapter 5).

3. **Moving from term-based to synset-based semantic relations, without using the extraction context** (chapter 7).

Also, even though the procedure for extracting semantic relations from dictionaries cannot be seen as novel, in chapter 4 of this thesis we have presented work on the **comparison of the structure and contents in different dictionaries of Portuguese**. For instance, we have shown that many regularities are kept across the definitions of each dictionary, which enabled us to use the same grammars for extracting information from all the three dictionaries.

Starting with a set of extracted semantic relations, and combining the aforementioned procedures in the appearing order, we proposed ECO, a **flexible approach for creating a wordnet-like lexical ontology** automatically from text. ECO was used for Portuguese but, considering that different methods can be used for the relation extraction step, it is language independent.

During this work, each of the previous procedures was used in the construction of several lexical-semantic resources. These resources, listed below, are **public domain** and may be used together with applications that we hope will contribute for advancing the state-of-the-art of the computational processing of Portuguese:

- CARTÃO: the largest term-based **lexical-semantic network** for Portuguese, larger that PAPEL, which it includes together with relations extracted from two other dictionaries (chapter 4).
- CLIP: the first **fuzzy thesaurus** for Portuguese, completely extracted from dictionaries (chapter 5).
- TRIP: the largest **synset-based thesaurus** for Portuguese, larger than TeP, which it includes together with synonymy information acquired automatically from dictionaries (chapter 6).
- Onto.PT: a new **wordnet-like lexical ontology** for Portuguese, extracted automatically from textual resources that covers more than 100,000 concepts (represented as synsets) and more than 170,000 semantic relations (chapter 8). Currently, Onto.PT contains information from five lexical resources, but the ECO approach enables the future integration of knowledge from other sources, and consequently its future expansion. It is an addition and/or an alternative to existing broad-coverage lexical-semantic resources for Portuguese.

The aforementioned contributions are described in the following scientific publications, presented in national and international events, including some highly selective ones. Together with the description of the publication venue, we present, when available, its acceptance rate and ERA ranking[1]:

- Automatic extraction of semantic relations from Portuguese definitions in collaboratively-created resources – Wikipedia, first, and Wiktionary, second:
  - Gonçalo Oliveira, H., Costa, H., and Gomes, P. (2010a). Extracção de conhecimento léxico-semântico a partir de resumos da Wikipédia. In *Actas do II Simpósio de Informática*, INFORUM 2010, pages 537–548, Braga, Portugal. Universidade do Minho *(40% acceptance rate)*

[1]Conference ranking by the Excellence in Research for Australia, see `http://core.edu.au/index.php/categories/conference\%20rankings/1` (August 2012)

  - Anton Pérez, L., Gonçalo Oliveira, H., and Gomes, P. (2011). Extracting lexical-semantic knowledge from the portuguese wiktionary. In *Proceedings of the 15th Portuguese Conference on Artificial Intelligence*, EPIA 2011, pages 703–717, Lisbon, Portugal. APPIA *(59% acceptance rate for Springer+APPIA proceedings, ERA 2010 ranking B)*

- Creation of the lexical-semantic network CARTÃO:

  - Gonçalo Oliveira, H., Antón Pérez, L., Costa, H., and Gomes, P. (2011). Uma rede léxico-semântica de grandes dimensões para o português, extraída a partir de dicionários electrónicos. *Linguamática*, 3(2):23–38 *(journal on the processing of iberian languages)*

- Clustering for synset discovery from synonymy networks/dictionaries:

  - Gonçalo Oliveira, H. and Gomes, P. (2010a). Automatic Creation of a Conceptual Base for Portuguese using Clustering Techniques. In *Proceedings of 19th European Conference on Artificial Intelligence (ECAI 2010)*, pages 1135–1136, Lisbon, Portugal. IOS Press *(40% acceptance rate for long+short papers, ERA 2010 ranking A)*
  - Gonçalo Oliveira, H. and Gomes, P. (2011a). Automatic Discovery of Fuzzy Synsets from Dictionary Definitions. In *Proceedings of 22nd International Joint Conference on Artificial Intelligence*, IJCAI 2011, pages 1801–1806, Barcelona, Spain. IJCAI/AAAI *(30% acceptance rate, ERA 2010 ranking A)*

- Enrichment of a thesaurus with synonymy relations extracted from text:

  - Gonçalo Oliveira, H. and Gomes, P. (2011b). Automatically enriching a thesaurus with information from dictionaries. In *Progress in Artificial Intelligence, Proceedings of 15th Portuguese Conference on Artificial Intelligence (EPIA 2011)*, volume 7026 of *LNCS*, pages 462–475, Lisbon, Portugal. Springer *(25% acceptance rate for Springer proceedings, ERA 2010 ranking B)*
  - Gonçalo Oliveira, H. and Gomes, P. ((submitted on) 2012b). Towards the Automatic Enrichment of a Thesaurus with Information in Dictionaries. *Expert Systems: The Journal of Knowledge Engineering (Indexed in the ISI Web of Knowledge, impact factor: 1.231 (2009), 0.684 (2011). Decision of the first revision phase: "minor revisions")*

- Algorithms for moving from term-based to synset-based relations (ontologising), without using the extraction context:

  - Gonçalo Oliveira, H. and Gomes, P. (2011c). Ontologising relational triples into a portuguese thesaurus. In *Proceedings of the 15th Portuguese Conference on Artificial Intelligence*, EPIA 2011, pages 803–817, Lisbon, Portugal. APPIA *(59% acceptance rate for Springer+APPIA proceedings, ERA 2010 ranking B)*
  - Gonçalo Oliveira, H. and Gomes, P. (2012a). Ontologising semantic relations into a relationless thesaurus. In *Proceedings of 20th European Conference on Artificial Intelligence (ECAI 2012)*, pages 915–916, Montpellier, France. IOS Press *(28% acceptance rate overall, 32% for short papers, ERA 2010 ranking A)*

- Early stages of Onto.PT:

  - Gonçalo Oliveira, H. (2009). Ontology learning for Portuguese. In *2nd Doctoral Symposium on Artificial Intelligence*, SDIA 2009, pages 21–30, Aveiro, Portugal *(Doctoral symposium)*
  - Gonçalo Oliveira, H. and Gomes, P. (2010c). Towards the automatic creation of a wordnet from a term-based lexical network. In *Proceedings of the ACL Workshop TextGraphs-5: Graph-based Methods for Natural Language Processing*, pages 10–18, Uppsala, Sweden. ACL Press *(∼60% acceptance rate)*
  - Gonçalo Oliveira, H. and Gomes, P. (2010b). Onto.PT: Automatic Construction of a Lexical Ontology for Portuguese. In *Proceedings of 5th European Starting AI Researcher Symposium (STAIRS 2010)*, volume 222 of *Frontiers in Artificial Intelligence and Applications*, pages 199–211. IOS Press *(ECAI 2010 collocated event, ∼50% acceptance rate, proceedings published as an IOS Frontiers in Artificial Intelligence volume)*

- Status of Onto.PT at the time:

  - Gonçalo Oliveira, H. and Gomes, P. (2011). Onto.PT: Construção automática de uma ontologia lexical para o português. In Luís, A. R., editor, *Estudos de Linguística*, volume 1, pages 161–180. Coimbra University Press, Coimbra *(published as a volume chapter by Universidade de Coimbra Press)*
  - Gonçalo Oliveira, H., Pérez, L. A., and Gomes, P. (2012c). Exploring Onto.PT. In *Demo Session of PROPOR 2012, 10th International Conference on the Computational Processing of the Portuguese Language*, Coimbra, Portugal
  - Gonçalo Oliveira, H., Antón Pérez, L., and Gomes, P. (2012a). Integrating lexical-semantic knowledge to build a public lexical ontology for Portuguese. In *Natural Language Processing and Information Systems, Proceedings of 17h International Conference on Applications of Natural Language to Information Systems (NLDB)*, volume 7337 of *LNCS*, pages 210–215, Groningen, The Netherlands. Springer *(39% acceptance rate for long+short papers, ERA 2010 ranking C)*

- Participation in the information retrieval task Págico, where Onto.PT was used for gathering synonyms:

  - Rodrigues, R., Gonçalo Oliveira, H., and Gomes, P. (2012). Uma abordagem ao Págico baseada no processamento e análise de sintagmas dos tópicos. *Linguamática*, 4(1):31–39 *(special issue of journal on the processing of iberian languages)*

9.2 Future work

We believe that important steps were given on the automatic creation and integration of lexical-semantic resources for Portuguese. However, as an automatic approach, there is still much room for improvement. We see the current version of Onto.PT as the first result of the application of the ECO approach, which can be further improved. There are many ideas that could not be explored during the period of this research. Some of them would complement it as well as its outputs.

In this section, we discuss some ideas that might be tackled in the future. Given that the results of this work are public and may be used at will, and that the followed methodology has been extensively described here, we see this discussion as

more than cues, not only for our future work, but also for others'. We sincerely hope that other researchers can use the resources we made available, and either use ECO to create new resources or to enrich the existing ones. Alternatively, they can give us feedback on their experience, which might as well lead to further improvements.

Regarding the flexibility of the ECO approach, we will devise the integration of other sources of knowledge in Onto.PT. A serious candidate is the Portuguese version of the collaborative encyclopedia Wikipedia, with which we have already made some preliminary work on the extraction of semantic relations (Gonçalo Oliveira et al., 2010a). However, in opposition to the previous work, next time we will probably not use handcrafted rules for extracting semantic relations from corpora text. Given the diversity of this kind of text, the idea is to follow a weakly supervised approach, similar to Espresso's (Pantel and Pennacchiotti, 2006), where CARTÃO relations in which we have higher confidence might be used as seeds.

Nevertheless, as most entries of an encyclopedia describe knowledge that is usually not present in dictionaries, instead of merging all kinds of extractions, we will devise the creation of a new layer in Onto.PT, with encyclopedic and world knowledge. It would also be interesting to find a possible mapping between Onto.PT and existing ontologies with that kind of knowledge, as DBPedia (Bizer et al., 2009), using frameworks that connect lexical and world knowledge, such as Lemon (Buitelaar et al., 2009). This way, it would be possible to connect Onto.PT to a huge quantity of linked data, which would enable several knowledge discovery tasks, including cross-lingual WSD and IR.

Despite the problems on the automatic translation of lexical-semantic resources from one language to another, machine translation of the knowledge in Princeton WordNet and knowledge bases in other languages can also be see as an additional source of information or, at least, additional hints, that may be used in the enrichment of Onto.PT. If Onto.PT is originally created for Portuguese, the problems typically related to the translation of lexical-semantic resources would be minimised. Moreover, the obtained information could be used together with other sources to compute the confidence of the knowledge encoded in Onto.PT.

One important limitation of Onto.PT is that its synsets do not contain glosses. Besides involving too much labour, the manual creation of glosses would not be practical because Onto.PT is not a static resource. If the glosses were ok for a certain instantiation of Onto.PT, inconsistencies would probably occur for other versions. Therefore, it would be interesting to automatically associate definitions from dictionaries, by matching synsets with dictionary entries, as Henrich et al. (2011) did for associating Wiktionary definitions with GermaNet (Kunze and Lemnitzer, 2002) synsets. As most of the knowledge in Onto.PT is acquired exactly from dictionaries, an alternative would be to collect definitions during the extraction step. This would however not be completely straightforward because all extracted knowledge is later merged, leading to a rearrangement of the covered concepts and their lexicalisations, and to potential inconsistencies.

The exploitation of a resource as Wiktionary for the acquisition of glosses for Onto.PT synsets could yet go further. For instance, as Henrich et al. (2012) also did, it could be used to obtain example sentences for contextualising the words in synsets. These sentences could then be used in the creation of a sense annotated corpus, similar to SemCor (Miller et al., 1994), a corpus where some words are annotated with the Princeton WordNet synset corresponding to their meaning.

Anyway, before growing in terms of new relations and exploited resources, Onto.PT will probably shrink, as we will try to minimise some problems, including incorrect extractions, that currently add noise to its contents. As it is generated by an automatic approach, and although it is very large, broad-coverage, and shown to be useful, Onto.PT is still far from being highly reliable. Therefore, some directions should be taken in order to improve its quality.

The manual evaluation of the extracted semantic relations is an important source for identifying specific problems, which might lead to future changes in the extraction grammars or in the filters applied after extraction. We can also exploit other sources of information, in order to compute the confidence of the extracted semantic relations. A common approach for this task relies on the application of similarity measures, based on the the occurrences of the relation arguments in corpora, or in the Web (Downey et al., 2005; Costa et al., 2011). These approaches could yet be combined with other kinds of information, including the frequency of the extracted relation, the confidence on the resource or method that lead to its extraction [2], and, as referred earlier, the occurrence of a relation in some resource in other language, after its translation. In a similar fashion to Wandmacher et al. (2007)'s work, this could lead to an integrated confidence.

Besides other benefits, a confidence value would enable the integration of only relations for which confidence is above a predefined threshold. This threshold could be an additional parameter to study in an extensive evaluation of ECO and Onto.PT. Such an evaluation would compare the impact of different parameters, including, but not limited to, the clustering thresholds and the similarity measures used in ontologisation. The idea would be to select different parameters to create different versions of the resource and then compare properties like covered lexical items and semantic relations, sense granularity or synset size. Another parameter could be a threshold on the corpus frequency of the integrated lexical items. Since much of the covered knowledge is extracted from dictionaries, Onto.PT contains some unfrequent, and possibly less useful, words. Besides being of no use for several applications, those words might as well work as additional sources of noise.

One final mention should be given to the adoption of the new spelling reform of Portuguese, agreed by the governments of the Portuguese speaking countries in 1990, but only started to be implemented in 2009. This reform aims to unify the orthography of the European and Brazilian variants of Portuguese. In Onto.PT, however, we have not adopted this reform because:

- All resources we have exploited are not yet converted. As some of them are written in the European variant of Portuguese, others in the Brazilian, and others in both, most of the written variations are covered, as well as some dropped forms;
- The transition period, where using dropped written forms is tolerated, is still going on in most of the countries. In Portugal, it ends in 2015.
- There is still a huge debate going on the adoption of this reform, and on its real benefits for Portuguese;

Nevertheless, we believe that an eventual conversion of Onto.PT to the new spelling

[2]Given that a thesaurus as TeP is created manually, there is certainly more confidence on a relation acquired directly from it than one extracted from text, by an automatic procedure.

reform would be quite straightforward, and consist of the application of rules based on regular expressions, in order to identify older forms and update their orthography.

9.3 Concluding remarks

We would like to conclude this thesis by reaffirming our expectations that Onto.PT and its further versions will be an important contribution to the computational processing of Portuguese. As referred in the previous section, there is plenty of work to do, and we sincerely hope that this project has the deserved continuity. However, there are aspects that do not depend solely on us.

We would also like to mention that it has always been our intention to work on Portuguese that is not only our language, in which we are proud to work on, but also one of the most spoken languages all over the world. If we, the native speakers, do nothing for our mother tongue, who will do?

It should however be stressed that it is very challenging to work on a non-English language. Despite important contributions to the development of computational tools and resources (hereafter, material) for the computational processing of Portuguese (e.g. by Linguateca[3]) there are still fewer, especially public, materials (e.g. taggers, parsers, several types of gold standards to be used as benchmarks ...) for Portuguese than for other languages, as English. This means that, when someone is willing to develop some system with Portuguese NLP capabilities, they have either a very limited choice of materials, or none, and ends up developing their own. Besides being time-consuming, the development of a new material usually requires its evaluation which, without benchmarks, has either to be done manually or after the creation of a gold standard for that task, also a time-consuming task. Moreover, when performing evaluation based on any of the previous, it is usually not possible to compare the obtained results directly against other approaches, reported in the literature. Therefore, we appeal to researchers and developers in Portuguese NLP to make their materials available, so that they can be used by others and thus enable the development of more complete systems, built on the top of available materials. More collaboration and less competition is the way to follow, in order to improve Portuguese NLP.

A consequent issue regarding the work on a non-English language is that, most of the times, we have to spend time creating materials that are already available for English, but usually not language independent, or with too much assumptions on the target language (e.g. the existence of other language-specific materials). Although the newly created materials are most of the times important contributions to the non-English language, they are often not seen as such by international reviewers. By not giving it its deserved value, these reviewers make it harder to publish our results in important scientific venues. Therefore, we conclude this thesis with a second appeal, this time to the scientific community, which might not always understand the aforementioned fact.

[3]See `http://www.linguateca.pt` (September 2012)

References

Agichtein, E. and Gravano, L. (2000). Snowball: Extracting relations from large plain-text collections. In *Proceedings of 5th ACM International Conference on Digital Libraries*, pages 85–94, New York, NY, USA. ACM Press.

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., and Soroa, A. (2009a). A study on similarity and relatedness using distributional and WordNet-based approaches. In *Proceedings Human Language Technologies: 2009 Annual Conference of the North American Chapter of ACL (NAACL-HLT)*, pages 19–27, Stroudsburg, PA, USA. ACL Press.

Agirre, E., Lacalle, O. L. D., and Soroa, A. (2009b). Knowledge-based WSD on specific domains: performing better than generic supervised WSD. In *Proceedings of 21st International Joint Conference on Artificial Intelligence*, IJCAI 2009, pages 1501–1506, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Agirre, E. and Soroa, A. (2009). Personalizing PageRank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL'09, pages 33–41, Stroudsburg, PA, USA. ACL Press.

Alshawi, H. (1987). Processing dictionary definitions with phrasal pattern hierarchies. *Computational Linguistics*, 13(3-4):195–202.

Alshawi, H. (1989). Analysing the dictionary definitions. *Computational lexicography for natural language processing*, pages 153–169.

Amancio, M. A., Watanabe, W. M., Jr., A. C., de Oliveira, M., Pardo, T. A. S., Fortes, R. P. M., and Alusio, S. M. (2010). Educational FACILITA: helping users to understand textual content on the Web. In *Extended Activities Proceedings of the 9th International Conference on Computational Processing of Portuguese Language (PROPOR)*, Porto Alegre/RS, Brazil.

Amsler, R. A. (1980). *The structure of the Merriam-Webster Pocket dictionary.* PhD thesis, The University of Texas at Austin.

Amsler, R. A. (1981). A taxonomy for English nouns and verbs. In *Proceedings of 19th annual meeting on Association for Computational Linguistics*, ACL'81, pages 133–138, Morristown, NJ, USA. ACL Press.

Anton Pérez, L., Gonçalo Oliveira, H., and Gomes, P. (2011). Extracting lexical-semantic knowledge from the portuguese wiktionary. In *Proceedings of the 15th Portuguese Conference on Artificial Intelligence*, EPIA 2011, pages 703–717, Lisbon, Portugal. APPIA.

Baeza-Yates, R. A. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval.* Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet project. In *Proceedings of the 17th international conference on Computational Linguistics*, pages 86–90, Morristown, NJ, USA. ACL Press.

Banerjee, S. and Pedersen, T. (2002). An adapted Lesk algorithm for word sense disambiguation using WordNet. In *Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2002)*, volume 2276 of *LNCS*, pages 136–145, London, UK. Springer.

Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., and Etzioni, O. (2007). Open information extraction from the Web. In *Proceedings of the International Joint Conference on Artificial Intelligence*, IJCAI 2007, pages 2670–2676.

Baptista, J., Costa, N., Guerra, J., Zampieri, M., Cabral, M., and Mamede, N. J. (2010). P-AWL: Academic Word List for Portuguese. In *Proceedings of Computational Processing of the Portuguese Language, 9th International Conference, PROPOR 2010*, volume 6001 of *LNCS*, pages 120–123. Springer.

Barriere, C. (1997). *From a children's first dictionary to a lexical knowledge base of Conceptual Graphs*. PhD thesis, Simon Fraser University, Burnaby, BC, Canada.

Bellare, K., Sharma, A. D., Sharma, A. D., Loiwal, N., and Bhattacharyya, P. (2004). Generic text summarization using WordNet. In *Proceedings of 4th International Conference on Language Resources and Evaluation*, LREC 2004, pages 691–694, Barcelona, Spain. ELRA.

Berger, A. L., Pietra, V. J. D., and Pietra, S. A. D. (1996). A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

Berland, M. and Charniak, E. (1999). Finding parts in very large corpora. In *Proceedings of 37th annual meeting of the Association for Computational Linguistics*, pages 57–64, Morristown, NJ, USA. ACL Press.

Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The Semantic Web. *Scientific American*, pages 34–43.

Bezdek, J. C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*. Kluwer Academic Publishers, Norwell, MA, USA.

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S. (2009). DBpedia - a crystallization point for the web of data. *Web Semantics*, 7(3):154–165.

Blum, A. and Mitchell, T. M. (1998). Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, COLT 1998, pages 92–100, Madison, Wisconsin, USA. ACM Press.

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., and Taylor, J. (2008). Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, SIGMOD 2008, pages 1247–1250, Vancouver, Canada. ACM Press.

Borin, L. and Forsberg, M. (2010). From the people's synonym dictionary to fuzzy synsets - first steps. In *Proceedings of LREC 2010 workshop on Semantic relations. Theory and Applications*, pages 18–25, La Valleta, Malta.

Brank, J., Grobelnik, M., and Mladenic', D. (2005). A survey of ontology evaluation techniques. In *Proceedings of Conference on Data Mining and Data Warehouses*, SiKDD 2005, pages 166–170.

Brewster, C. and Wilks, Y. (2004). Ontologies, taxonomies, thesauri: Learning from texts. In Deegan, M., editor, *Proceedings of The Use of Computational Linguistics in the Extraction of Keyword Information from Digital Library Content Workshop*, London, UK. Centre for Computing in the Humanities, Kings College.

Brin, S. (1998). Extracting patterns and relations from the World Wide Web. In *Proceedings of 6th International Conference on Extending Database Technology, EDBT'98*, pages 172–183, London, UK. Springer.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks*, 30(1-7):107–117.

Briscoe, T. (1991). Lexical issues in natural language processing. In *Proceedings of Natural Language and Speech Symposium*, pages 39–68. Springer.

Broekstra, J., Kampman, A., and van Harmelen, F. (2002). Sesame: A generic architecture

for storing and querying RDF and RDF Schema. In *The Semantic Web – ISWC 2002: First International Semantic Web Conference*, volume 2342 of *LNCS*, pages 54–68. Springer, Sardinia, Italy.

Bruce, R. and Guthrie, L. (1992). Genus disambiguation: A study in weighted preference. In *Proceedings of the 14th conference on Computational Linguistics*, COLING'92, pages 1187–1191, Nantes, France. ACL Press.

Buitelaar, P., Cimiano, P., Haase, P., and Sintek, M. (2009). Towards linguistically grounded ontologies. In *Proceedings of the 6th European Semantic Web Conference on The Semantic Web: Research and Applications*, ESWC 2009, pages 111–125, Heraklion, Crete, Greece. Springer.

Calzolari, N. (1977). An empirical approach to circularity in dictionary definitions. In *Cahiers de Lexicologie*, pages 118–128.

Calzolari, N. (1984). Detecting patterns in a lexical data base. In *Proceedings of the 22nd annual meeting on Association for Computational Linguistics*, pages 170–173, Morristown, NJ, USA. ACL Press.

Calzolari, N., Pecchia, L., and Zampolli, A. (1973). Working on the italian machine dictionary: a semantic approach. In *Proceedings of 5th Conference on Computational Linguistics*, COLING'73, pages 49–52, Morristown, NJ, USA. ACL Press.

Caraballo, S. A. (1999). Automatic construction of a hypernym-labeled noun hierarchy from text. In *Proceedings of 37th annual meeting of the Association for Computational Linguistics*, pages 120–126, Morristown, NJ, USA. ACL Press.

Cardoso, N. (2012). Medindo o precipício semântico. *Linguamática*, 4(1):41–48.

Carletta, J. (1996). Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2):249–254.

Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Estevam R. Hruschka Jr, and Mitchell, T. M. (2010a). Toward an architecture for Never-Ending Language Learning. In *Proceedings of 24th Conference on Artificial Intelligence*, AAAI 2010. AAAI Press.

Carlson, A., Betteridge, J., Wang, R. C., Hruschka Jr., E. R., and Mitchell, T. M. (2010b). Coupled semi-supervised learning for information extraction. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining*, WSDM 2010, pages 101–110, New York, NY, USA. ACM Press.

Cederberg, S. and Widdows, D. (2003). Using LSA and noun coordination information to improve the precision and recall of automatic hyponymy extraction. In *Proceedings of 7th Conference on Computational Natural Language Learning*, CoNLL 2003, pages 111–118, Morristown, NJ, USA. ACL Press.

Chein, M. and Mugnier, M.-L. (2008). *Graph-Based Knowledge Representation: Computational Foundations of Conceptual Graphs*. Advanced Information and Knowledge Processing. Springer, 1st edition.

Chinchor, N. and Robinson, P. (1997). MUC-7 named entity task definition. In *Proceedings of 7th Message Understanding Conference*, MUC-7.

Chodorow, M. S., Byrd, R. J., and Heidorn, G. E. (1985). Extracting semantic hierarchies from a large on-line dictionary. In *Proceedings of the 23rd annual meeting on Association for Computational Linguistics*, ACL'85, pages 299–304, Morristown, NJ, USA. ACL Press.

Church, K. W. and Hanks, P. (1989). Word association norms, mutual information and lexicography. In *Proceedings of 27th Annual Meeting of the Association for Computational Linguistics*, pages 76–83, Vancouver, BC, Canada. ACL Press.

Cimiano, P. and Wenderoth, J. (2007). Automatic acquisition of ranked qualia structures from the Web. In *Proceedings of 45th Annual Meeting of the Association for Computational Linguistics*, pages 888–895. ACL Press.

Clark, P., Fellbaum, C., and Hobbs, J. (2008). Using and extending WordNet to support question-answering. In *Proceedings of 4th Global WordNet Conference*, GWC 2008, pages 111–119, Szeged, Hungary.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37.

Correia, R., Baptista, J., Eskenazi, M., and Mamede, N. (2012). Automatic generation of cloze question stems. In *Proceedings of the 10th International Conference on Computational Processing of the Portuguese Language (PROPOR 2012)*, volume 7243 of *LNCS*, pages 168–178, Coimbra, Portugal. Springer.

Correia, R., Baptista, J., Mamede, N., Trancoso, I., and Eskenazi, M. (2010). Automatic generation of cloze question distractors. In *Second Language Studies: Acquisition, Learning, Education and Technology*, Tokyo, Japan.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

Costa, H. (2011). O desenho do novo Folheador. Technical report, Linguateca.

Costa, H., Gonçalo Oliveira, H., and Gomes, P. (2011). Using the Web to validate lexico-semantic relations. In *Progress in Artificial Intelligence, Proceedings of 15th Portuguese Conference on Artificial Intelligence (EPIA 2011)*, volume 7026 of *LNCS*, pages 597–609. Springer, Lisbon, Portugal.

Cruse, D. A. (1986). *Lexical Semantics.* Cambridge University Press, Cambridge University Press.

de Melo, G. and Weikum, G. (2008). On the utility of automatically generated wordnets. In *Proceedings of 4th Global WordNet Conference*, GWC 2008, pages 147–161, Szeged, Hungary. University of Szeged.

de Paiva, V. and Rademaker, A. (2012). Revisiting a brazilian wordnet. In *Proceedings of 6th Global Wordnet Conference*, GWC 2012, Matsue, Japan. Tribun EU, Brno.

de Souza, M. M. (2010). Análise lexicográfica na FrameNet Brasil. *Gatilho*, 11(4).

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41:391–407.

Defays, D. (1977). An efficient algorithm for a complete link method. *The Computer Journal*, 20(4):364–366.

Dias-Da-Silva, B. C. and de Moraes, H. R. (2003). A construção de um thesaurus eletrônico para o português do Brasil. *ALFA*, 47(2):101–115.

Dias da Silva, B. C., de Oliveira, M. F., and de Moraes, H. R. (2002). Groundwork for the Development of the Brazilian Portuguese Wordnet. In *Advances in Natural Language Processing (PorTAL 2002)*, LNAI, pages 189–196, Faro, Portugal. Springer.

Dias-da Silva, B. C., Di Felippo, A., and Hasegawa, R. (2006). Methods and tools for encoding the wordnet.br sentences, concept glosses, and conceptual-semantic relations. In *Proceedings of the 7th International conference on Computational Processing of the Portuguese Language (PROPOR 2006)*, volume 3960 of *LNCS*, pages 120–130, Itatiaia, Brazil. Springer.

Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische Mathematik*, 1:269–271.

DLP (2005). *Dicionário PRO da Língua Portuguesa.* Porto Editora, Porto.

Dolan, W., Vanderwende, L., and Richardson, S. D. (1993). Automatically deriving structured knowledge bases from online dictionaries. In *Proceedings of the 1st Conference of the Pacific Association for Computational Linguistics*, PACLING'93, pages 5–14.

Dolan, W. B. (1994). Word sense ambiguation: clustering related senses. In *Proceedings of*

15th International Conference on Computational Linguistics, COLING'94, pages 712–716, Morristown, NJ, USA. ACL Press.

Dorow, B. (2006). *A Graph Model for Words and their Meanings*. PhD thesis, Institut fur Maschinelle Sprachverarbeitung der Universitat Stuttgart.

Downey, D., Etzioni, O., and Soderland, S. (2005). A probabilistic model of redundancy in information extraction. In *Proceedings of the 19th International Joint Conference on Artificial intelligence*, IJCAI 2005, pages 1034–1041, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Edmonds, P. and Hirst, G. (2002). Near-synonymy and lexical choice. *Computational Linguistics*, 28(2):105–144.

Elberrichi, Z., Rahmoun, A., and Bentaalah, M. A. (2006). Using WordNet for text categorization. *International Arab Journal of Information Technology*, 5(1):3–37.

Esuli, A. and Sebastiani, F. (2007). PageRanking WordNet synsets: An application to opinion mining. In *Proceedings of 45th Annual Meeting of the Association of Computational Linguistics*, ACL'07, pages 424–431. ACL Press.

Etzioni, O., Cafarella, M. J., Downey, D., Kok, S., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D. S., and Yates, A. (2004). Web-scale information extraction in KnowItAll: (preliminary results). In *Proceedings of the 13th international conference on World Wide Web*, WWW 2004, pages 100–110, New York, NY, USA.

Etzioni, O., Fader, A., Christensen, J., Soderland, S., and Mausam (2011). Open information extraction: The second generation. In *Proceedings of 22nd International Joint Conference on Artificial Intelligence*, IJCAI 2011, pages 3–10, Barcelona, Spain. IJCAI/AAAI.

Fader, A., Soderland, S., and Etzioni, O. (2011). Identifying relations for open information extraction. In *Proceedings of the Conference of Empirical Methods in Natural Language Processing*, EMNLP 2011, Edinburgh, Scotland, UK. ACL Press.

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

Fellbaum, C. (2010). WordNet. In *Theory and Applications of Ontology: Computer Applications*, chapter 10, pages 231–243. Springer.

Ferreira, L., Teixeira, A., and da Silva Cunha, J. P. (2008). REMMA - Reconhecimento de entidades mencionadas do MedAlert. In Mota, C. and Santos, D., editors, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas*, pages 213–229. Linguateca.

Fillmore, C. J. (1982). Frame semantics. In *Linguistics in the morning calm*. Seoul: Hanshin Publishing Co.

Francopoulo, G., Bel, N., George, M., Calzolari, N., Monachini, M., Pet, M., and Soria, C. (2009). Multilingual resources for NLP in the Lexical Markup Framework (lmf). *Language Resources and Evaluation*, 43(1):57–70.

Freitas, C., Santos, D., Gonçalo Oliveira, H., and Quental, V. (2012). VARRA: Validação, Avaliação e Revisão de Relações semânticas no AC/DC. In *Livro do IX Encontro de Linguística de Corpus*, ELC 2010, page In press, Rio Grande do Sul, Brasil.

Freitas, C., Santos, D., Mota, C., Gonçalo Oliveira, H., and Carvalho, P. (2009). Detection of relations between named entities: report of a shared task. In *Proceedings of NAACL-HLT, Semantic Evaluations: Recent Achievements and Future Directions Workshop*, SEW 2009, Boulder, Colorado. ACL Press.

Freitas, M. C. (2007). *Elaboração automática de ontologias de domínio: discussão e resultados*. PhD thesis, Pontifícia Universidade Católica do Rio de Janeiro.

Gale, W. A., Church, K. W., and Yarowsky, D. (1992). One sense per discourse. In *Proceedings of the HLT'91 workshop on Speech and Natural Language*, pages 233–237, Morristown, NJ, USA. ACL Press.

Gangemi, A., Guarino, N., Masolo, C., and Oltramari, A. (2010). Interfacing WordNet with DOLCE: towards OntoWordNet. In *Ontology and the Lexicon: A Natural Language Processing Perspective*, Studies in Natural Language Processing, chapter 3, pages 36–52. Cambridge University Press.

Geeraerts, D. (2010). *Theories of Lexical Semantics (Oxford Linguistics)*. Oxford University Press.

Gfeller, D., Chappelier, J.-C., and De Los Rios, P. (2005). Synonym Dictionary Improvement through Markov Clustering and Clustering Stability. In *Proceedings of International Symposium on Applied Stochastic Models and Data Analysis*, ASMDA 2005, pages 106–113, Brest, France.

Girju, R., Badulescu, A., and Moldovan, D. (2003). Discovery of manner relations and their applicability to question answering. In *Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering*, pages 54–60, Morristown, NJ, USA. ACL Press.

Girju, R., Badulescu, A., and Moldovan, D. (2006). Automatic discovery of part-whole relations. *Computational Linguistics*, 32(1):83–135.

Girju, R. and Moldovan, D. (2002). Text mining for causal relations. In *Proceedings of 15th International Florida Artificial Intelligence Research Society Conference*, FLAIRS 2002, pages 360–364.

Gomes, P., Pereira, F. C., Paiva, P., Seco, N., Carreiro, P., Ferreira, J. L., and Bento, C. (2003). Noun Sense Disambiguation with WordNet for Software Design Retrieval. In *Proceedings of Advances in Artificial Intelligence, 16th Conference of the Canadian Society for Computational Studies of Intelligence*, pages 537–543, Halifax, Canada.

Gonçalo Oliveira, H. (2009). Ontology learning for Portuguese. In *2nd Doctoral Symposium on Artificial Intelligence*, SDIA 2009, pages 21–30, Aveiro, Portugal.

Gonçalo Oliveira, H. (2012). PoeTryMe: a versatile platform for poetry generation. In *Proceedings of the ECAI 2012 Workshop on Computational Creativity, Concept Invention, and General Intelligence*, C3GI 2012, Montpellier, France.

Gonçalo Oliveira, H., Antón Pérez, L., Costa, H., and Gomes, P. (2011). Uma rede léxico-semântica de grandes dimensões para o português, extraída a partir de dicionários electrónicos. *Linguamática*, 3(2):23–38.

Gonçalo Oliveira, H., Antón Pérez, L., and Gomes, P. (2012a). Integrating lexical-semantic knowledge to build a public lexical ontology for Portuguese. In *Natural Language Processing and Information Systems, Proceedings of 17h International Conference on Applications of Natural Language to Information Systems (NLDB)*, volume 7337 of *LNCS*, pages 210–215, Groningen, The Netherlands. Springer.

Gonçalo Oliveira, H., Costa, H., and Gomes, P. (2010a). Extracção de conhecimento léxico-semântico a partir de resumos da Wikipédia. In *Actas do II Simpósio de Informática*, INFORUM 2010, pages 537–548, Braga, Portugal. Universidade do Minho.

Gonçalo Oliveira, H., Costa, H., and Santos, D. (2012b). Folheador: browsing through Portuguese semantic relations. In *Proceedings of 12th Conference of the European Chapter of the Association for Computational Linguistics (Demos Session)*, EACL 2012, Avignon, France. ACL Press.

Gonçalo Oliveira, H. and Gomes, P. (2010a). Automatic Creation of a Conceptual Base for Portuguese using Clustering Techniques. In *Proceedings of 19th European Conference on Artificial Intelligence (ECAI 2010)*, pages 1135–1136, Lisbon, Portugal. IOS Press.

Gonçalo Oliveira, H. and Gomes, P. (2010b). Onto.PT: Automatic Construction of a Lexical Ontology for Portuguese. In *Proceedings of 5th European Starting AI Researcher*

Symposium (STAIRS 2010), volume 222 of *Frontiers in Artificial Intelligence and Applications*, pages 199–211. IOS Press.

Gonçalo Oliveira, H. and Gomes, P. (2010c). Towards the automatic creation of a wordnet from a term-based lexical network. In *Proceedings of the ACL Workshop TextGraphs-5: Graph-based Methods for Natural Language Processing*, pages 10–18, Uppsala, Sweden. ACL Press.

Gonçalo Oliveira, H. and Gomes, P. (2011a). Automatic Discovery of Fuzzy Synsets from Dictionary Definitions. In *Proceedings of 22nd International Joint Conference on Artificial Intelligence*, IJCAI 2011, pages 1801–1806, Barcelona, Spain. IJCAI/AAAI.

Gonçalo Oliveira, H. and Gomes, P. (2011b). Automatically enriching a thesaurus with information from dictionaries. In *Progress in Artificial Intelligence, Proceedings of 15th Portuguese Conference on Artificial Intelligence (EPIA 2011)*, volume 7026 of *LNCS*, pages 462–475, Lisbon, Portugal. Springer.

Gonçalo Oliveira, H. and Gomes, P. (2011c). Ontologising relational triples into a portuguese thesaurus. In *Proceedings of the 15th Portuguese Conference on Artificial Intelligence*, EPIA 2011, pages 803–817, Lisbon, Portugal. APPIA.

Gonçalo Oliveira, H. and Gomes, P. (2011). Onto.PT: Construção automática de uma ontologia lexical para o português. In Luís, A. R., editor, *Estudos de Linguística*, volume 1, pages 161–180. Coimbra University Press, Coimbra.

Gonçalo Oliveira, H. and Gomes, P. (2012a). Ontologising semantic relations into a relationless thesaurus. In *Proceedings of 20th European Conference on Artificial Intelligence (ECAI 2012)*, pages 915–916, Montpellier, France. IOS Press.

Gonçalo Oliveira, H. and Gomes, P. ((submitted on) 2012b). Towards the Automatic Enrichment of a Thesaurus with Information in Dictionaries. *Expert Systems: The Journal of Knowledge Engineering*.

Gonçalo Oliveira, H., Gomes, P., Santos, D., and Seco, N. (2008). PAPEL: a dictionary-based lexical ontology for Portuguese. In *Computational Processing of the Portuguese Language, 8th International Conference, Proceedings (PROPOR 2008)*, volume 5190, pages 31–40, Aveiro, Portugal. Springer.

Gonçalo Oliveira, H., Pérez, L. A., and Gomes, P. (2012c). Exploring Onto.PT. In *Demo Session of PROPOR 2012, 10th International Conference on the Computational Processing of the Portuguese Language*, Coimbra, Portugal.

Gonçalo Oliveira, H., Santos, D., and Gomes, P. (2009). Relations extracted from a portuguese dictionary: results and first evaluation. In *Proceedings of 14th Portuguese Conference on Artificial Intelligence (EPIA)*, EPIA 2009, pages 541–552. APPIA.

Gonçalo Oliveira, H., Santos, D., and Gomes, P. (2010b). Extracção de relações semânticas entre palavras a partir de um dicionário: o PAPEL e sua avaliação. *Linguamática*, 2(1):77–93.

Green, A. M. (1997). Kappa statistics for multiple raters using categorical classifications. In *Proceedings of the 22nd Annual Conference of SAS Users Group*, San Diego, USA.

Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, Norwell, MA, USA.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220.

Guarino, N. (1998). Formal ontology and information systems. In *Proceedings of the 1st International Conference on Formal Ontologies in Information Systems (FOIS'98)*, pages 3–15. IOS Press.

Gurevych, I., Eckle-Kohler, J., Hartmann, S., Matuschek, M., Meyer, C. M., and Wirth, C. (2012). UBY - a large-scale unified lexical-semantic resource. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*,

EACL 2012, pages 580–590, Avignon, France. ACL Press.

Guthrie, L., Slater, B. M., Wilks, Y., and Bruce, R. (1990). Is there content in empty heads? In *Proceedings of the 13th conference on Computational Linguistics*, volume 3 of *COLING'90*, pages 138–143, Helsinki, Finland. ACL Press.

Harabagiu, S. M. and Moldovan, D. I. (2000). Enriching the wordnet taxonomy with contextual knowledge acquired from text. In *Natural language processing and knowledge representation: language for knowledge and knowledge for language*, pages 301–333. MIT Press, Cambridge, MA, USA.

Harary, F., Norman, R. Z., and Cartwright, D. (1965). *Structural models: an introduction to the theory of directed graphs*. Wiley, New York.

Harris, Z. S. (1968). *Mathematical Structures of Language*. Wiley, New York, NY, USA.

Havasi, C., Speer, R., and Alonso, J. (2009). ConceptNet: A lexical resource for common sense knowledge. In *Selected papers from Recent Advances in Natural Language Processing 2007*, RANLP 2007, pages 269—280. John Benjamins Publishing Co., Borovets, Bulgaria.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of 14th Conference on Computational Linguistics*, COLING 92, pages 539–545, Morristown, NJ, USA. ACL Press.

Hearst, M. A. (1998). Automated Discovery of WordNet Relations. In *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*, pages 131–151. The MIT Press.

Hemayati, R., Meng, W., and Yu, C. (2007). Semantic-based grouping of search engine results using WordNet. In *Proceedings of the joint 9th Asia-Pacific web and 8th international conference on web-age information management Conference on Advances in Data and Web Management*, APWeb/WAIM'07, pages 678–686. Springer.

Hendrickx, I., Kim, S. N., Kozareva, Z., Nakov, P., Séaghdha, D. O., Padó, S., Pennacchiotti, M., Romano, L., and Szpakowicz, S. (2010). Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval 2010, pages 33–38, Stroudsburg, PA, USA. ACL Press.

Henrich, V., Hinrichs, E., and Vodolazova, T. (2011). Semi-automatic extension of germanet with sense definitions from wiktionary. In *Proceedings of 5th Language & Technology Conference*, LTC 2011, pages 126–130, Poznan, Poland.

Henrich, V., Hinrichs, E., and Vodolazova, T. (2012). WebCAGe – a web-harvested corpus annotated with germanet senses. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 387–396, Avignon, France. ACL Press.

Herbelot, A. and Copestake, A. (2006). Acquiring ontological relationships from Wikipedia using RMRS. In *Proceedings of ISWC 2006 Workshop on Web Content Mining with Human Language Technologies*.

Hervás, R., Camara Pereira, F., Gervás, P., and Cardoso, A. (2006). A text generation system that uses simple rhetorical figures. *Procesamiento de Lenguaje Natural*, 37:199–206.

Hirst, G. (2004). Ontology and the lexicon. In Staab, S. and Studer, R., editors, *Handbook on Ontologies*, International Handbooks on Information Systems, pages 209–230. Springer.

Hoffart, J., Suchanek, F. M., Berberich, K., Lewis-Kelham, E., de Melo, G., and Weikum, G. (2011). Yago2: exploring and querying world knowledge in time, space, context, and many languages. In *Proceedings of the 20th International Conference on World Wide Web (Companion Volume)*, WWW 2011, pages 229–232, Hyderabad, India.

Hovy, E., Hermjakob, U., and yew Lin, C. (2001). The use of external knowledge in factoid QA. In *Proceedings of the 10th Text REtrieval Conference*, TREC 2001, pages 644–652.

Ide, N. and Véronis, J. (1995). Knowledge extraction from machine-readable dictionaries: An evaluation. In *Machine Translation and the Lexicon*, volume 898 of *LNAI*. Springer.

Ittoo, A. and Bouma, G. (2010). On learning subtypes of the part-whole relation: Do not mix your seeds. In *Proceedings of 48th Annual Meeting of the Association for Computational Linguistics*, ACL 2010, pages 1328–1336, Uppsala, Sweden. ACL Press.

Jarmasz, M. and Szpakowicz, S. (2003). Roget's thesaurus and semantic similarity. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2003)*, CILT, pages 212–219, Borovets, Bulgaria. John Benjamins, Amsterdam/Philadelphia.

Jing, H. (1998). Usage of WordNet in natural language generation. In *Proceedings of the COLING-ACL'98 workshop on "Usage of WordNet in Natural Language Processing Systems"*, pages 128–134, Quebec, Canada. ACL Press.

Jurafsky, D. and Martin, J. H. (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Prentice Hall series in artificial intelligence. Prentice Hall, Pearson Education International, Englewood Cliffs, NJ, 2nd edition.

Kilgarriff, A. (1996). 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*, pages 193–200.

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

Kozareva, Z. and Hovy, E. (2010). A semi-supervised method to learn and construct taxonomies using the Web. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1110–1118, Stroudsburg, PA, USA. ACL Press.

Kunze, C. and Lemnitzer, L. (2002). GermaNet - representation, visualization, application. In *Proceedings of 3rd International Conference on Language Resources and Evaluation*, LREC 2002, pages 1485–1491, Las Palmas, Spain.

Kwong, O. Y. (1998). Bridging the gap between dictionary and thesaurus. In *Proceedings of 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (Student Papers)*, COLING-ACL'98, pages 1487–1489, Montréal, Quebec, Canada. Morgan Kaufmann Publishers / ACL Press.

Lafferty, J. D., McCallum, A., and Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Lenat, D. B. (1995). CyC: a large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38.

Lesk, M. (1986). Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. In *Proceedings of the 5^{th} Annual International Conference on Systems documentation*, SIGDOC '86, pages 24–26, New York, NY, USA. ACM.

Levin, B. (1993). *English Verb Classes and Alternations A Preliminary Investigation*. University of Chicago Press, Chicago and London.

Lin, D. (1998). Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on Computational linguistics*, COLING'98, pages 768–774, Montreal, Quebec, Canada. ACL Press.

Lin, D. and Pantel, P. (2002). Concept discovery from text. In *Proceedings of 19th International Conference on Computational Linguistics*, COLING 2002, pages 577–583.

Liu, H. and Singh, P. (2004). ConceptNet: A practical commonsense reasoning toolkit. *BT Technology Journal*, 22(4):211–226.

Liu, S., Liu, F., Yu, C., and Meng, W. (2004). An effective approach to document retrieval via utilizing wordnet and recognizing phrases. In *Proceedings of the 27th annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 266–272, New York, NY, USA. ACM Press.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press.

Markowitz, J., Ahlswede, T., and Evens, M. (1986). Semantically significant patterns in dictionary definitions. In *Proceedings of the 24th annual meeting on Association for Computational Linguistics*, ACL'86, pages 112–119, Morristown, NJ, USA. ACL Press.

Marques, C. J. L. (2011). Syntactic REAP.PT. Master's thesis, Instituto Superior Técnico, Lisboa, Portugal.

Marrafa, P. (2001). *WordNet do Português: uma base de dados de conhecimento linguístico*. Instituto Camões.

Marrafa, P. (2002). Portuguese WordNet: general architecture and internal semantic relations. *DELTA*, 18:131–146.

Marrafa, P., Amaro, R., and Mendes, S. (2011). WordNet.PT Global – extending WordNet.PT to Portuguese varieties. In *Proceedings of the 1st Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, pages 70–74, Edinburgh, Scotland. ACL Press.

Matuszek, C., Cabral, J., Witbrock, M., and DeOliveira, J. (2006). An introduction to the syntax and content of CyC. In *Proceedings of the 2006 AAAI Spring Symposium on Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering*.

Maziero, E. G., Pardo, T. A. S., Felippo, A. D., and Dias-da-Silva, B. C. (2008). 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.

McCallum, A., Nigam, K., Rennie, J., and Seymore, K. (1999). A Machine Learning Approach to Building Domain-Specific Search Engines. In *Proceedings of 16th International Joint Conference on Artificial Intelligence*, IJCAI'99, pages 662–667, Stockholm, Sweden. Morgan Kaufmann.

McGuinness, D. L. and van Harmelen, F. (2004). OWL Web Ontology Language overview. Published: W3C Recommendation.

Medelyan, O., Milne, D., Legg, C., and Witten, I. H. (2009). Mining Meaning from Wikipedia. *International Journal of Human-Computer Studies*, 67(9):716–754.

Michiels, A., Mullenders, J., and Noël, J. (1980). Exploiting a large data base by Longman. In *Proceedings of the 8th conference on Computational Linguistics*, COLING'80, pages 374–382, Morristown, NJ, USA. ACL Press.

Miller, E. and Manola, F. (2004). RDF primer. Published: W3C Recommendation.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the*

ACM, 38(11):39–41.

Miller, G. A., Chodorow, M., Landes, S., Leacock, C., and Thomas, R. G. (1994). Using a semantic concordance for sense identification. In *Proceedings of ARPA Human Language Technology Workshop*, Plainsboro, NJ, USA.

Mitkov, R., Evans, R., Orasan, C., Barbu, C., Jones, L., and Sotirova, V. (2000). Coreference and anaphora: developing annotating tools, annotated resources and annotation strategies. In *Proceedings of the Discourse, Anaphora and Reference Resolution Conference*, DAARC 2000, pages 49–58, Lancaster, UK.

Moens, M.-F. (2006). *Information Extraction: Algorithms and Prospects in a Retrieval Context (The Information Retrieval Series)*. Springer-Verlag New York, Secaucus, NJ, USA.

Mohamed, T. P., Hruschka, Jr., E. R., and Mitchell, T. M. (2011). Discovering relations between noun categories. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP 2011, pages 1447–1455, Stroudsburg, PA, USA. ACL Press.

Moldovan, D. I. and Mihalcea, R. (2000). Using WordNet and lexical operators to improve internet searches. *IEEE Internet Computing*, 4(1):34–43.

Montemagni, S. and Vanderwende, L. (1992). Structural patterns vs. string patterns for extracting semantic information from dictionaries. In *Proceedings of the 14th conference on Computational linguistics*, COLING'92, pages 546–552, Morristown, NJ, USA. ACL Press.

Mota, C. and Santos, D., editors (2008). *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca.

Mota, C., Simões, A., Freitas, C., Costa, L., and Santos, D. (2012). Págico: Evaluating Wikipedia-based information retrieval in Portuguese. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, LREC 2012, Istanbul, Turkey. ELRA.

Murphy, M. L. (2003). *Semantic Relations and the Lexicon*. Cambridge University Press.

Naber, D. (2004). Openthesaurus: Building a thesaurus with a Web community. http://www.openthesaurus.de/download/openthesaurus.pdf (retrieved on August 2012).

Nadig, R., Ramanand, J., and Bhattacharyya, P. (2008). Automatic evaluation of WordNet synonyms and hypernyms. In *Proceedings of 6th International Conference on Natural Language Processing*, ICON 2008, Pune, India.

Nakamura, J.-I. and Nagao, M. (1988). Extraction of semantic information from an ordinary English dictionary and its evaluation. In *Proceedings of the 12th conference on Computational linguistics*, COLING'88, pages 459–464, Morristown, NJ, USA. ACL Press.

Nancy Ide, J. V. (1998). Introduction to the special issue on word sense disambiguation: The state of the art. *Computational Linguistics*, 24:1–40.

Nastase, V., Strube, M., Boerschinger, B., Zirn, C., and Elghafari, A. (2010). WikiNet: A very large scale multi-lingual concept network. In *Proceedings of the International Conference on Language Resources and Evaluation*, LREC 2010. ELRA.

Nastase, V. and Szpakowicz, S. (2003). Augmenting WordNet's structure using LDOCE. In *Computational Linguistics and Intelligent Text Processing, 4th International Conference, CICLing*, volume 2588 of *LNCS*, pages 281–294. Springer.

Navarro, E., Sajous, F., Gaume, B., Prévot, L., Hsieh, S., Kuo, T. Y., Magistry, P., and Huang, C. R. (2009). Wiktionary and NLP: Improving synonymy networks. In *Proceedings of Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, pages 19–27, Suntec, Singapore. ACL Press.

Navigli, R. (2009a). Using cycles and quasi-cycles to disambiguate dictionary glosses. In *Proceedings of the 12th Conference on European chapter of the Association for Computational Linguistics*, EACL'09, pages 594–602, Athens, Greece. ACL Press.

Navigli, R. (2009b). Word sense disambiguation: A survey. *ACM Computing Surveys*, 41(2):1–69.

Navigli, R. (2012). A quick tour of word sense disambiguation, induction and related approaches. In *Proceedings of Theory and Practice of Computer Science, 38th Conference on Current Trends in Theory and Practice of Computer Science*, volume 7147 of *LNCS*, pages 115–129, Spindleruv Mlýn, Czech Republic. Springer.

Navigli, R. and Ponzetto, S. P. (2010). BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL 2010, pages 216–225, Uppsala, Sweden. ACL Press.

Navigli, R. and Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Navigli, R. and Velardi, P. (2003). An analysis of ontology-based query expansion strategies. In *Proceedings of the ECML 2003 Workshop on Adaptive Text Extraction and Mining (ATEM) in the 14th European Conference on Machine Learning*, pages 42–49, Cavtat-Dubrovnik, Croatia.

Navigli, R., Velardi, P., Cucchiarelli, A., and Neri, F. (2004). Extending and enriching WordNet with OntoLearn. In *Proceedings of 2nd Global WordNet Conference (GWC)*, pages 279–284, Brno, Czech Republic. Masaryk University.

Nichols, E., Bond, F., and Flickinger, D. (2005). Robust ontology acquisition from machine-readable dictionaries. In *Proceedings of 19th International Joint Conference on Artificial Intelligence*, IJCAI 2005, pages 1111–1116. Professional Book Center.

Niemann, E. and Gurevych, I. (2011). The people's Web meets linguistic knowledge: Automatic sense alignment of Wikipedia and WordNet. In *Proceedings of International Conference on Computational Semantics*, IWCS 2011, pages 205–214, Oxford, UK.

Nunberg, G. D. (1978). *The pragmatics of reference*. PhD thesis, City University of New York.

O'Hara, T. P. (2005). *Empirical Acquisition of Conceptual Distinctions via Dictionary Definitions*. PhD thesis, NMSU CS.

Oliveira Santos, J., Oliveira Alves, A., Câmara Pereira, F., and Henriques Abreu, P. (2012). Semantic enrichment of places for the Portuguese language. In *Proceedings of INFORUM 2012, Simpósio de Informática*, pages 407—418, Lisbon, Portugal.

Padró, M., Bel, N., and Necsulescu, S. (2011). Towards the automatic merging of lexical resources: Automatic mapping. In *Recent Advances in Natural Language Processing*, RANLP 2011, pages 296–301, Hissar, Bulgaria. RANLP 2011 Organising Committee.

Pantel, P. (2005). Inducing ontological co-occurrence vectors. In *Proceedings of 43rd annual meeting of the Association for Computational Linguistics*, ACL 2005, pages 125–132. ACL Press.

Pantel, P. and Lin, D. (2002). Discovering word senses from text. In *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 613–619.

Pantel, P. and Pennacchiotti, M. (2006). Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of 21st International Conference on Computational Linguistics and 44th annual meeting of the Association for Computational Linguistics*, pages 113–120, Sydney, Australia. ACL Press.

Pantel, P. and Ravichandran, D. (2004). Automatically labeling semantic classes. In *Proceedings of Human Language Technology/North American Chapter of the Association for Computational Linguistics (HLT/NAACL)*, HLT-NAACL 2004, pages 321–328.

ACL Press.

Partee, B. H., ter Meulen, A., and Wall, R. E. (1990). *Mathematical Methods in Linguistics*. Kluwer, Dordrecht.

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

Paulo-Santos, A., Gonçalo Oliveira, H., Ramos, C., and Marques, N. C. (2012). A bootstrapping algorithm for learning the polarity of words. In *Proceedings of Computational Processing of the Portuguese Language - 10th International Conference (PROPOR 2012)*, volume 7243 of *LNCS*, pages 229–234, Coimbra, Portugal. Springer.

Pease, A. and Fellbaum, C. (2010). Formal ontology as interlingua: the SUMO and WordNet linking project and global WordNet linking project. In *Ontology and the Lexicon: A Natural Language Processing Perspective*, Studies in Natural Language Processing, chapter 2, pages 25–35. Cambridge University Press.

Pennacchiotti, M. and Pantel, P. (2006). Ontologizing semantic relations. In *Proceedings of 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, COLING/ACL 2006, pages 793–800. ACL Press.

Peters, W., Peters, I., and Vossen, P. (1998). Automatic Sense Clustering in EuroWordNet. In *Proceedings of 1st International Conference on Language Resources and Evaluation*, LREC'98, pages 409–416, Granada.

Pianta, E., Bentivogli, L., and Girardi, C. (2002). MultiWordNet: developing an aligned multilingual database. In *1st International Conference on Global WordNet*.

Plaza, L., Díaz, A., and Gervás, P. (2010). Automatic summarization of news using WordNet concept graphs. *International Journal on Computer Science and Information System (IADIS)*, V:45–57.

Ponzetto, S. P. and Navigli, R. (2009). Large-scale taxonomy mapping for restructuring and integrating Wikipedia. In *Proceedings of 21st International Joint Conference on Artificial Intelligence*, IJCAI 2009, pages 2083–2088, Pasadena, California. AAAI Press.

Ponzetto, S. P. and Navigli, R. (2010). Knowledge-rich word sense disambiguation rivaling supervised systems. In *Proceedings of 48th Annual Meeting of the Association for Computational Linguistics*, ACL 2012, pages 1522–1531, Uppsala, Sweden. ACL Press.

Prestes, K., Wilkens, R., Zillio, L., and Villavicencio, A. (2011). Extração e validação de ontologias a partir de recursos digitais. In *Proceedings of the 6th International Workshop on Metamodels, Ontologies and Semantic Technologies*, pages 183–188, Gramado, Brazil.

Prévot, L., Huang, C.-R., Calzolari, N., Gangemi, A., Lenci, A., and Oltramari, A. (2010). Ontology and the lexicon: a multi-disciplinary perspective (introduction). In Huang, C.-R., Calzolari, N., Gangemi, A., Lenci, A., Oltramari, A., and Prévot, L., editors, *Ontology and the Lexicon: A Natural Language Processing Perspective*, Studies in Natural Language Processing, chapter 1, pages 3–24. Cambridge University Press.

Pustejovsky, J. (1991). The generative lexicon. *Computational Linguistics*, 17(4):409–441.

Pustejovsky, J. and Boguraev, B., editors (1996). *Lexical semantics: The problem of polysemy*. Oxford, Clarendon Press.

Pustejovsky, J., Castaño, J. M., Zhang, J., Kotecki, M., and Cochran, B. (2002). Robust relational parsing over biomedical literature: Extracting inhibit relations. In *Pacific Symposium on Biocomputing*, pages 362–373.

Pustejovsky, J., Havasi, C., Littman, J., Rumshisky, A., and Verhagen, M. (2006). Towards a generative lexical resource: The Brandeis Semantic Ontology. In *Proceedings of the 5th*

Language Resource and Evaluation Conference, LREC 2006, pages 1702–1705. ELRA.

Quillian, R. (1968). Semantic memory. In *Semantic Information Processing*, pages 216–270. MIT Press, Cambridge, MA, USA.

Rabiner, L. L. (1989). A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of IEEE 77*, volume 77, pages 257–286. The IEEE Computer Society.

Recasens, M., Màrquez, L., Sapena, E., Martí, M. A., Taulé, M., Hoste, V., Poesio, M., and Versley, Y. (2010). Semeval-2010 task 1: Coreference resolution in multiple languages. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval 2010, pages 1–8, Stroudsburg, PA, USA. ACL Press.

Resnik, P. (1995). Disambiguating Noun Groupings with Respect to WordNet Senses. In *Proceedings of 3rd Workshop on Very Large Corpora*, pages 54–68. Cambridge, MA, USA.

Richardson, S., Vanderwende, L., and Dolan, W. (1993). Combining dictionary-based and example-based methods for natural language analysis. In *Proceedings of the 5th International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 69–79, Kyoto, Japan.

Richardson, S. D. (1997). *Determining Similarity and Inferring Relations in a Lexical Knowledge Base*. PhD thesis, The City University of New York, New York, NY, USA.

Richardson, S. D., Dolan, W. B., and Vanderwende, L. (1998). MindNet: Acquiring and structuring semantic information from text. In *Proceedings of 17th International Conference on Computational Linguistics*, COLING'98, pages 1098–1102.

Riloff, E. and Shepherd, J. (1997). A corpus-based approach for building semantic lexicons. In *Proceedings of 2nd Conference on Empirical Methods in Natural Language Processing*, EMNLP'97, pages 117–124.

Roark, B. and Charniak, E. (1998). Noun-phrase co-occurrence statistics for semiautomatic semantic lexicon construction. In *Proceedings 17th International Conference on Computational Linguistics*, COLING'98, pages 1110–1116, Morristown, NJ, USA. ACL Press.

Rocha, P. A. and Santos, D. (2000). CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. In *V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR 2000)*, pages 131–140, São Paulo. ICMC/USP.

Rodrigues, M., Dias, G. P., and Teixeira, A. (2011). Criação e acesso a informação semântica aplicada ao Governo Eletrónico. *Linguamática*, 3(2):55–68.

Rodrigues, R., Gonçalo Oliveira, H., and Gomes, P. (2012). Uma abordagem ao Págico baseada no processamento e análise de sintagmas dos tópicos. *Linguamática*, 4(1):31–39.

Roget, P. M. (1852). *Roget's Thesaurus of English Words and Phrases*. Longman, London.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Reviews*, 65(6):386–408.

Rosso, P., Ferretti, E., Jiménez, D., and Vidal, V. (2004). Text categorization and information retrieval using WordNet senses. In *Proceedings of 2nd Global Wordnet Conference*, GWC 2004, pages 299–304.

Ruiz-Casado, M., Alfonseca, E., and Castells, P. (2005). Automatic assignment of Wikipedia encyclopedic entries to WordNet synsets. In *Proceedings Advances in Web Intelligence 3rd International Atlantic Web Intelligence Conference, AWIC 2005*, volume 2663 of *LNAI*, pages 380–386. Springer.

Ruiz-Casado, M., Alfonseca, E., and Castells, P. (2007). Automatising the learning of lexical patterns: An application to the enrichment of wordnet by extracting semantic relationships from wikipedia. *Data & Knowledge Engineering*, 61(3):484–499.

Russell, S. and Norvig, P. (1995). *Artificial Intelligence: A Modern Approach*. Prentice-Hall.

Saias, J. M. G. (2010). *Contextualização e Activação Semântica na Selecção de Resultados em Sistemas de Pergunta-Resposta*. PhD thesis, Universidade de Évora, Évora, Portugal.

Sajous, F., Navarro, E., Gaume, B., Prévot, L., and Chudy, Y. (2010). Semi-automatic Endogenous Enrichment of Collaboratively Constructed Lexical Resources: Piggybacking onto Wiktionary. In *Advances in Natural Language Processing, 7th International Conference on NLP, IceTAL 2010*, volume 6233 of *LNCS*, pages 332–344, Reykjavik, Iceland. Springer.

Salomao, M. M. M. (2009). Framenet Brasil: Um trabalho em progresso. *Calidoscópio*, 7(3):171–182.

Sampson, G. (2000). Review of Fellbaum (1998). *International Journal of Lexicography*, 13(1):54–59.

Santos, D. (1992). Natural Language and Knowledge Representation. In *Proceedings of the ERCIM Workshop on Theoretical and Experimental Aspects of Knowledge Representation*, pages 195–197.

Santos, D. (1997). The importance of vagueness in translation: Examples from English to Portuguese. *Romansk Forum*, 5:43–69. Revised as Santos 1998.

Santos, D. (2011). Linguateca's infrastructure for Portuguese and how it allows the detailed study of language varieties. *OSLa: Oslo Studies in Language*, 3(2):113–128. Volume edited by J.B.Johannessen, Language variation infrastructure.

Santos, D., Barreiro, A., Freitas, C., Gonçalo Oliveira, H., Medeiros, J. C., Costa, L., Gomes, P., and Silva, R. (2010). Relações semânticas em português: comparando o TeP, o MWN.PT, o Port4NooJ e o PAPEL. In *Textos seleccionados. XXV Encontro Nacional da Associação Portuguesa de Linguística*, pages 681–700. APL, Lisboa, Portugal.

Santos, D. and Bick, E. (2000). Providing Internet access to Portuguese corpora: the AC/DC project. In *Proceedings of 2nd International Conference on Language Resources and Evaluation*, LREC 2000, pages 205–210.

Santos, D., Mota, C., Freitas, C., and Costa, L. (2012). Linguamática – edição especial Págico (eds.).

Santos, D. and Rocha, P. (2001). Evaluating CETEMPúblico, a free resource for Portuguese. In *Proceedings of 39th Annual Meeting of the Association for Computational Linguistics*, ACL 2001, pages 442–449. ACL Press.

Sarmento, L. (2010). *Definition and Computation of Similarity Operations between Web-specific Lexical Items*. PhD thesis, Universidade do Porto.

Sarmento, L., Teixeira, J., and Oliveira, E. (2008). Experiments with query expansion in the Raposa (Fox) question answering system. In *Working Notes for the Cross Evaluation Forum*, CLEF 2008, Aarhus, Denmark.

Schaeffer, S. E. (2007). Graph clustering. *Computer Science Review*, 1(1):27–64.

Schuler, K. K. (2006). *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. PhD thesis, University of Pennsylvania.

Seco, N., Veale, T., and Hayes, J. (2004). An intrinsic information content metric for semantic similarity in WordNet. In *Proceedings of 16th European Conference on Artificial Intelligence*, ECAI 2004, pages 1089–1090, Valencia, Spain. IOS Press.

Shen, D., Zhang, J., Su, J., Zhou, G., and Tan, C. L. (2004). Multi-criteria-based active learning for named entity recognition. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, ACL 2004, pages 589–596, Barcelona, Spain. ACL Press.

Shi, L. and Mihalcea, R. (2005). Putting pieces together: Combining FrameNet, VerbNet and WordNet for robust semantic parsing. In *Computational Linguistics and Intelligent Text Processing (CICLing 2005)*, volume 3406 of *LNCS*, pages 100–111. Springer.

Sibson, R. (1973). SLINK: an optimally efficient algorithm for the single-link cluster method. *The Computer Journal*, 16(1):30–34.

Silva, A., Marques, C., Baptista, J., Ferreira, A., and Mamede, N. (2012a). REAP.PT serious games for learning portuguese. In *Proceedings of the 10th International Conference on Computational Processing of the Portuguese Language (PROPOR 2012)*, volume 7243 of *LNCS*, pages 248–259, Coimbra, Portugal. Springer.

Silva, M. J., Carvalho, P., and Sarmento, L. (2012b). Building a sentiment lexicon for social judgement mining. In *Proceedings of Computational Processing of the Portuguese Language - 10th International Conference (PROPOR 2012)*, volume 7243 of *LNCS*, pages 218–228, Coimbra, Portugal. Springer.

Simões, A., ao Almeida, J. J., and Farinha, R. (2010). Processing and extracting data from Dicionário Aberto. In *Proceedings of International Conference on Language Resources and Evaluation*, LREC 2010, Malta. ELRA.

Simões, A. and Farinha, R. (2011). Dicionário Aberto: Um novo recurso para PLN. *Vice-Versa*, pages 159–171.

Simões, A., Sanromán, A. I., and ao Almeida, J. J. (2012). Dicionário-Aberto: A source of resources for the Portuguese language processing. In *Proceedings of Computational Processing of the Portuguese Language, 10th International Conference (PROPOR 2012), Coimbra Portugal*, volume 7243 of *LNCS*, pages 121–127. Springer.

Smith, B. (2001). Ontology and information systems. `http://ontology.buffalo.edu/ontology(PIC).pdf` (draft paper: retrieved March 2012).

Smullyan, R. (1995). *First-Order Logic*. Dover Publications.

Snow, R., Jurafsky, D., and Ng, A. Y. (2005). Learning syntactic patterns for automatic hypernym discovery. In *Advances in Neural Information Processing Systems*, pages 1297–1304. MIT Press, Cambridge, MA.

Soderland, S. and Mandhani, B. (2007). Moving from textual relations to ontologized relations. In *Proceedings of AAAI Spring Symposium on Machine Reading*.

Sowa, J. (1999). *Knowledge Representation: Logical, Philosophical and Computational Foundations*. Thomson Learning, New York, NY, USA.

Sowa, J. F. (1992). Conceptual Graphs summary. In Eklund, P., Nagle, T., Nagle, J., and Gerholz, L., editors, *Conceptual structures: current research and practice*, pages 3–51. Ellis Horwood.

Stamou, S., Oflazer, K., Pala, K., Christoudoulakis, D., Cristea, D., Tufis, D., Koeva, S., Totkov, G., Dutoit, D., and Grigoriadou, M. (2002). BalkaNet: A multilingual semantic network for the balkan languages. In *Proc. 1st Global WordNet Conference*, GWC'02.

Suchanek, F. M., Kasneci, G., and Weikum, G. (2007). Yago: a core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web*, WWW 2007, pages 697–706, Alberta, Canada. ACM Press.

Teixeira, J., Sarmento, L., and Oliveira, E. (2010). Comparing verb synonym resources for Portuguese. In *Proceedings of Computational Processing of the Portuguese Language, 9th International Conference, PROPOR 2010*, volume 6001 of *LNCS*, pages 100–109. Springer.

Tokunaga, T., Syotu, Y., Tanaka, H., and Shirai, K. (2001). Integration of heterogeneous language resources: A monolingual dictionary and a thesaurus. In *Proceedings of the 6th Natural Language Processing Pacific Rim Symposium*, NLPRS 2001, pages 135–142, Tokyo, Japan.

Tonelli, S. and Pighin, D. (2009). New features for FrameNet: WordNet mapping. In

Proceedings of 13th Conference on Computational Natural Language Learning, CoNLL 2009, pages 219–227, Stroudsburg, PA, USA. ACL Press.

Toral, A., Muñoz, R., and Monachini, M. (2008). Named entity WordNet. In *Proceedings International Conference on Language Resources and Evaluation*, LREC 2008, Marrakech, Morocco. ELRA.

Turney, P. D. (2001). Mining the Web for synonyms: PMI–IR versus LSA on TOEFL. In *Proceedings of 12th European Conference on Machine Learning, ECML 2001*, volume 2167 of *LNCS*, pages 491–502. Springer.

Turney, P. D., Littman, M. L., Bigham, J., and Shnayder, V. (2003). Combining independent modules to solve multiple-choice synonym and analogy problems. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2003)*, CILT, pages 482–489, Borovets, Bulgaria. John Benjamins, Amsterdam/Philadelphia.

van Assem, M., Gangemi, A., and Schreiber, G. (2006). RDF/OWL representation of WordNet. W3c working draft, World Wide Web Consortium.

van Dongen, S. M. (2000). *Graph Clustering by Flow Simulation*. PhD thesis, University of Utrecht.

Vanderwende, L. (1994). Algorithm for automatic interpretation of noun sequences. In *Proceedings of the 15th Conference on Computational linguistics*, COLING'94, pages 782–788, Morristown, NJ, USA. ACL Press.

Vanderwende, L. (1995). Ambiguity in the acquisition of lexical information. In *Proceedings of the AAAI 1995 Spring Symposium*, Working notes of the symposium on representation and acquisition of lexical knowledge, pages 174–179.

Vanderwende, L., Kacmarcik, G., Suzuki, H., and Menezes, A. (2005). MindNet: An automatically-created lexical resource. In *Proceedings of HLT/EMNLP 2005 Interactive Demonstrations*, pages 8–9, Vancouver, Canada. ACL Press.

Veale, T. (2006). Tracking the lexical zeitgeist with wordnet and wikipedia. In *Proceedings of 17th European Conference on Artificial Intelligence (ECAI 2006)*, pages 56–60, Riva del Garda, Italy. IOS Press.

Veiga, A., Lopes, C., Celorico, D., Proença, J., Perdigão, F., and Candeias, S. (2012). O desafio da participação humana do IT-Coimbra no Págico. *Linguamática*, 4(1):49–51.

Velldal, E. (2005). A fuzzy clustering approach to word sense discrimination. In *Proceedings of 7th International Conference on Terminology and Knowledge Engineering*, Copenhagen, Denmark.

Verhagen, M., Saurí, R., Caselli, T., and Pustejovsky, J. (2010). Semeval-2010 task 13: Tempeval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 57–62, Stroudsburg, PA, USA. Association for Computational Linguistics.

Voorhees, E. M. (1998). Using WordNet for Text Retrieval. In *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*, pages 285–303. The MIT Press.

Vossen, P. (1997). EuroWordNet: a multilingual database for information retrieval. In *Proceedings of DELOS workshop on Cross-Language Information Retrieval*, Zurich.

Vossen, P., Maks, I., Segers, R., and VanderVliet, H. (2008). Integrating Lexical Units, Synsets and Ontology in the Cornetto Database. In *Proceedings of the International Conference on Language Resources and Evaluation*, LREC 2008, Marrakech, Morocco. ELRA.

Wandmacher, T., Ovchinnikova, E., Krumnack, U., and Dittmann, H. (2007). Extraction, evaluation and integration of lexical-semantic relations for the automated construction of a lexical ontology. In *Proceedings of 3rd Australasian Ontology Workshop (AOW)*,

volume 85 of *CRPIT*, pages 61–69, Gold Coast, Australia. ACS.

Weale, T., Brew, C., and Fosler-Lussier, E. (2009). Using the Wiktionary graph structure for synonym detection. In *Proceedings of 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, People's Web '09, pages 28–31, Stroudsburg, PA, USA. ACL Press.

Wiegand, M., Roth, B., and Klakow, D. (2012). Web-based relation extraction for the food domain. In *Natural Language Processing and Information Systems, Proceedings of 17h International Conference on Applications of Natural Language to Information Systems (NLDB)*, volume 7337 of *LNCS*, pages 222–227, Groningen, The Netherlands. Springer.

Wilks, Y. (2000). Is word sense disambiguation just one more NLP task? *Computers and the Humanities*, 34:235–243.

Wilks, Y., Fass, D., ming Guo, C., Mcdonald, J. E., Plate, T., and Slator, B. M. (1988). Machine tractable dictionaries as tools and resources for natural language processing. In *Proceedings of the 12th conference on Computational linguistics*, COLING'88, pages 750–755, Morristown, NJ, USA. ACL Press.

Williams, G. K. and Anand, S. S. (2009). Predicting the polarity strength of adjectives using WordNet. In *Proceedings of the 3rd International Conference on Weblogs and Social Media*, ICWSM 2009, San Jose, California, USA. AAAI Press.

Winston, M. E., Chaffina, R., and Herrmann, D. (1987). A taxonomy of part-whole relations. *Cognitive Science*, 11(4):417–444.

Woodsend, K. and Lapata, M. (2011). Wikisimple: Automatic simplification of wikipedia articles. In *Proceedings of 25th AAAI Conference on Artificial Intelligence*, AAAI 2011, San Francisco, California, USA. AAAI Press.

Wu, F. and Weld, D. S. (2010). Open information extraction using Wikipedia. In *Proceedings of 48th Annual Meeting of the Association for Computational Linguistics*, pages 118–127, Uppsala, Sweden. ACL Press.

Zesch, T., Müller, C., and Gurevych, I. (2008a). Extracting lexical semantic knowledge from Wikipedia and Wiktionary. In *Proceedings of 6th International Conference on Language Resources and Evaluation*, LREC 2008, Marrakech, Morocco.

Zesch, T., Müller, C., and Gurevych, I. (2008b). Using Wiktionary for computing semantic relatedness. In *Proceedings of 23rd National Conference on Artificial Intelligence*, AAAI 2008, pages 861–866. AAAI Press.

Appendix A

Description of the extracted semantic relations

This appendix lists the semantic relations extracted in the scope of Onto.PT. After the relation names, we present the relation subtypes, a natural language attempt to describe it, and an example of each subtype. Besides Onto.PT, these relation types are the same as in PAPEL and CARTÃO.

Hypernymy

- x hiperonimoDe y
  - x is a kind/instance of y
  - y is a category that x belongs to
    - e.g. x = *animal*, y = *dog*

Part-of

- x parteDe y
  - x is a part of y
  - y has the part x
    - e.g. x = *wheel*, y = *car*
- x parteDeAlgoComPropriedade y
  - x is part of something that is y
  - y is a property typically held by something with the part x
    - e.g. x = *oil*, y = *oily*

Member-of

- x membroDe y
  - x is a member of y
  - y includes x
    - e.g. x = *deputy*, y = *parliament*
- x membroDeAlgoComPropriedade y

  - x is a member of something that is y
  - y is a property typically held by something with that includes x
    - e.g. x = *entity*, y = *collective*

- x propriedadeDeAlgoMembroDe y
  - x is a property typically held by something that is member of y
  - y includes something that is x
    - e.g. x = *partner*, y = *partnership*

Contains

- x contidoEm y
  - x is contained in y
  - y contains x
    - e.g. x = *heart*, y = *chest*
- x contidoEmAlgoComPropriedade y
  - x is contained in something that is y
  - y is a property typically held by something that contains x
    - e.g. x = *conclusion*, y = *conclusive*

Material

- x materialDe y
  - x is a material of y
  - y is made of x
    - e.g. x = *fabric*, y = *dress*

Location

- x localOrigemDe y
  - x is the place where y is originally from or can be found in
  - y is originally from/can be found in x
    - e.g. x = *Portugal*, y = *portuguese*

Causation

- x causadorDe y
  - x is the causation of y/causes y to happen
  - y can be the result of x/because of y
    - e.g. x = *flu*, y = *fever*
- x causadorDeAlgoComPropriedade y
  - x is the causation of something that is y
  - y is a property typically held by something that is the result of x
    - e.g. x = *paixão*, *passional* = *conclusive*

- x propriedadeDeAlgoQueCausa y
  - x is a property typically held by something that causes y to happen
  - y can be the result of something that is x
    - e.g. x = *despicable*, y = *disdain*
- x accaoQueCausa y
  - x is an action that causes y to happen
  - y can be the result of x
    - e.g. x = *move*, y = *movement*
- x causadorDaAccao y
  - x is the causation of the action y
  - y is an action that may be the result of x
    - e.g. x = *hot*, y = *melt*

Producer

- x produtorDe y
  - x is the producer/may produce y
  - y is produced by x
    - e.g. x = *grapevine*, y = *grape*
- x produtorDeAlgoComPropriedade y
  - x is the producer of something that is y
  - y is a property typically held by something that is produced by x
    - e.g. x = *oral*, y = *mouth*
- x propriedadeDeAlgoProdutorDe y
  - x is a property typically held by something that produces y
  - y is produced by something that is x
    - e.g. x = *nauseous*, y = *nausea*

Purpose

- x fazSeCom y
  - x is performed/obtained with y
  - y is a means for performing/obtaining x
    - e.g. x = *transport*, y = *lorry*
- x fazSeComAlgoComPropriedade y
  - x is performed/obtained with something that is y
  - y is a property typically held by a means for performing/obtaining x
    - e.g. x = *inspiration*, y = *inspiratory*
- x finalidadeDe y

  - x is the purpose of y
  - y is a means/instrument for x
    - e.g. x = *drink*, y = *glass*

- x finalidadeDeAlgoComPropriedade y
  - x is a purpose of something that is y
  - y is a property typically held by a means/instrument for x
    - e.g. x = *compensar*, y = *compensatory*

Quality

- x temQualidade y
  - x has the quality y
  - y is a quality of x
    - e.g. x = *soft*, y = *softness*
- x devidoAQualidade y
  - x is it because it has the quality y
  - y is a quality of what is x
    - e.g. x = *digestible*, y = *digestibility*

State

- x temEstado y
  - x has the state y
  - y is a possible state of x
    - e.g. x = *serene*, y = *serenity*
- x devidoAEstado y
  - x is it because it has the state y
  - y is a possible state of what is x
    - e.g. x = *positive*, y = *positivity*

Antonymy

- x antonimoNDe y
  - x is the opposite of y
    - e.g. x = *front*, y = *back*
- x antonimoVDe y
  - x is the opposite of y
    - e.g. x = *advance*, y = *retreat*
- x antonimoAdjDe y
  - x is the opposite of y
    - e.g. x = *long*, y = *short*

- x antonimoAdjDe y
  - x is the opposite of y
    - e.g. x = *quickly*, y = *slowly*

Property

- x dizSeSobre y
  - x is a property characteristic of/related with y
  - y is related to/is referred by x
    - e.g. x = *sympathetic*, y = *sympathy*
- x dizSeDoQue y
  - x is a property characteristic of/related with the action y
  - y is an action related to/that may characterize something that is x
    - e.g. x = *smiling*, y = *smile*

Manner

- x maneiraPorMeioDe y
  - x is a manner related with y
  - y is related to the manner x
    - e.g. x = *repeatedly*, y = *repetition*
- x maneiraComPropriedade y
  - x is a manner related with what is y
  - y is a property related with the manner x
    - e.g. x = *shamefully*, y = *shameful*

Manner without

- x maneiraSem y
  - x is a manner characterised by not being related with y
  - y is not related to the manner x
    - e.g. x = *doubtless*, y = *doubt*
- x maneiraSemAccao y
  - x is a manner characterised by not being related with the action y
  - y is an not related with the manner x
    - e.g. x = *silently*, y = *make_noise*

Appendix B

Coverage of EuroWordNet base concepts

In this appendix, we present the closest matches between the Onto.PT synsets and EuroWordNet 164 base concrete concepts, as indicated in:

- `http://www.globalwordnet.org/gwa/ewn_to_bc/corebcs.html`

These matches were selected manually, after observation. Table B.1 shows the mappings defined for the 66 concrete concepts (all nouns), and table B.2 for the 98 abstract concepts (nouns and verbs).

Table B.1: Mapping between Onto.PT and WordNet concrete base concepts.

WordNet synset	**Onto.PT synset**
amount.1, measure.1, quantity.1, quantum.1 • how much there is of anything	*porção, tanto, soma, quantidade, cifra, quantia*
animal.1, animate being.1, beast.1, brute.1, creature.1, fauna.1 • a living organism characterized by voluntary movement	*animal, bicho, balada, animália, béstia, alimal, minante*
apparel.1, clothes.1, clothing.1, vesture.1, wear.2, wearing apparel.1 • covering designed to be worn on a person's body	*vestido, traje, fato, costume, vestuário, cobertura, pálio, revestimento, veste, vestimenta, vestidura, roupa, toilette, trajo, indumentária, encadernação, fatiota, farpela, véstia, trajadura, fardamenta, vestiaria, vestimento, induto, indúvia, vestes, indumento, entraje, vestia, fateco, trajar, trem, rebanho, manada, gado*
artefact.1, artifact.1 • a man-made object	*fabrico, artefacto, manufactura, artefato, manufatura*
article of furniture.1, furniture.1, piece of furniture.1 • artifacts that make a room ready for occupancy	*móvel, objecto, peça*
asset.2 • anything of material value or usefulness	*capital, bem, propriedade, património, riqueza*
being.1, life form.1, living thing.1, organism.1 • any living entity	*vida, existência, ser, espírito, alma, essência*
beverage.1, drink.2, potable.1 • any liquid suitable for drinking	*líquido, bebida, beberagem, bebedura, beber, bromo, poção, poto, pingalho*
body.3, organic structure.1, physical structure.1 • the entire physical structure of an organism (especially an animal or human being)	*corpo, carne*
bound.2, boundary.2, bounds.2 • the line or plane indicating the limit or extent of something	*margem, marca, limite, gol, marco, baliza, fronteira, meta, raia, órbita, divisa, confins, contorno, linda, estrema, circumdamento, linde, circundamento, têrmo, xebre, alfa, medida, últimas, slogan, moeda*
building.3, edifice.1 • a structure that has a roof and walls	*imóvel, casa, edifício, prédio, bloco, herdade, aranha-céu*
causal agency.1, causal agent.1, cause.1 • any entity that causes events to happen	*agente, causa, factor, fator*
chemical compound.1, compound.4 • a substance formed by chemical union of two or more elements or ingredients in definite proportion by weight	*químico*
chemical element.1, element.6 • any of the more than 100" known substances (of which 93 occur naturally) that cannot be separated into simpler substances and that singly or in combination constitute all matter	N/A
cloth.1, fabric.1, material.1, textile.1 • something made by weaving or felting or knitting or crocheting natural or synthetic fibers	*tecido, fazenda, pano, bona, fali, panho, xitaca, chela, haver, algo*

Continued on next page...

Table B.1 – continued from previous page

WordNet synset	Onto.PT synset
commodity.1, goods.1 • articles of commerce	*artigo, mercadoria, mercancia, veniaga, merce, marçaria, mercer*
construction.4, structure.1 • a thing constructed; a complex construction or entity	*obra, estrutura, construção, edificação, edifício, fábrica, prédio, erecção, edificamento*
consumer goods.1 • goods (as food or clothing) intended for direct use or consumption	N/A
covering.4 • an artifact that protects or shelters or conceals	*coberta, protecção, cobertura, fronha, toldo, velame, tolda, ensombro*
creation.3 • something that has been brought into existence by someone	*obra, produção, criação, descoberta, invenção, invento*
decoration.2, ornament.1 • something used to beautify	*adornamento, enfeite, decoração, ataviamento, ornamentação, aformoseamento, adereçamento, enfeitamento, decoramento, aformosentamento, ornatos*
device.2 • an instrumentality invented for a particular purpose	*material, máquina, instrumento, aparelho, maquinismo, utensílio, apetrecho, equipamento, amanho, apresto, aparelhamento, apeiro, alfaias, arreios, estromento, utênsil, aprestos, apetrechos, petrechos, telefone, documento, avião, acta, munições*
document.2, papers.1, written document.1 • writing providing information; esp. of an official nature	*patente, escrito, carta, título, documento, diploma, canudo, pretexo, quirógrafo*
dry land.1, earth.3, ground.7, land.6, solid ground.1, terra firma.1 • the solid part of the Earth's surface	*terra, solo, chão, país, território, campina, poeira*
entity.1 • something having concrete existence; living or nonliving	*existência, existencia, ente, ser, realidade, sêr*
extremity.3 • the outermost or farthest region or point	*extremo, extremidade, ponta*
flora.1, plant.1, plant life.1 • a living organism lacking the power of locomotion	*vegetal, planta, pranta, plantas, mapa, melancieira, caruru, cameleira*
fluid.2 • a substance that is fluid at room temperature and pressure	*fluido*
food.1, nutrient.1 • any substance that can be metabolized by an organism to give energy and build tissue	*alimentação, sustento, mantença, mantimento, alimento, comida, nutrição, paparoca, nutrimento, buxa, pábulo, sustentação*
furnishings.2 • the furniture and appliances and other movable accessories (including curtains and rugs) that make a home (or other building) livable	*mobiliário, mobília, arreamento, adereços*
garment.1 • an article of clothing	*hábito, vestuário, veste, vestimenta, vestidura*
group.1, grouping.1 • any number of entities (members) considered as a unit	*aglomerado, grupo, colônia, coleção, agrupamento, morganho*
human.1, individual.1, mortal.1, person.1, someone.1, soul.1 • a human being	*mortal, homem, humanidade, ser_humano*
inanimate object.1, object.1, physical object.1 • a nonliving entity	*objecto, coisa, ente*
instrument.2 • a device that requires skill for proper use	*ferramenta, instrumento, documento, utensílio, ferramental, apetrecho*
instrumentality.1, instrumentation.2 • an artifact (or system of artifacts) that is instrumental in accomplishing some end	*orquestração, instrumentação*
language unit.1, linguistic unit.1 • one of the natural units into which linguistic messages can be analyzed	*termo, têrmo, expressão, vocábulo*
line.21 • a spatial location defined by a real or imaginary unidimensional extent	*directriz, linha, coordenada, diretriz, diretiva, plica, lĩa, online*
line.26 • a length (straight or curved) without breadth or thickness; the trace of a moving point	*extremo, traçado, rofo, alinhamento, risco, linha, traçamento, raia, traço, rasgo, risca, assomo, estria, arraia, riscadura, riscamento, feições, espelde, serradura, sulco*
liquid.4 • a substance that is liquid at room temperature and pressure	N/A
location.1 • a point or extent in space	*local, lado, lugar, sítio, situação, localidade, básis*
material.5, stuff.7 • the tangible substance that goes into the makeup of a physical object; "coal is a hard black material"; "wheat is the stuff they use to make bread"	*matéria, corporalidade, materialidade, bruteza, fisicalidade, prosaísmo, corporeidade, animalização*
matter.1, substance.1 • that which has mass and occupies space; "an atom is the smallest indivisible unit of matter"	*substância, matéria*
medium of exchange.1, monetary system.1 • anything that is generally accepted as a standard of value and a measure of wealth in a particular country or region	*dinheiro, moeda, níquel, toura, numo, dinheiros-secos*
mixture.5 • a substance consisting of two or more substances mixed together (not in fixed proportions and not with chemical bonding)	*misto, mistura, mescla, amálgama, promiscuidade, impurezas, anguzada*
money.2 • a medium of exchange that functions as legal tender	*pastel, ouro, massa, dinheiro, finanças, verba, quantia, maquia, tutu, guita, pasta, metal, milho, tostão, cobre, arame, grana, cacau, pingo, bagaço, china, bago, pataco, teca, pecúnia, chelpa, boro, pilim, massaroca, roço, jimbo, bagalhoça, baguines, parrolo, marcaureles, bagalho, bilhestres, janda-cruz, cum-quibus, mussuruco, zerzulho, calique, dieiro, pila, matambira, gimbo, cunques, fanfa, maco, jibungo, patacaria, carcanhol, espécie, caroço, pecunia, pecuniária, estilha, gaita, guines, painço*
natural object.1 • an object occurring naturally; not made by man	N/A

Continued on next page...

Table B.1 – continued from previous page

WordNet synset	Onto.PT synset
opening.4 • a vacant or unobstructed space	*espaço, intervalo, aberta, trecho, interrupção*
part.3, portion.2 • something less than the whole of a human artifact	*parte, peça, pedaço, fragmento, porção*
part.9, region.2 • the extended spatial location of something	*parte, noite, zona*
part.12, portion.5 • something determined in relation to something that includes it; "he kept all the parts because he had no idea which part was effective"	*parte, divisão, fragmento, porção, quinhão, troço, fracção, secção, parcela, segmento, seção, retalho, fração, pagela, metâmero, quebrado, artelho*
passage.6 • a path or channel through or along which someone or something may pass	*comunicação, acesso, passagem, porta, passadouro, passadoiro*
piece of work.1, work.4 • something produced or accomplished through the effort or activity or agency of a person or thing	*obra, trabalho, serviço, feitura, opifício*
place.13, spot.10, topographic point.1 • a point located with respect to surface features of some region	*ponto, lugar, sítio, altura*
point.12 • the precise location of something	*situação, posição, localização, mansão*
possession.1 • anything owned or possessed	*possessão, propriedade, mão, posse, domínio, senhorio, feudo, pertence, especialidade, senhoria*
product.2, production.2 • an artifact that has been produced by someone or some process	*obra, produção, produto, fabrico, output*
representation.3 • a visual or tangible rendering of someone or something	*imagem, descrição, diegese, representação, figuração, retrato, retratação, reprodução, raconto, cópia*
surface.1 • the outer boundary of an object or a material layer constituting or resembling such a boundary	*exterior, face, superfície*
surface.4 • the extended two-dimensional outer boundary of a three-dimensional object	N/A
symbol.2 • an arbitrary sign (written or printed) that has acquired a conventional significance	*imagem, signo, sinal, símbolo, sino, atributo, alegoria*
way.4 • any road or path affording passage from one place to another; "he said he was looking for the way out"	*itinerário, via, caminho, estrada, veia, trajetória, tendência, alfazar, norma, viela, tramite*
word.1 • a unit of language that native speakers can identify	*termo, têrmo, signo, palavra, fala, vocábulo, dicção, vocabro, mo, verbo, sílaba, palabra, lexema, parávoa, parávora*
worker.2 • a person who has employment	*trabalhador, videiro, obreiro, operário, ganhador, jornaleiro, ganhadinheiro, obregão*
writing.4, written material.1 • reading matter; anything expressed in letters of the alphabet (especially when considered from the point of view of style and effect)	*composição, redação, redacção, misto*
written communication.1, written language.1 • communication by means of written symbols	*comunicado, memorando, comunicação, informação, notícia, nota, mensagem, anúncio, aviso, informe, participação*

Table B.2: Mapping between Onto.PT and WordNet abstract base concepts.

WordNet synset	Onto.PT synset
(n) ability.2, power.3 • possession of the qualities (especially mental qualities) required to do something or get something done	*possibilidade, poder, capacidade, faculdade, habilidade*
(n) abstraction.1 • a concept formed by extracting common features from examples	*abstracção, conceptualização, abstraimento*
(n) act.1, human action.1, human activity.1 • something that people do or cause to happen	*feito, obra, ação, ato, acto, realização, acção, auto, atuação, aução, acções, inoperação*
(v) act.12, do something.1, move.19, perform an action.1, take a step.2, take action.1, take measures.1 • carry out an action; be an agent; carry into effect	*proceder, funcionar, agir, obrar, operar, trabalhar, actuar, atuar, andar, manobrar, trabucar*
(v) act together.2, act towards others.1, interact.1 • act together with others	*interagir, interatuar, interactuar*
(n) action.1 • something done (usually as opposed to something said)	*acto, fenómeno, passo, facto, negócio, coisa, cousa, espécie, realidade, mistério*
(n) activity.1 • any specific activity or pursuit	*atividade, actividade, dinamismo, diligência, expedição, agilidade, prontidão*
(n) aim.4, bearing.5, heading.2 • the direction or path along which something moves or along which it lies	*percurso, caminho, viagem, trajecto, distância, curso, rumo, discurso, varadouro, trajeto, trajetória, trajectória, correnteza, decorrer, ...*
(v) allow.6, let 7, permit.5 • make it possible for something to happen	*deixar, conceder, facilitar, facultar, franquear, permitir, ceder, doar, ofertar, consentir, autorizar, outorgar, possibilitar, viabilizar, ...*
(v) alter.2, change.12, vary.1 • make or become different in some particular way, without permanently losing one's or its former characteristics or essence	*mudar, variar, modificar, alterar, transmutar, transmudar*
(n) amount of time.1, period.3, period of time.1 • a length of time	*tempo, período, prazo, timing, dilação*
(n) attitude.3, mental attitude.1 • a complex mental orientation involving beliefs and feelings and values and dispositions to act in certain ways	*comportamento, atitude, conduta, procedimento, reacção, posicionamento*

Continued on next page...

Table B.2 – continued from previous page

WordNet synset	Onto.PT synset
(n) attribute.1 • an abstraction belonging to or characteristic of an entity	*atributo, característica, apanágio, prerogativa, partilha, apanagem*
(n) attribute.2, dimension.3, property.3 • a construct whereby objects or individuals can be distinguished	*propriedade, caraterística, atributo, qualidade, faculdade, predicado, valor, capacidade, virtude, calibre, carácter, condão, jaez, mérito, génio, característica, pano, prenda, validade, naipe, apanágio, calidade, gênio, ...*
(v) be 4, have the quality of being.1 • copula, used with an adjective or a predicate noun	*ser, consistir*
(v) be 9, occupy a certain area.1, occupy a certain position.1 • be somewhere	*ser, sêr, haver, estar, existir*
(v) cause 6, get 9, have 7, induce.2, make.12, stimulate.3 • cause to do; cause to act in a specified manner	*gerar, promover, ocasionar, produzir, importar, determinar, incubar, causar, provocar, puxar, trazer, dar, desenvolver, suscitar, custar, derivar, originar, criar, sugerir, procriar, retirar, implicar, render, lucrar, proporcionar, desencadear, surtir, uberar, seer, chamar, lançar, propor, crescer, interessar, transportar, animar, facilitar, atrair, inspirar, arrastar, comandar, motivar, inventar, evocar, captar, pregar, instituir, retornar, chover, acarretar, fomentar, induzir, desferir, infligir, abotoar, avivar, incutir, alevantar, germinar, predispor, infundir, ensejar, carrear, catalisar, engenhar, acarear, carretar, carrar, acarrear, chimpar, carrejar, carretear, agomar, acarrejar, antemover, determinhar, levantar, bracejar, mover, pupular, ser, azar*
(v) cause.7, do.5, give rise to.1, make.17 • make a big stink	*gerar, produzir, causar, originar, fabricar*
(v) cease.3, discontinue.2, give up.12, lay off.2, quit.5, stop.20 • put an end to a state or an activity	*interromper, cessar, deixar, suspender, parar, paralisar, descontinuar*
cerebrate.1, cogitate.1, think.4 • use or exercise the mind in order to make a decision or arrive at a solution	*formular, pensar, meditar, pretender, intentar, projectar, tencionar, projetar, planear, cogitar, intencionar, maquinar, matutar, desejar, apetecer, dever, querer*
(n) change.1 • the act of changing something	*mudança, alteração, modificação, mutação, substituição, vicissitude, viragem, reviravolta, reviramento, rectificação, metamorfismo, imutação*
(v) change.11 • undergo a change; become different in essence; losing one's or its original nature	*transformar, converter, mudar, cambiar, modificar, transfazer, alterar, desengatilhar, voltar, imutar, elaborar, mexer*
(v) change magnitude.1, change size.1 • change in size or magnitude	N/A
(v) change of location.1, motion.1, move 4, movement.1 • the act of changing your location from one place to another	*deslocação, deslocamento, movimentação, circulação, locomoção, mobilização, desarticulação, luxação*
(v) change of position.1, motion.2, move.5, movement.2 • motion that does not entail a change of location	*movimento, animação, movimentação, agitação, moto, vibração, alvoroço, efervescência, vida, trepidação*
(v) change of state.1 • the act of changing something into something different in essential characteristics	*mudança, troca, transformação, permuta, substituição, transmutação, conversão, imutação, câmbio, comutação, convertimento, resmuda, comuta*
(n) character.2, lineament.2, quality 4 • a characteristic property that defines the apparent individual nature of something	*natureza, perfil, carácter, compleição, jaez, índole, temperamento, crase, cariz, carnadura, natura, idiossincrasia, naturaleza, compreição, catástase, mescla*
(n) cognition.1, knowledge.1 • the psychological result of perception and learning and reasoning	*estudo, conhecimento, saber, cultura*
(n) cognitive content.1, content.2, mental object.1 • the sum or range of what has been perceived, discovered, or learned	*informação, conhecimento, saber, instrução, racionalidade, ciência, sabedoria, erudição, consciencialização, conscientização, gnose, sofia, noção, perícia, sensatez*
(n) color.2, coloring.2, colour.2, colouring.2 • a visual attribute of things that results from the light they emit or transmit or reflect	*colorido, cor, tom, matiz, pintura, coloração, tonalidade, tinta, pincel, pigmento, color, caiadela, cromia*
(v) communicate.1, intercommunicate.1, transmit feelings.1, transmit thoughts.1 • transmit thoughts or feelings	*transmitir, comunicar, participar, noticiar, legar, apegar, vocalizar*
(n) communication.1 • something that is communicated between people or groups	*diálogo, convivência, comunicação, conversação, colóquio, contato, convívio*
(n) concept.1, conception.3 • an abstract or general idea inferred or derived from specific instances	*opinião, ideia, idéia, noção, conceito, concepção*
(n) condition.5, status.2 • a condition or state at a particular time	*estado, caso, situação, circunstância, coisa, condição, posição, modo, maneira, disposição, colocação, postura*
(n) consequence.3, effect 4, outcome.2, result.3, upshot.1 • a phenomenon that follows and is caused by some previous phenomenon	*consectário, resultado, consequência, conclusão, decorrência, derivação*
(v) consume.2, have 8, ingest.2, take.16 • serve onself to, or consume regularly	*consumir, tomar, chupar, exaurir, exaustar, absorver, ensecar, ingerir, haurir, esvaziar, devorar, beber, gastar, desfalcar, obliterar, maquiar, desfalcoar, dissipar, esgotar, sumir, roubar, absumir*
(v) convey.1, impart.1 • make known; pass on, of information	*passar, transmitir, comunicar, propagar, contagiar, inocular*
(n) course 7, trend.3 • general line of orientation	*corrente, decurso, fluxo, duração, curso, correnteza*
(v) cover.16 • provide with a covering	*proteger, cobrir, defender, acobertar, encobertar*
(v) create.2, make.13 • cause to be or to become	*gerar, criar, produzir, incubar, causar, provocar, desenvolver, derivar, originar, implicar, proporcionar, desencadear, propor, crescer, interessar, transportar, animar, facilitar, atrair, inspirar, arrastar, comandar, motivar, inventar, evocar, captar, pregar, instituir, acarretar, fomentar, induzir, incutir, germinar, infundir, ensejar, carrear, engenhar, surtir, procriar, determinar, suscitar, acarear, carretar, acarrear, render, custar, catalisar, carrejar, carretear, acarrejar, uberar, chamar, lançar, antemover, chimpar, importar, predispor, avivar, carrar, agomar, levantar, alevantar, puxar, determinhar, trazer, dar, ser, azar, mover, promover, ocasionar, sugerir, desferir, infligir, bracejar, retirar, abotoar, retornar, chover, lucrar*
(v) decrease.5, diminish.1, fall.11, lessen.1 • decrease in size, extent, or range	*minuir, diminuir, reduzir, encolher, encurtar, acurtar, minguar, decrescer*

Continued on next page...

Table B.2 – continued from previous page

WordNet synset	Onto.PT synset
(n) definite quantity.1 • a specific measure of amount	N/A
(n) development.1 • act of improving by expanding or enlarging or refining	*desenvolvimento, extensão, ampliação, expandidura, alargamento, incremento, prolongamento, dilatação, crescimento, expansão, prorogação*
(n) direction 7, way.8 • a line leading to a place or point	*sentido, destino, direcção, via, orientação, caminho, linha, direção, rumo, roteiro, rota, derrota, senda, coordenada, tramontana, travia, corrume, guião, esteira, método, regra, senso, lado, vieiro*
(n) disorder.1 • a disturbance of the peace or of public order	*agitação, desordem, alvoroço, bulha, distúrbio, revolução, trabuzana, perlenga, embrulhada, comoção, perturbação, motim, conturbação, revolta, tumulto, espalhafato, rebuliço, bernarda, alvoroto, trinta-e-um, desestabilização, borborinho, amotinação, matinada, parlenga, rebúmbio, alevanto, revoldaina, emoção*
(n) distance.1 • the space between two objects or points	*aberto, espaço, distância, abertura, afastamento, separação*
(v) emit.2, express audibly.1, let loose.1, utter.3 • utter sounds; not necessarily words	*proferir, largar, emitir, soltar, exalar, desatar, desfechar, rutilar, dar, despedir*
(n) event.1 • something that happens at a given place and time	*situação, acontecimento, evento, vicissitude, possibilidade, lance, acaso, acidente, eventualidade, peripécia, contingência*
(v) evince.1, express.6, show.10 • give expression to	*parecer, expor, passar, contar, proferir, manifestar, expressar, exprimir, mostrar, palrar, palrear, descortinar, desencobrir, figurar, patentear, expirar, comunicar, respirar, adicar, revelar, vender, divulgar, desembuçar, descerrar, desencerrar, evidenciar, esclarecer, aclarar, desvelar, desvendar, desenterrar, clarear, demostrar, delatar, estiar, deslacrar, desaferrolhar, desembaciar, escogitar, desenfardar, desencovar, ostender, dessoterrar, desencantoar, dessepultar, palear, deseclipsar, empenhar, franquear, patentizar, descobrir, confiar, romper, entregar, perfurar, destapar, desnudar, descapota, inaugurar, detectar, cicatrizar, cheirar, categorizar, quantiar, desentaipar, esfossilizar*
(v) experience 7, get.18, have.11, receive 8, undergo.2 • of mental or bodily states or experiences	*ver, viver, experienciar, passar, provar, experimentar, vivenciar, colher, ensaiar, saborear, chincar*
(v) express.5, give tongue to.1, utter.1 • express verbally	*articular, dizer, fazer, ler, pronunciar, proferir, recitar, declamar*
(n) feeling.1 • the psychological feature of experiencing affective and emotional states	*alma, sentimento, convicção, sentir, desejo, sensação, coração, sensibilidade, pêsames, estesia, estese*
(n) form.1, shape.1 • the spatial arrangement of something as distinct from its substance	*físico, aspecto, forma, figura, tremenho, perfil, configuração, morfologia, construtura, compleição, formato, feição, talhe, feitio, guisa, silhueta, figuração, laia, carácter, hábito, maneira*
(n) form.6, pattern.5, shape.5 • a perceptual structure	*exemplo, forma, modelo, norma, molde, gabarito, cofragem, modêlo, fôrma*
(v) furnish.1, provide.3, render.12, supply.6 • provide or furnish with	*fornecer, abastar, suprir, repor, prover, abastecer, munir, guarnecer, rechear, provisionar, aprovisionar, fornir, municionar, vitualhar, equipar, municiar, contribuir, dotar, sortir, refazer, aperceber*
(v) get hold of.2, take.17 • get into one's hands, take physically	*colher, tirar, tomar, retirar, sacar, recolher*
(v) give.16 • transfer possession of something concrete or abstract to somebody	*dar, transferir, conceder, oferecer, atribuir, deitar, devolver, conferir, permitir, ceder, emprestar, doar, transmitir, imprimir, outorgar, deferir, tributar, arbitrar, dispensar, brindar, presentear, dadivar, depositar, deliberar, adjudicar, creditar, desasir, alvidrar, ...*
(v) go.14, locomote.1, move.15, travel.4 • change location; move, travel, or proceed	*ver, andar, vêr, cobrir, copular, frequentar, caminhar, transitar, cursar, visitar, viajar, percorrer, tramitar, peregrinar*
(n) happening.1, natural event.1, occurrence.1 • an event that happens	*sucedido, ocorrido, acontecido, facto, ocorrência, sucesso, êxito, acção, acontecimento, evento, efeméride, intercorrência, feiteira, sucedimento, sucedo, sucedenho*
(v) have.12, have got.1, hold.19 • have or possess, either in a concrete or an abstract sense	*dispor, ter, dominar, possuir, apossuir*
(n) idea.2, thought.2 • the content of cognition; the main thing you are thinking about	*entendimento, opinião, ideia, espírito, pensamento, conceito, concepção, aviso, apreciação, juízo, pensar, vêr*
(n) improvement.1 • the act of improving something	*melhoramento, melhoria, bem-feitoria, melhora*
(v) increase.7 • become bigger or greater in amount	*desenvolver, medrar, aumentar, incrementar*
(n) information.1 • knowledge acquired through study or experience	*informação, conhecimento, saber, instrução, noção, racionalidade, sensatez, ciência, perícia, sabedoria, erudição, consciencialização, conscientização, gnose, sofia*
(v) kill.5 • cause to die	*jugular, matar, chacinar, assassinar, massacrar, fuzilar, trucidar*
(n) know-how.1, knowhow.1 • the knowledge and skill required to do something	*uso, exercício, atividade, prática, experiência, perícia, aprendizado, tirocínio, tarimba*
(n) locomotion.1, travel.1 • self-propelled movement	*deslocação, deslocamento, movimentação, circulação, mobilização, desarticulação, locomoção, luxação*
(n) magnitude relation.1 • a relation between magnitudes	N/A
(n) message.2, content.3, subject matter.1, substance.4 • what a communication that is about something is about	*fundo, contento, conteúdo, assunto, matéria, teor, texto, pratinho*
(n) method.2 • a way of doing something, esp. a systematic one; implies an orderly logical arrangement (usually in steps)	*modo, maneira, sistema, processo, método, fórmula, ordem, conduta, técnica, procedimento, arrumação, signo*
(n) motion.5, movement.6 • a natural event that involves a change in the position or location of something	*movimento, mudança, deslocação, deslocamento, moção, remoção, translação*
(v) need.5, require.3, want.5 • have need of	*precisar, falecer, necessitar, carecer*
(v) need.6 • have or feel a need for	*querer, procurar, requerer, precisar, padecer, reclamar, pedir, exigir, necessitar, carecer, requisitar, demandar*
(n) path.3, route.2 • an established line of travel or access	*giro, decurso, itinerário, carreira, passagem, percurso, caminho, viagem, trajecto, distância, curso, rumo, derrota, discurso, varadouro, trajeto, currículo, trajetória, trajectória, correnteza, decorrer*
(n) phenomenon.1 • any state or process known through the senses rather than by intuition or reasoning	*fenómeno, fenômeno, milagre, maravilha, prodígio, bons-dias*

Continued on next page...

Table B.2 – continued from previous page

WordNet synset	Onto.PT synset
(n) production.1 • the act of producing something	*produção, fabrico, criação, construção, composição, elaboração*
(n) property.2 • an attribute shared by objects	N/A
(n) psychological feature.1 • a feature of the mental life of a living organism	*carácter, caráter, propriedade, expressão*
(n) quality.1 • essential attribute of something or someone	*particular, característico, variante, entidade, individualidade, justeza, particularidade, peculiaridade, propriedade, característica, caraterística, especialidade, singularidade, especificidade, congruência, pontualidade, anomalia, particularismo, pormenorização, vernaculidade, invulgaridade, celebreira*
(n) ratio.1 • the relative magnitudes of two quantities (usually expressed as a quotient)	*termo, razão, taxa, percentagem, proporção*
(n) relation.1 • an abstraction belonging to or characteristic of two entities or parts together	*relação, semelhança, afinidade, analogia*
(n) relationship.1 • often used where "relation" would serve; preferred usage of "relationship" is for personal relations or states of relatedness	*relação, ligação, vinculação, concernência*
(n) relationship.3 • a state of connectedness between people (especially an emotional connection)	*romanço, romance, caso, conto, narrativa, história, relacionamento, narração, novela, descrição, fábula, rimance, diegese, raconto, mussosso*
(v) remember.2, think of.1 • keep in mind for attention or consideration	*ver, vêr, celebrar, evocar, mentalizar, rever, mentar, lembrar, recordar, ementar, inventariar, amentar, reviver, escordar, comemorar, repassar, relembrar, revistar, reconstituir, revisitar, rememorar, assoprar, memorar, alembrar, revivescer, prègar, remomerar, revivecer, remembrar, reviviscer*
(v) remove.2, take 4, take away.1 • remove something concrete, as by lifting, pushing, taking off, etc.; or remove something abstract	*separar, tirar, eliminar, estremar, desviar, levar, deslocar, abrir, arrancar, extrair, sacar, extractar, exturquir, demitir, escolher, excluir, revirar, avocar, demover, repartir, remover, afastar, apartar, distrair, desunir, repelir, delimitar, vastar, arredar, subtrair, abduzir, disjungir, esconjuntar, desachegar, abjugar, distanciar, frustrar, divergir, dissuadir, baldar, despregar, espaçar, marginar, espaciar, despartir, espacejar, entrelinhar, faiar, desaconchegar, deflectir, desapartar, desaviar, desconchegar, desaproximar, desaquinhoar, aleixar, desarredar, amover*
(v) represent.3 • serve as a means of expressing something	*representar, descrever, figurar, debuxar, retratar*
(v) say.8, state.7, tell.7 • express an idea, etc. orally, in writing, or with gestures	*dizer, ter, dar, dirigir, brotar, pronunciar, proferir, publicar, declarar, indicar, manifestar, emitir, soltar, expressar, mencionar, decretar, exprimir, rezar, enunciar, enumerar, declinar, caducar, exteriorizar, externar*
(n) sign.3, signal.1, signaling.1 • any communication that encodes a message	*sinal, mostra, amostra, indicação*
(n) situation.4, state of affairs.1 • the general state of things; the combination of circumstances at a given time	*passo, ponto, situação, conjuntura, circunstância, contexto*
(n) social relation.1 • a relation between living organisms; esp between people	*comércio, convivência, trato, relação, relações, respondência*
(n) space.1 • the unlimited 3-dimensional expanse in which everything is located	*universo, espaço, região, área, domínio, terra, orbe, esfera, grandeza, campo, âmbito, círculo, reino, abrangência*
(n) spacing.1, spatial arrangement.1 • the property possessed by an array of things that have space between them	*longada, distância, afastamento, irradiação, distanciamento, espaçamento, espacejamento, desvizinhança, longinquidade*
(n) spatial property.1, spatiality.1 • any property relating to or occupying space	N/A
(n) state.1 • the way something is with respect to its main attributes	*estado, caso, situação, circunstância, coisa, condição, posição, modo, maneira, disposição, colocação, postura, circunstâncias*
(n) structure 4 • the complex composition of knowledge as elements and their combinations	*forma, organização, estrutura, constituição, configuração, morfologia, conformação, construtura, arquitectónica*
(n) time.1 • the continuum of experience in which events pass from the future through the present to the past	*idade, tempo, período, época, temporada, ocasião, estação, fase, século, etapa, era, lua, quadra*
(n) unit.6, unit of measurement.1 • any division of quantity accepted as a standard of measurement or exchange	*medida, regra, norma, cânone, marco, baliza, bitola, escantilhão, linda, padrão, cravo, compasso, diapasão, termómetro, craveira, mesura, estalão, medida-padrão, cômpito, xeura, gueja, mensura, predisposições, comensuração, ferrete, nível*
(n) visual property.1 • an attribute of vision	N/A