• U  C •

Thesis Proposal
Doctoral Program in Information Sciences and Technologies
Artificial Intelligence

Ontology Learning for Portuguese

Hugo Gonçalo Oliveira

Thesis Advisor
Paulo Jorge de Sousa Gomes

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

25th September 2009

Abstract

Having in mind both the importance that semantic information plays nowadays in natural language processing, as well as the work involved in creating lexical resources from the scratch, this research aims to create a lexical ontology for Portuguese by semi-automatic means.

While, for English, WordNet (Fellbaum (1998)) established as the standard model of a lexical ontology, for Portuguese, the few existing similar resources, created manually, are either on earlier stages of development or not publicly available for download and entire use.

Therefore, as an alternative to manual creation and maintenance of such resources, the work proposed here is concerned with the development of computational tools capable of extracting lexico-semantic knowledge from Portuguese textual resources. The knowledge acquired will then be organised into a public domain lexical ontology.

The extraction procedures will be based on the detection of textual patterns that are indicative of lexico-semantic relations between terms. Machine-readable dictionaries will be used as the primary source of knowledge, since they are already structured around words and their meanings, they typically use simple vocabulary, they were created by experts and they are the main source of general knowledge. However, this work will not be limited by processing dictionaries so, textual corpora will be used as the second source of knowledge, in order to enrich the the ontology in several more specific domains.

Furthermore, the quality and utility of the resources developed will be assessed. Besides manual evaluation, and considering the time needed to perform the latter, automatic evaluation methodologies, inspired by related work, will be applied. In the end of this research, important contributions to the computational processing of Portuguese are expected, such as a new public domain lexical resource and computational tools capable of learning lexico-semantic information from text.

Resumo

Actualmente, a informação semântica desempenha um papel muito importante no processamento de linguagem natural. No entanto, os recursos lexicais são normalmente o resultado de trabalho manual intensivo. Tendo isto em conta, nesta investigação pretende-se construir, de forma semi-automática, uma ontologia lexical para a língua portuguesa.

Enquanto que, para o inglês, a WordNet (Fellbaum (1998)) se estabeleceu como o modelo paradigmático de ontologia lexical, para o português, os poucos recursos existentes, criados manualmente, ou se encontram ainda numa fase inicial ou então a sua utilização completa não é gratuita.

Por isso, como alternativa à criação e manutenção manual de recursos lexicais, no trabalho aqui proposto pretendem-se desenvolver ferramentas computacionais capazes de extrair conhecimento léxico-semântico a partir de recursos textuais escritos em português. O conhecimento extraído será depois organizado numa ontologia lexical de domínio público.

A extracção será baseada na detecção de padrões textuais indicadores de relações léxico-semânticas entre termos. Dicionários electrónicos serão o primeiro recursos a ser explorado, tendo em conta que já se encontram estruturados de acordo com as palavras e os seus significados, usam normalmente vocabulário simples, foram criados por especialistas e são talvez a principal fonte de conhecimento geral. No entanto, este trabalho não se limitará a processar dicionários, por isso, corpora textual será utilizado como uma segunda fonte de conhecimento, de forma a enriquecer a ontologia em vários domínios mais específicos.

Pretende-se ainda que a qualidade e utilidade dos recursos desenvolvidos sejam avaliadas. Além da habitual avaliação manual, e tendo em conta que esta será morosa, serão aplicadas metodologias de avaliação automática de ontologias, inspiradas em trabalho relacionado. No fim deste trabalho esperam-se contribuições importantes para o processamento computational da língua portuguesa, de onde se destaca um novo recursos lexical do domńio público e ainda ferramentas capazes de extrair informação léxico-semântica a partir de texto.

Contents

List of Figures

List of Tables

Acknowledgements

Hugo Gonçalo Oliveira is supported by the FCT scholarship grant SFRH/BD/44955/2008.

Glossary

- **AHD3**: American Heritage Dictionary, 3rd Edition
- **AHFD**: American Heritage First Dictionary
- **AI**: Artificial intelligence
- **CBC**: Clustering By Committee
- **LDOCE**: Longman Dictionary of Contemporary English
- **LKB**: Lexical Knowledge Base
- **LSA**: Latent Semantic Analysis
- **MPD**: Merry Webster's Pocket Dictionary
- **MRD**: Machine readable dictionary
- **NER**: Named entity recognition
- **NLP**: Natural language processing
- **PMI**: Pointwise mutual information
- **POS**: Part of Speech
- **Q&A**: Question & answering
- **RMRS**: Robust Minimal Recursion Semantics
- **W7**: Webster's Seventh New Collegiate Dictionary
- **WSD**: Word sense disambiguation

Chapter 1

Introduction

There is a growing number of computer applications that perform tasks where semantic knowledge is needed. Tasks that go from automatic generation of text to intelligent search and machine translation, as well as writing aids. These applications demonstrate that natural language processing (NLP) (Jurafsky and Martin (2000)) is becoming more and more dependent on semantic information and thus, computational access to such knowledge is needed.

For instance, consider the following sentence:

The cat has four wheels.

Despite there is a problem in this sentence, a regular spell checker would not be able to detect it, because this problem occurs at the semantic level. Semantic knowledge is usually encoded in lexico-semantic resources, such as lexical ontologies, that are important tools to help the achievement of NLP tasks where understanding the meaning of texts is critical. Lexical ontologies are models aiming to represent the lexical structure and thus, the meaning of a language, as opposing to terminologies or domain ontologies, whose purpose is to describe specific topics or domains.

In order to find out the problem in the example sentence, a lexical ontology could be used along with a spell checker. It would provide semantic information that would make it possible to find out that there were no known relationships between *cat* and *wheels*, which suggests that their co-occurrence in one sentence is quite odd. At the same time, a simple algorithm could be used to find out that *wheels* have however a relationship with a word whose spelling is very close to *cat*, more precisely *car*, which is actually known to have *wheels*. Therefore, a semantically-aware spell checker would be able to suggest the writer of the sentence to change *cat* into *car*.

1.1 Motivation

While for English, despite some known issues, WordNet (Fellbaum (1998)) established as the standard model of a lexical ontology, for Portuguese, and other non-English languages, the situation is quite different. Similar resources for Portuguese are either currently in earlier stages of development (WordNet.BR (Dias da Silva et al. (2002))) or not publicly available for download and free usage (WordNet.PT (Marrafa (2002)) or MultiWordNet.PT[1]). Moreover, all of these resources are the

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

result of time-consuming manual effort.

The truth is that while there is intensive labour involved in manually encoding lexical entries, lexical capabilities of NLP systems will always be weak (Briscoe (1991)). Handcrafting ontologies is impractical and undesirable and we should take advantage of available NLP tools in order to automate part of this task, reducing the need of manual input (Brewster and Wilks (2004)).

This is why we believe that contributions regarding the automatic, or semi-automatic, construction of such a resource for Portuguese should be considered as an alternative and the subject of research, in order to avoid time-consuming human work in its construction and maintenance. Moreover, we also believe that this kind of resource should be in the public domain, and thus available for the Portuguese NLP community, or other researchers and developers that work with the Portuguese language. This is probably the best way to establish a broad community of users, from which feedback can be gathered.

In addition to the example given in the beginning of this chapter, the integration of a lexical ontology in an application can be used to accomplish much more NLP tasks, such as semantically driven information retrieval. In a more complex utilisation example, consider the following three text snippets:

Snippet A	`A gripe é causada por um vírus altamente contagioso que afecta aves e mamíferos. Tipicamente, a gripe é transmitida por mamíferos infectados por meio do ar e por aves infectadas por meio de suas secreções.`
Snippet B	`A varicela é causada por um vírus altamente contagioso que afecta essencialmente crianças. Tipicamente, a varicela é transmitida através da inalação de gotículas presentes no ar, que contêm o vírus.`
Snippet C	`O mal pode facilmente atingir várias pessoas e os seus principais sintomas são calafrios, febre alta, dores de garganta, dores de cabeça e fadiga.`

Regarding that a user wanted to group these snippets according to their similarity, a typical keyword and frequency based approach would most probably put snippets A and B in the same group, and leave snippet C in a different one. Nevertheless, depending on the purpose of the search, snippet C can be considered to be very related to the previous ones, because while A and B describe how two diseases can be transmitted, C describes the symptoms of some disease.

Assuming that the search algorithm has access to a Portuguese lexical ontology with semantic information on health or medicine, it would be possible to notice two facts, or relations, suggesting that snippets A and B are both about *doenças* (diseases), more precisely:

gripe is-a *doença*
varicela is-a *doença*

While in one hand, no relations would be found directly between snippets A and B, on the other hand, there are many relations that suggest that snippets A and C are closely related, such as the following:

gripe is-a *doença*
doença has-synonym *mal*
afectar has-synonym *atingir*
pessoa is-a *mamífero*
gripe causes *calafrios*
gripe causes *febre alta*
gripe causes *dores de cabeça*
gripe causes *dores de garganta*
gripe causes *fadiga*

So, if the user searches for texts about *doenças*, a semantic driven approach would be able to return all the three snippets. Otherwise, if the user queries the system for texts about *gripe* (flu), only snippets A and C would be returned because symptoms of *gripe* are different than the symptoms of *varicela* (chickenpox), encoded in the following relations:

varicela causes *erupções cutâneas*
varicela causes *febre baixa*

Among a substantial quantity of semantically-driven NLP tasks, the access to these relations can as well be useful for giving answers in a generic question & answering (Q&A) (Strzalkowski and Harabagiu (2006)) system. For instance, it would be possible to answer the following questions:

- *O que é a gripe?*
  – Uma doença.

- *Que sintomas tem a gripe?*
  – Calafrios, febre alta, dores de cabeça, dores de garganta e fadiga.

Furthermore, information could be crossed in order to give a slightly more complex answer to the first question:

- *O que é a gripe?*
  – Uma doença ou mal que provoca calafrios, febre alta, dores de cabeça, dores de garganta e fadiga.

1.2 Research Goals

The main goal of this research is to **create a lexical ontology for Portuguese**, which will be called **Onto.PT**, by semi-automatic means. Therefore, this research will focus on the design and development of the adequate computational tools for exploiting machine-readable textual sources, in order to acquire the knowledge needed for the construction of Onto.PT. These tools shall also provide further updates to the base ontology, after processing previously unprocessed textual resources.

The development of the extraction tools will be based the earlier intuition (Hearst (1992)) that the presence of certain textual patterns can indicate a particular semantic relationship between two terms. For instance, after analysing the following sentence:

A fábula ***é um tipo de*** *narrativa que* ***tem o objectivo de*** *entreter e aconselhar.*

It is possible to take advantage of the patterns in bold to acquire lexico-semantic knowledge, which can be translated by the following relations:

fábula is-a *narrativa*
fábula has-pupose *entreter*
fábula has-pupose *aconselhar*

Despite much of the related work being more concerned with learning similar terms and simple relations between them, such as the is-a relation, we are also concerned with the extraction of other interesting relations, such as the has-purpose relation and other relations involving an agent or process and an effect or result.

The knowledge acquired can furthermore be organised and structured into a lexical ontology to ease its integration with other applications. It is important to state that we intend Onto.PT, so as the developed tools, to be in the public domain and freely available for download, so that in a near future they can be used by the Portuguese NLP community and also by other researchers that need Portuguese lexical knowledge in their work.

It is also our intention to apply several validation methodologies, or to develop some new methodologies, to ensure the quality of our results. The validation procedures should require the least human intervention possible, so that they can be repeated as many times and whenever needed (e.g. for each new available version). Moreover, human judgements are always more prone to subjectivity than automatic procedures. However, regarding its reliability, we do not discard manual evaluation completely.

One of the challenges involved is that for Portuguese, as well as for the majority of other non-English languages, the amount of existing NLP resources are scarce, so we will have to come up with new ideas or eventually recycle old ones.

1.3 Approach and Expected Contributions

In order to extract the knowledge needed to create Onto.PT, two different kinds of textual resources will be exploited:

- Machine readable dictionaries (MRDs) will be used as the primary source of knowledge, since they are perhaps the most important source of general knowledge.

- As suggested by several authors (Hearst (1992)), textual corpora will be explored in order to enrich the base resource, because general knowledge seems to be insufficient or inadequate for most NLP tasks (Riloff and Shepherd (1997); Roark and Charniak (1998); Caraballo (1999)).

The knowledge extracted from both resources will then be adequately merged and organised in a lexical ontology. During the research process, several evaluations will take place, and it is also our aim to automate some of the evaluation procedures. Even though, we believe that, at least in some point, manual evaluation should have to be performed, as a consequence of its reliability.

To sum it up, at the end of this research, the following contributions are expected:

- Onto.PT, a new lexical ontology for Portuguese, created by semi-automatic means;
- Computational tools for semi-automatic:
  - Extraction of lexico-semantic knowledge from MRDs;
  - Extraction of lexico-semantic knowledge from textual corpora;
  - Organisation of lexico-semantic knowledge (extracted from different sources) into Onto.PT;
  - Browsing and updating Onto.PT.
- Methodologies to evaluate lexical ontologies.
- Several scientific papers about the most relevant conclusions and results, and also a PhD thesis, describing all the work done.

1.4 Outline of the proposal

In **Section 2**, background concepts, important to understand this research, are introduced. This includes an introduction to NLP (Section 2.1) and a section that converges to the notion of lexical ontology (Section 2.2). The latter section comprises the topic of lexical semantics (Section 2.2.1), the most important lexical resources and representations (Section 2.2.2), a discussion concerning the notion of ontology (Section 2.2.3) and finally presents several well-known lexical ontologies (Section 2.2.4).

In **Section 3**, an overview on related work is made. It comprises, more precisely, research work on the acquisition of lexical and semantic knowledge from MRDs (Section 3.1), on ontology learning from textual corpora (Section 3.2) and on the evaluation of ontologies (Section 3.3).

In **Section 4**, the approach to be followed during this research is discussed. This comprises the description of several tasks involved in each one of the research phases.

In **Section 5**, the research working plan is presented along with the reference to some current work and also to conferences and journals where we aim to publish some our results and conclusions.

Finally, **Section 6** concludes this proposal with some remarks.

Chapter 2

Background

This chapter addresses background concepts, important for understanding this research. Since NLP (Jurafsky and Martin (2000)) is the basis of this work, the chapter starts with a brief introduction to this topic. All the levels involved in NLP are presented and followed by a concise description of the major NLP tasks and a section dedicated to Portuguese NLP, which is the language focused in this research.

Another important concept concerning this thesis is the notion of lexical ontology, which is addressed right after NLP. Existing lexical ontologies are presented after introducing the topic of lexical semantics, possible ways to represent and organise lexical knowledge into one resource, and discussing what is after all an ontology.

2.1 Natural Language Processing

The topic of Natural Language Processing (NLP) (Jurafsky and Martin (2000)) is commonly introduced by 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 HAL in Stanley's Kubrick *2001: A Space Odyssey* or Bender and other robots in Matt Groening's *Futurama*.

NLP is a field of artificial intelligence (AI) 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. Since natural language, used by humans for communication, is probably the most natural way for encoding, transmitting and reasoning about knowledge, most knowledge repositories are in a written form (Santos (1992)). Therefore, the emergence of the NLP field from AI should not seem surprising.

One of the main problems concerning natural languages is that it differs from formal languages (e.g. programming languages) because in the latter, each symbol has only one possible meaning while in the former a symbol can have different meanings depending on the context it is used at. Ambiguity occurs when it is not possible assign a single meaning to a form of communication, because it can be interpreted in more than one way. In natural language, ambiguity can occur at several levels, namely phonology, morphology, syntax, semantics, pragmatics and discourse. A good example of a sentence that can be ambiguous in all the levels is given in Jurafsky and Martin (2000):

I made her duck.

The following sections will introduce all the six levels of knowledge needed to achive complete NLP. After presenting each level, it will be pointed out how the example sentence can be ambiguous in that level.

2.1.1 Phonology

Phonology involves analysing the sounds of speech and converting them into symbols. We will not go further in this level, since this research will only be dealing with written text.

For the example sentence, the words *I* and *made* are ambiguous at the phonetics level, because they can also be converted to *eye* and *maid*, respectively.

2.1.2 Morphology

Morphology deals with the identification, analysis and description of the structure of words. Usually considered as the smallest unit of a sentence, words can be related with other words, according to the patterns used in their construction. The regularities involved enable the determination of each word's morphological category (or categories), the identification of its base form or headword, which is called the lemma, and also some other characteristics depending on the word's category, such as its gender, number or tense (for verbs). For nouns, the lemma corresponds to the word form in the masculine gender and singular number, while for verbs, it corresponds to the infinitive form.

Here are two examples of the morphological analysis of two nouns and two verbs:

- The words *car* and *cars* are both masculine noun forms of the lemma *car*, but the former is in the singular while the latter is in the plural.
- The words *makes* and *making* are all verb forms of the lemma *make*. The former is the third person of the singular (he/she/it) of the present tense and the latter is the gerund form.

Back to the example sentence, the word *duck* is morphological ambiguous because it can either be a noun or a verb. The word *her* is also morphological ambiguous because it can be a dative pronoun or a possessive pronoun.

2.1.3 Syntax

Syntax deals with the study of structural relationships between words in a sentence. Words can be grouped according to their functions and, depending on their place and on their neighbours, they can have different parts of speech (POS) that are usually one of the possible morphological categories for the word.

Syntactic analysers can be limited to classify words according to their POS which is per se very important, since it gives us substantial information about the word and its most probable neighbours. For example, possessive pronouns are likely to be followed by nouns while personal pronouns are usually followed by verbs. Moreover, the POS can tell us something about how the word is pronounced, it can be very helpful in more complex NLP tasks, such as information retrieval (IR) or word sense

```
Sentence → NP VP
VP → Verb NP
NP → the man
NP → the book
Verb → took
```

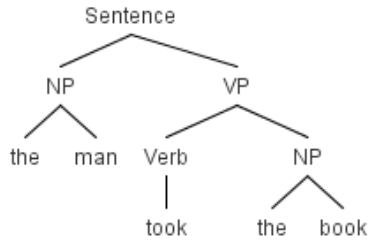

Figure 2.1: The first context-free grammar parse tree (adapted from Chomsky (1956)).

disambiguation (WSD), and also for frequency studies of particular constructions. Here is an example of POS tagging:

The/det *car*/n *has*/v *wheels*/n.

Other syntactic analysers, typically based on context-free grammars (Chomsky (1956)), achieve full parsing and identify not only the POS of words but also:

- the relations between the constituents of the sentence, and group them into phrases (e.g. noun phrase, verb phrase, prepositional phrase);
- the function of a word inside a sentence (e.g. subject or object);
- dependencies between words or phrases.

Given a context-free grammar, a sentence can be derived and the above information is represented as a syntactic tree. Figure 2.1 shows a grammar and its derivation tree for the sentence *The man took the book*.

Ambiguities in syntactic analysis may occur in very distinct situations. In the example given in the beginning of this section, syntactic ambiguities occur together with the morphological ambiguity of the words *duck* and *her*. Additionally there is an ambiguity in the verb *make*: it can take only one direct object, and thus be transitive; but it can also take two objects (*her* and *made*), meaning the first object got made into the second; or it can take a direct object (*her*) and a verb (*made*), meaning that the object caused to perform the verb.

2.1.4 Semantics

Semantics studies the meaning of a language. In order to do that, natural language is mapped to a formal language, enabling the interpretation of words, phrases, sentences and texts by machines.

There are several ways for representing meaning in a formal language, namely logical predicates (Smullyan (1995)), directed graphs and semantic frames (Fillmore (1982)). Figures 2.2, 2.3 and 2.4 show different representations of the meaning of the sentences:

The bottle contains wine. Wine is a beverage.

```
contains(bottle, wine)
isa(wine, beverage)
```

Figure 2.2: Meaning representation 1: Logical predicates

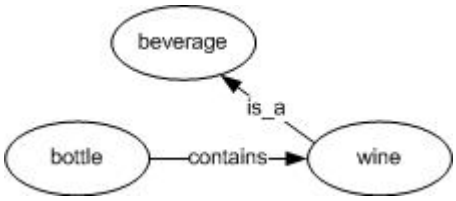

Figure 2.3: Meaning representation 2: directed graph.

Having in mind the principle of compositionality (Frege (1960)), transposed to the scope of words and sentences – the meaning of a sentence is composed from the meaning of its parts – the syntactic structure of a sentence can be used as an input for semantic analysis, leading to a syntax-driven semantic analysis. However, the syntactic structure is sometimes not well-suited for semantic analysis, because key semantic elements are often widely spread across syntactic structures; some syntactically motivated constituents play no essential role in semantic processing; as well as the general nature of many syntactic constituents that may result in semantic attachments that give rise to meaning representations without any useful purpose.

Semantic grammars (Brown and Burton (1975)) can be used to perform a more direct and sensible semantic analysis. The grammars, where rules and constituents are designed to deal directly with the entities and relations of the domain, have semantically driven rules with the key semantic components occurring together. There are however several limitations, since these grammars are usually specific for some kind of text or domain, which limits their potential of being reused.

It should however be pointed out that semantic grammars follow the same principles and formalisms as syntactic grammars, and they are both used to analysed text. The only difference relies on their final purpose, that may change the way rules are designed.

Furthermore on semantics, it should be referred that the same sentence can have incompatible meaning representations, depending, for instance, on the situation it occurs, which leads to ambiguities at the semantic level. In the example sentence, the verb *make* is semantically ambiguous because it can either mean *create* or *cook* and both of these verbs would give rise to different meaning representations. WSD is the process for determining which sense of a word is being used. This topic is slightly discussed in Section 2.2.1 of this proposal. Also in Section 2.2.1, the topic of lexical semantics (Cruse (1986)), which the subfield of semantics that studies the

```
wine
  isa:  beverage
  container:  bottle
```

Figure 2.4: Meaning representation 3: frame *wine*.

words and their meanings, is further discussed.

2.1.5 Pragmatics and discourse

Pragmatics studies how language is used to accomplish goals, according to the context. People, objects and situations involved can all lead to different ways to convey the same message. In order to acquire information about the context, knowledge about the whole discourse, as opposing to knowledge on specific parts or inside the same sentence, is needed. Discourse analysis studies inter-sentence relationships or, in other words, it identifies relations between units larger than a sentence.

Analysing words that can have a different meaning than the usual (e.g. figures of speech or stylistic and rhetorical devices) is typically involved in discourse analysis. One example of this kind of analysis is anaphora resolution, that deals with the identification of expressions referring to other expressions. For example, in the following sentences, the word *he* is related with *Joe*.:

Joe did not go to work. He is sick.

2.1.6 Major NLP Tasks

Combining the aforementioned basic levels of NLP for different purposes, more complex tasks can be developed. The following are among the major NLP tasks:

- *Machine translation* (MT) (Hutchins and Somers (1992)): automatic translation of text written in one natural language to another.
- *Word Sense Disambiguation* (WSD) (Nancy Ide (1998)): selection of the most adequate sense of a word in a context.
- Information Retrieval (IR) (Salton and McGill (1983)): a task concerned with locating documents, other natural language resources, or information within them, according to some user's query.
- *Information Extraction* (IE) (Grishman (1997)): the generic task of automatically extracting structured information from unstructured natural language inputs. IE generally encompasses several steps, where other NLP tasks, such as named entity recognition (NER), relation detection or temporal analysis, are performed.
- *Named Entity Recognition* (NER) (Chinchor and Robinson (1997)): identification and (sometimes) classification of proper nouns, more precisely names of persons, organisations, places, events and pieces of art, expressions of time, quantities and monetary values, and sometimes even abstractions.
- *Question & Answering* (Q&A) (Strzalkowski and Harabagiu (2006)): automatic answering to natural language questions.
- *Anaphora resolution* (Mitkov (1999)): identification of anaphoras and determination of the expressions or entities they are referring to.

- *Automatic Summarisation* (Mani and Maybury (1998)): automatic creation of a shortened, summarised version of a natural language text.
- *Natural language generation* (NLG) (Reiter and Dale (1997)): automatic generation of natural language.
- *Speech recognition* (hui Lee and hwang Juang (1996)): conversion of spoken words into a machine-readable input.

2.1.7 Portuguese NLP

NLP "was born" for English language and sometimes it is still only seen as the processing of English, almost ignoring that other languages exist. As it is supported by Santos (1999), the computational processing of a non-English language, in this case, Portuguese, should not be limited to do exactly the same things that have been done for English. For instance, each language has a different lexicon and a different grammar. Additionally, each language develops in a different culture, so reasoning and methodology should be applied for studying the target language, instead of assuming that what has been for English is appropriate. NLP should then be specific to each language, in our case Portuguese, instead of adopting "general NLP". Moreover, there seems to be no doubt that English is by far the most spoken and understood language in the world, so, there is no surprise that, as opposing to English, for Portuguese, and all other languages, the amount of NLP resources is relatively small.

In the field of developing and providing access to resources for Portuguese NLP, Linguateca[1] plays an important role. Linguateca is a distributed network for fostering the computational processing of the Portuguese language (see Santos (2000), Veiga and Santos (2001), Santos (2002), Santos et al. (2004) and Santos (2009) for different snapshots of this project). Besides, Linguateca has been instrumental in fostering evaluation of Portuguese systems and tools, by organising several evaluation contests for Portuguese and helping disseminate evaluation in the Portuguese-speaking community, namely Morfolimpíadas, HAREM and CLEF (see Santos (2007a), Santos and Cardoso (2007), Peters et al. (2008), and Mota and Santos (2008) for the evaluation effort).

In Brazil, as the biggest country with Portuguese as first language, there are also groups working on Portuguese NLP, for instance NILC[2]. The Núcleo Interinstitucional de Lingüística Computacional (NILC) fosters research and development projects in Computational Linguistics and NLP. It includes computer cientists, linguists and other researchers from Universidade de São Paulo (USP) in São Carlos, Universidade Federal de São Carlos (UFSCar) and Universidade Estadual Paulista (UNESP) of Araraquara, all in Brazil.

Anyway, some NLP resources exist for Portuguese. The following are a small selection:

- Corpora, such as CETEMPúblico (Rocha and Santos (2000); Santos and Rocha (2001)) which is publicly available and contains about 180 million words in a compilation of newspaper text of the Portuguese daily newspaper Público;

[1] http://www.linguateca.pt
[2] http://www.nilc.icmc.usp.br/nilc/index.html

Corpus do Português[3], which contains more than 45 million words in Portuguese texts from the 1300s to the 1900s;

- Morphological analysers, such as Jspell (Simões and Almeida (2002)), which is a generic morphological analyser that, given a word, returns its morphological characteristics (e.g. grammatical category, gender, number, ...);
- Broad-coverage parsers, such as PALAVRAS (Bick (2000)), which is a lexicon and rule-based morpho-syntactic dependency parser;
- Treebanks, such as Floresta Sintá(c)tica (Freitas et al. (2008)), which is a publicly available corpus, syntactically-annotated by PALAVRAS, including a manually revised subset, called Bosque;
- Electronic Thesaurus, such as Tep (Maziero et al. (2008)), which is a publicly available lexical database, comprising 19,888 synonym sets and also antonymy links with Brazilian Portuguese word forms;
- Q&A systems, such as Esfinge (Costa and Cabral (2008)), which is a general domain question answering system which uses the information on the Web as an additional resource when searching for the answers.

2.2 Lexical Ontologies

In this section, the topics discussed will lead to the notion of one of the final goals of this thesis – lexical ontologies – which includes a review on two supporting concepts, namely lexical semantics and ontologies. Therefore, in Section 2.2.1, the topic of lexical semantics is presented with special focus on lexico-semantic relations, while in Section 2.2.2, several common ways of structuring semantic information are introduced. Section 2.2.3 is about ontologies, and includes a discussion on the controversy around its definition, a classification of ontologies according different dimensions and also a brief overview on the ontology construction process. Finally, in Section 2.2.4, some state of the art lexical ontologies are presented.

2.2.1 Lexical Semantics

According to the principle of compositionality (Frege (1960)), words, individually, are simple symbolic fragments that do not refer to the world and can hence be said to really have no meanings. In this view, words are just pieces used to construct a meaning representation, contributing in this manner to the meaning of the sentences in which they occur. This notion is, nevertheless, quite narrow and sees the vocabulary of a language, also known as the lexicon, as a simple unstructured set of words.

A different view on this subject is provided by the theory of lexical semantics (Cruse (1986)), which is the subfield of semantics that studies the words of a language and their meanings. The lexicon is hence seen as a finite list of lexical items (usually words or expressions) with a highly systematic structure that controls what words can mean. It can be seen as the bridge between a language and the knowledge

[3]http://www.corpusdoportugues.org/

expressed in that language (Sowa (1999)). Concerning the identification of a word's meaning in some context, dictionaries are very helpful tools, since they contain a collection of words and the description of their meanings.

The conceptual model of a language is structured around lexical items, their meaning (often referred as sense) and lexico-semantic relations held between the latter. To deal with the meaning of a language it is important to study these relations. Here, the most representative and studied relations are introduced.

Synonymy

The synonymy relation holds among different lexical items that have the same meaning, for instance:

car synonym_of *automobile*

A more practical definition will state 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.

Homonymy

Homonymy occurs when lexical items have the same form but different meanings, for instance:

bank: financial institution.
bank: sloping land.

When the meanings of two homonyms are somehow related, it is usually considered that we are in presence of a single lexical item with different meanings and the relation between these meanings is called polysemy (Pustejovsky and Boguraev (1996)). This is what happens between the word referring to a person native of some country and a word referring to the language spoken in that country. For instance:

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 taken into consideration to define how related the items are.

The process of identifying which overall meaning, also known as sense, of a word is being used in a sentence is called word sense disambiguation (WSD) (Nancy Ide (1998)). WSD is however very dependent on the purpose (Wilks (2000)) because sense division is not straightforward regarding there is no consensus, and probably there will never be, around this topic – even dictionaries cannot be seen as the ultimate truth, as different lexicographers, or system developers, divide senses differently (Kilgarriff (1996)).

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 true for all animals

Table 2.1: Replacement of hyponyms and hypernyms.

Hyponymy and hypernymy

When a lexical item is a subclass or specific kind of another we are in the presence of a hyponymy relation, also known as the *is-a* relation, for instance:

dog hyponym_of *mammal*.

On the other hand, hypernymy is the inverse relation of hyponymy:

mammal hypernym_of *dog*.

These relations are used to build up taxonomies, introduced in Section 2.2.2. In other words, an hyponym is a specification of its hypernym an inherits all its properties. A true meaningful sentence should remain true if we replace some concept by its hyponym, but it might not remain true if the concept is changed by its hypernym (see the example in Table 2.1).

Hypernymy can also occur between verbs, for instance:

move hypernym_of *walk*.

However, in this case, its inverse relation is called troponymy, so:

walk troponym_of *move*.

Meronymy and holonymy

When a lexical item is part, piece or member of another, a meronymy (or *part-of*) relation holds between them. For instance:

wheel meronym_of *car*

If we go in the opposite direction, a holonym is the whole that owns or has the part:

car holonym_of *wheel*

Besides hyponymy, it is also possible to build up taxonomies out of meronyms.

Other relations

Besides the relations already presented, which are the most referred in the literature, it is possible to define many more lexico-semantic relations between lexical items. Here we present other relations we are also interested in studying. All of them are non-taxonomic relations:

- **Causation:** one lexical is caused by another, for instance:

  virus causation_of *flu*

- **Purpose:** one lexical item is the purpose of another, for instance:

  find purpose_of *search*

- **Manner:** one lexical item is performed in another's manner, for instance:

  quickly manner_of *walk*

- **Localisation:** one lexical item is located in another, for instance:

  CISUC located_in *Coimbra*

The Generative Lexicon

The theory of the Generative Lexicon (Pustejovsky (1991)) is commonly referred to as an important contribution to account 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 such as synonymy, antonymy, metonymy[4] and others.

It is argued that lexical meaning can best be 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 a word or phrase;
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 in the lexicon.

[4]Metonymy is a figure of speech in which a thing is not called by its own name, but by the name of something intimately associated with that thing. 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 President of the USA, who lives in the *White House*.

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

- **Constitutive**: the parts or components of an object;
- **Agentative**: action which typically brings the object into existence;
- **Formal**: distinguishing information about the object, its hypernyms;
- **Telic**: the purpose or function of an object.

As a result, the Generative Lexicon only needs to store a single entry for every polysemous word and is able to generate the appropriate sense when placed in some context.

2.2.2 Lexical Resources and Representations

Lexical information can be organised and structured in different ways, giving rise to a lexical resource that is basically a representation of the lexicon. When in a machine-readable format, these resources can be useful for NLP applications. Here some of the most typical lexical resources (or representations) are introduced, namely dictionaries, thesauri, taxonomies, ontologies and lexical knowledge bases. The last kind of resource presented, (domain and lexical) ontologies, will be described in more detail in the further sections.

Dictionary

A dictionary is an organised repository of words and the description of their possible meanings, as definitions. Typically, a dictionary entry contains other useful pieces of information about the words, such as etymologies, pronunciation, morphological category, syllabic division, domain and examples of usage. In Figure 2.5 there is the example of an entry of the Longman Dictionary of Contemporary English (LDOCE)[4].

General language dictionaries contain a representative collection of the vocabulary of a language. They are compiled, written and edited by lexicographers, who are experts in analysing and describing the semantic, syntagmatic and paradigmatic relationships within the lexicon of a language. Bilingual dictionaries have a similar format to language dictionaries but their main purpose is to translate words or phrases from one language to another. Electronic versions of dictionaries, Machine Readable Dictionaries (MRDs), are introduced in Section 3.1 of this thesis.

Thesaurus

Similarly to a dictionary, a thesaurus is a repository of words where, in some cases, their definitions are also provided. The difference is that, in a thesaurus, words are associated to their synonyms (or close synonyms), and sometimes to their antonyms. For each word, a thesaurus contains an entry for all its possible meanings and, each

[4]The current version of the LDOCE is available for online search in `http://www.ldoceonline.com/`

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 2.5: Entry for the word *dictionary* in the LDOCE[4].

entry consists of a group of words that, in some context, have the same meaning and are thus synonyms. Roget's Thesaurus (Roget (1852)) is the first ever and a widely-used English thesaurus. Figure 2.6 contains some of the entries for the word *thesaurus* in the online service *Thesaurus.com*[5].

Concerning Portuguese, more precisely Brazilian Portuguese, Tep[6] (Maziero et al. (2008)) is an existing electronic thesaurus. An example of the entries for the word *canto* is shown in Figure 2.7.

Taxonomy

A taxonomy is basically a classification of a certain group of entities, such as plants, academical degrees, musical genres, and so on. It can be seen as a hierarchical tree where the top nodes are the more general and the lowest the more specific. Concerning the lexicon, taxonomies are often used to represent hierarchical relations, such as hypernymy, where each node in the hierarchy inherits all the properties of its *father*-node. See Figure 2.8 for an example of a taxonomy of animals.

Smith (2004) sets forth a list of principles for taxonomy well-formedness:

1. A taxonomy should take the form of a tree in the mathematical sense. This assures that the tree corresponds to a connected graph without cycles where the nodes represent categories at greater and lesser levels of generality, and branches connecting nodes represent the relations of inclusion of a lower category in a higher one.

2. A taxonomy should have a basis in minimal nodes, representing lowest categories in which no sub-categories are included. Basis, in the mathematical sense, assures that leaf nodes exhaust the maximal category in every way possible. Smith (2004) gives as an example a chemical classification of the noble gases, that is exhausted by the nodes *Helium*, *Neon*, *Argon*, *Krypton*, *Xenon* and *Radon*. This principle ensures also that every intermediate node in the tree is identifiable as a combination of minimal nodes.

3. A taxonomy should be unified in the sense that it should have a single top-most or maximal node, representing the maximum category. In other words, there should exist a (maximal) category that includes all the categories represented

[5]Available from the URL `http://thesaurus.reference.com`
[6]Available from the URL `http://www.nilc.icmc.usp.br/tep2/index.htm`

- **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, wordbook, 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 2.6: Some of the entries for the word *thesaurus* in Thesaurus.com (Roget's Thesaurus).

by the nodes lower down the tree. Otherwise, it would not be one taxonomy at all, but rather two separate and perhaps competing taxonomies.

Ontologies and lexical knowledge bases

In the computer science domain, ontologies (Gruber (1993); Guarino (1998)), further discussed in Section 2.2.3, have several proposed definitions, but can be said to be representations of explicit and formal knowledge with its meaning encoded to allow the exchange of information. Both taxonomies and thesauri are related with

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 2.7: Entries for the word *canto* in Tep.

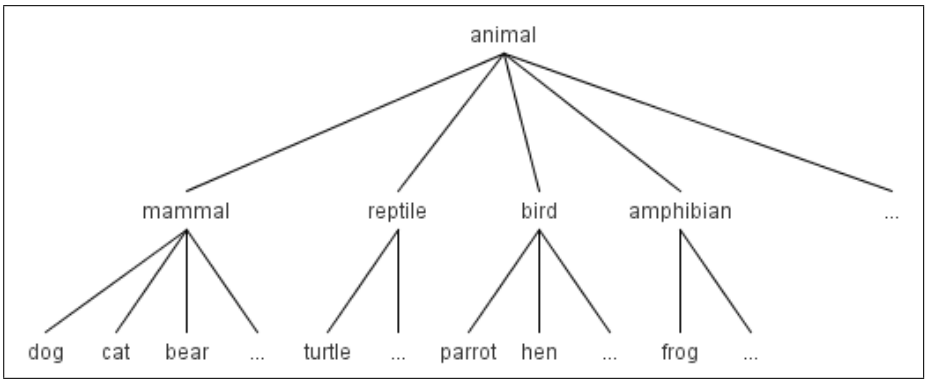

Figure 2.8: Taxonomy of animals.

ontologies, since they can be said to be ontologies where there is only one type of relation between the entities. While in the former the entities are connected through a hierarchical relation, in the latter similar entities are associated.

Lexical databases or lexical knowledge bases are structures where lexical items, their meanings and lexico-semantic relations between them are organised according to a specific theory of lexical semantics. These resources can be viewed as ontologies, more precisely lexical ontologies, which try to cover the whole lexicon of a language, or several languages in the case of multilingual lexical databases. Some existing lexical resources are presented later, in Section 2.2.4.

2.2.3 Ontologies

Knowledge engineering and representation, knowledge management and organisation, language engineering, information modeling and integration, information retrieval and extraction, and database design are just some of the fields of computer science where the importance of the use of ontologies has been recognised (Guarino (1998)). Using ontologies in applications of very distinct areas, such as natural language translation, medicine, electronic commerce or geographical information systems has also been reported, which made the construction of an ontology highly interdisciplinary process (Guarino (1998)). This led to working groups with people from different areas (e.g. computer scientists, philosophers, linguists...) and

made it important to come up with a terminology consensus or at least mutual understanding on the different terminologies.

Despite having its origins in philosophy, the term "ontology" has widespread into the computer science community (at least) in the last twenty years. When adapted to computer science, the notion of ontology was reinterpreted, leading to some disagreement concerning its definition.

Nevertheless, there is no doubt that ontologies are efficient tools to represent and share knowledge. They are basic components of the Semantic Web (Berners-Lee et al. (2001)) and, ten years ago, were almost seen as a synonym for the solution to many problems concerning the fact that computers do not understand human language – *if there were an ontology and every document were marked up with it and we had agents that would understand the markup, then computers would finally be able to process our queries in a really sophisticated way* (Biemann (2005)). This vision has however not come true (at least not yet).

Here, following an approach on the controversy around the definition of ontology, several classifications of ontologies are introduced. Then, a methodology and some issues for the construction of an ontology are presented and finally some criteria that should be considered for automatic acquisition of ontologies from text are referred.

Definition

The origin of the term "ontology" is reported to Aristoteles and has a Greek etymology *ón*, *óntos* – being, existing, essence – and *logos* – science, study, theory. In the philosophy domain, ontology represents a branch dedicated to the study and description of existence and reality (Zúñiga (2001)).

On the other hand, and considering the computer science domain, a commonly cited definition is given by Gruber (1993):

> *"An ontology is an explicit specification of a conceptualization."*

Based on the notion of conceptualisation by Genesereth and Nilsson (1987), "the objects, concepts, and other entities that are presumed to exist in some area of interest and the relationships that hold them", Gruber says that a conceptualisation is an abstract and simplified view of the world that we wish to represent for some purpose.

After analysing several other definitions of ontologies in the artificial intelligence domain, van Heijst et al. (1997) suggest the following definition:

> *"An ontology is an explicit knowledge-level specification of a conceptualization, i.e. the set of distinctions that are meaningful to an agent. The conceptualization and therefore the ontology may be affected by the particular domain and the particular task it is intended for."*

Still, in these definitions, the relation between a conceptualisation and an ontology is not completely clear. The problem seems to be in the definition of conceptualisation where it is not evident if both concepts (i.e. models) and objects (i.e. instances) are at the same level. Having this in mind, Guarino (1998) defines an ontology, in the artificial intelligence domain, as:

> *"An ontology is an engineering artifact, constituted by a specific vocabulary used to describe a certain reality, plus a set of explicit assumptions regarding the intended meaning of the vocabulary words."*

He believes that, in the computer science domain, a conceptualisation is basically a philosophical ontology and refines its notion as "a set of conceptual relations defined on a domain space". In order to refine Gruber's definition and clarify the difference between an ontology and a conceptualisation, he gives another definition of ontology:

> *"An ontology is a logical theory accounting for the intended meaning of a formal vocabulary, i.e. its ontological commitment to a particular conceptualization of the world. The intended models of a logical language using such a vocabulary are constrained by its ontological commitment. An ontology indirectly reflects this commitment (and the underlying conceptualization) by approximating these intended models."*

Guarino (1998) also emphasises that, in the computer science domain, an ontology is dependent on a vocabulary, in opposition to a conceptualisation, that is a vision of world that is independent of a language. This makes it possible for two different ontologies to share the same conceptualisation.

Both definitions of ontology were discussed by Zúñiga (2001), who contributes with a unified definition that aims to be understood interdisciplinarily, and ease the work among heterogeneous teams:

> *"An ontology is an axiomatic theory made explicit by means of a specific formal language. The IS ontology*[7] *is designed for at least one specific and practical application. Consequently, it depicts the structure of a specific domain of objects, and it accounts for the intended meaning of a formal vocabulary or protocols that are employed by the agents of the domain under investigation."*

She also gives a new interdisciplinary definition for conceptualisation: *"A conceptualization is the universe of discourse at work in every possible state of affairs for the particular domain (or domain space) of objects that is targetted by the IS ontology"*.

Classification

Different systems need different ontologies, capable of representing different kinds of knowledge and thus different views of the world.

Ontologies can hence be classified according to, at least two, different dimensions: the amount and type of structure of the conceptualisation, and the subject of the conceptualisation (van Heijst et al. (1997)). Concerning the first dimension, van Heijst et al. (1997) distinguish three categories:

- *Terminological Ontologies*: those that specify the terms that are used to represent knowledge in the domain of discourse (e.g. domain-based lexicons).

[7] Zúñiga refers to ontologies in the computer science/information systems as IS ontologies.

Axioms:
food(brie), food(camembert), food(turkey), food(meatballs), food(chili con carne), meat(turkey), meat(minced meat), part_of(minced meat, chili con carne), part_of(minced meat, meatballs)

$veg\_food(x) = x|food(x) \wedge (\neg part\_of(y,x) \wedge meat(y)) \vee \neg meat(x)$
$non\_veg\_food(x) = x|food(x) \wedge (part\_of(y,x) \wedge meat(y)) \wedge meat(x)$

Figure 2.9: Formal ontology, where it is possible to derive, for instance, that turkey and chili con carne are non-vegetarian foods. (adapted from Biemann (2005))

- *Information Ontologies*: those that specify the record structure of databases (e.g. database schemata).
- *Knowledge Modeling Ontologies*: usually with a richer structure, these ontologies specify conceptualisations that are optimised for a particular use of the knowledge that they describe.

Also according to the type of structure of its conceptualisation, Sowa (1999) classifies ontologies into three kinds:

- *Axiomatised or Formal Ontologies*: those that were the subject of Gruber (1993)'s and Guarino (1998)'s definitions, where categories are distinguished by axioms and definitions. The common-sense knowledge base Cyc (Lenat (1995)) suits this definition. In Figure 2.9 a piece of a formal ontology is shown.
- *Prototype-based Ontologies*: those where categories are distinguished by typical instances or prototypes, rather than by axioms and definitions in logic. Thesaurus belong to this category of ontologies, because they contain sets of related terms that resemble the prototype of a category.
- *Terminological Ontologies*: those where the categories do not need to be fully specified by axioms and definitions. WordNet (Fellbaum (1998)) can be viewed as a *terminological ontology* where categories are partly specified by taxonomic or other lexico-semantic relations and concepts are described by labels or synonyms rather that prototypical instances. Taxonomies can also be considered to be terminological ontologies.

As for the second dimension, the subject of the conceptualisation, Biemann (2005) defines two main levels of ontologies:

- *Upper ontologies*: those which describe the most general entities, contain very generic specifications and serve as a foundation for specialisations. These ontologies, typically contain entries like *space*, *object*, *event*, *action* or *time*. As a source of lexical knowledge, where the general meaning of a language is encoded, lexical ontologies like WordNet (Fellbaum (1998)) and others introduced in Section 2.2.4 suit this description of top-level ontologies.
- *Domain ontologies*: those which describe subject domains or, in other words, have a particular perception of the world (e.g. the world of medicine, the world

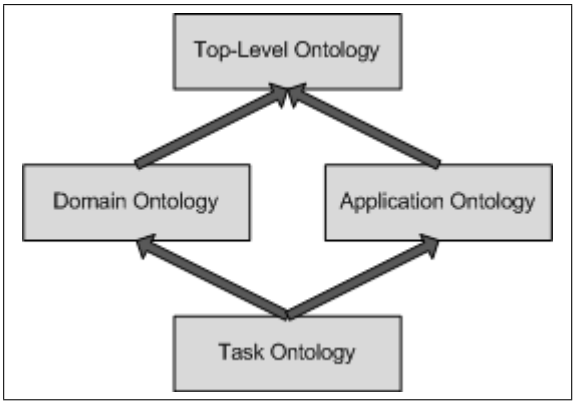

Figure 2.10: Specialisation relationships between different kinds of ontologies, according to their level of dependence on a particular task or point of view.

of automobiles...). If we think of the "one sense per domain" assumption, the more specialised the domain is, the less is the influence of WSD.

Guarino (1998) goes further on the level of specialisation of an ontology and gives the following classification, according to its level of generality (see the diagram in Figure 2.10):

- *Top-level ontologies*: basically the same as Biemann (2005)'s *upper ontologies*.
- *Domain ontologies*: defined in the same way as Biemann (2005)'s *domain ontologies*.
- *Task ontologies*: which are ontologies that have the same level of specialisation as the *domain ontologies*, but describe task or activity-specific vocabulary (e.g. diagnosing, selling ...).
- *Application ontologies*: those which describe concepts that correspond to roles played by domain entities while performing a certain task and are thus dependent on both a domain ontology and a task ontology.

At least theoretically, more general ontologies have a unified content which is therefore more consensual and leads to larger communities of users. On the other hand, more specific ontologies have more specialised content and are sometimes developed to represent particular kinds of knowledge with less intentions of being reused. This model leads to an easier integration of different information systems that agree, at least, with the top-level ontology.

This thesis is more concerned with top-level ontologies, more precisely lexical ontologies. Nevertheless, as will be referred in Section 3.2, domain-specific knowledge, extracted from domain-specific texts, will also be dealt in order to be used for the enrichment of the main ontology.

Construction

Regarding the achievement of representing knowledge that is meant to be shared, but is basically a view on the world, the construction of an ontology is far from being a simple task. There is no right way or methodology for the construction of ontologies

and planing this process depends on the purpose of the ontology itself. Among various alternatives for ontology development, a general methodology, commonly followed, is proposed by Noy and McGuinness (2000). It is an iterative approach consisting of the following steps:

1. *Determine the domain and scope of the ontology*: definition of the domain that will be covered by the ontology, and of the purpose and future applications of the ontology. All these elements may change during the design process, but fixing the scopre of the model can benefit from their definition.

2. *Consider reusing existing ontologies*: ontologies are meant to be shared so it might be possible to reuse existing ontologies and avoid the repetition of work that has already been done. Additionally, systems using the same ontology can easily interact with each other.

3. *Enumerate important terms in the ontology*: writing down some terms that are closely related to the target domain. It is useful to do it in an earlier stage, concerning the definition of the kind of instances needed and also their properties, without thinking about relations or term overlapping.

4. *Define the classes and the class hierarchy*: this step can follow one of three approaches, namely top-down, bottom-up or a combination of both. On the one hand, a top-down approach starts with the definition of most general concepts in the domain, followed by their specialisation. On the other hand, a bottom-up approach starts by defining the most specific classes, followed by their association into more general classes.

5. *Define the properties of classes-slots*: definition of the internal structure of each class, more precisely, their specific properties (referred to as slots attached to classes).

6. *Define the facets of the slots*: definition of the restrictions of each slots. For instance, the value type, allowed values and cardinality of the values.

7. *Create instances*: production of individual instances of the classes in the hierarchy. A class is first chosen, then the instance is created and finally its respective slot values are filled.

Despite the existing methodologies, the development of reusable ontologies is not always achieved due to two major difficulties: hugeness and interaction (van Heijst et al. (1997)). There is an overwhelming amount of knowledge in the world and often domain knowledge cannot be represented independently from particular assumptions of how it will be used in reasoning. Following these difficulties, van Heijst et al. (1997) refer four issues in ontology construction:

1. *Language*: the means to specify the ontology

2. *Modularity*: the cohesion within modules should be maximal and the interaction between them minimal

3. *Alternative definitions*: definitions should be viewed as conceptualisations that have been proven useful for solving problems. Thus, it is sometimes important to allow for alternative, sometimes inconsistent, definitions of a concept.

4. *The need for a higher-order language*: where higher-order expressions are allowed. In order to hold the principle of modularity, the more generic aspects of a concept must be defined in a core theory, while the more domain-specific aspects of those concepts must be in a more peripheral theory.

Another problem in ontology construction and sharing is the common assumption that, given the existence of an ontology, people will be willing to tag their own work (Brewster and Wilks (2004)). However, experience tells us that authors tag their work inadequately or inappropriately.

Most ontologies are handcrafted (Brewster and Wilks (2004)) which leads to considerable problems. Considering the aforementioned steps in ontology development, there is much human effort involved in manual construction and maintenance. Besides, the knowledge that intends to be captured is usually changing and developing continuously and it is always a personal view on the world that is hardly consensual, given the difficulties of agreement on world categorisation.

Learning

Bearing the remarks given in the above section, automatic acquisition of knowledge from electronic sources should be viewed as viable alternative. Brewster and Wilks (2004) suggest a set of criteria for ontology learning from text, willing to guide both the choice of taxonomy (here expanded to ontology) construction and evaluation methods:

1. *Coherence*: in the user's point of view, the ontology should have a coherent and common sense organisation of concepts or terms. Coherence is dependant on the terms and on the associations between them, that should be part of the shared conceptualisation referred in the previous section. Encompassing coherence is however different for different applications. For example, in a linguistically coherent thesaurus we expect to find groupings of similar words, while in a coherent taxonomy concepts or terms are organised into categories and subcategories that are established according to specific properties. It is thus very difficult to evaluate an ontology from the coherence angle.

2. *Multiple Inheritance*: as pointed out by Noy and McGuinness (2000), a term can occur multiple times and in different positions of the same taxonomy. Brewster and Wilks (2004) are in agreement and state that methods for ontology learning from text should take into account the different senses of a term and position them adequately in the ontology.

3. *Ease of Computation*: having in mind the general problem of maintenance of a knowledge base, it is important that the construction method does not have a high computational complexity, providing the output as soon as possible whether it is for updating, evaluation or deployment purposes.

4. *Single labels*: all nodes in an ontology need to have single labels, despite being or not composed by more than one word. Groups of words identified by only one word are easily understood by the users. Although synonyms can be defined as a set of terms, their labels should have only one of the terms or, alternatively, all terms can act independently as possible labels for the same concept.

5. *Data Source*: an ontology should be constructed out of data coming from both documents (primary sources), and seed ontologies, that may work as an initial base structure and should thus represent more consensual knowledge.

2.2.4 State of the Art Lexical Ontologies

In the last two decades, there have been many efforts to create a large database where words and their meanings were represented along with connections held between them, in order to structure lexical knowledge. Lexical databases, lexical knowledge bases or lexical ontologies are some of the names given to the resources resulting from these efforts. Some of them are presented in this section.

In general, the construction of a lexical ontology is aided by information in dictionaries, thesauri or other textual resources like corpora and can be achieved either by handcrafting or by automatically acquiring information from text. The structures of these resources usually follows one of the three formalisms to represent the meaning of a language, introduced in Section 2.1.4: logical predicates (Smullyan (1995)), directed graphs and semantic frames (Fillmore (1982)).

NLP capabilities of a language rely heavily on the existence of these resources that, among the large set of NLP tasks, are useful for:

- Inferring similarity (Richardson (1997), Seco et al. (2004));
- WSD (Gomes et al. (2003), Cañas et al. (2003));
- Q&A (Clark et al. (2008), Kaisser (2005));
- Cross-lingual text retrieval (Gonzalo et al. (1998));
- Intelligent search (Moldovan and Mihalcea (2000), Liu et al. (2004));
- Machine translation (Chatterjee et al. (2005));
- Creative text generation (Hervás et al. (2006)).

More applications of WordNet are reported by Morato et al. (2004) and ideas for applying Cyc are referred by Lenat and Guha (1991).

Princeton WordNet

Princeton WordNet (Fellbaum (1998)) is a resource that combines traditional lexicographic information with modern computation, in a lexical resource based on psycolinguistic principles. It is freely available, widely used in computational linguistics and NLP, and probably the most important reference when it comes to lexical ontologies in English. It was however manually created.

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 2.11: Entry for the word *bird* in Princeton WordNet 3.0, after extending the direct hyponyms of one of its senses.

In the WordNet's lexicon, the words are clearly divided into nouns, verbs, adjectives, adverbs and functional words. The basic structure in WordNet is the *synset*, which is a set of synonym words that can be used to represent one concept. The synsets are organised in a network of semantic relations, such as hyponymy and meronymy (between nouns), and troponymy and entailment (between verbs). For illustrative purposes, Figure 2.11 contains the Princeton WordNet entry for the word *bird* and some of its direct hyponyms. As one can see, WordNet has five different noun senses and one verb sense for *bird*.

Multilingual wordnets

EuroWordNet (Vossen (1997)) and MultiWordNet (Pianta et al. (2002)) are two multilingual databases, created by two distinct models. One of the main purposes of these kind of resources is multilingual information retrieval.

The EuroWordNet's database consists of wordnets for English, Spanish, Dutch

and Italian, each one of them structured in the same lines as Princeton Wordnet. This means that synonyms are grouped into synsets, which are related by means of semantic relations. Additionally, each meaning is linked to a Princeton Wordnet synset by means of an equivalence relation. The monolingual wordnets are interconnected via an unstructured list of synsets, called the Inter-Lingual-Index (ILI).

In MultiWordnet, wordnets of different languages are strictly aligned with Princeton Wordnet. The first wordnet to be aligned was the Italian, but several other languages (namely Spanish, Portuguese, Romanian, Latin and Hebrew) have also joined the project, which led to the inclusion of their aligned wordnets in the MultiWordNet database.

The main difference between EuroWordNet and MultiWordNet is that, while in the construction of the former a team tries to find correspondences between the existing wordnets for different languages, in the latter, language specific wordnets are built keeping as much as possible the semantic relations available in Princeton WordNet.

Cyc

Cyc (Lenat and Guha (1989)) was not created in order to become a linguistic resource, but it is frequently cited and used in the NLP community. Opposing to WordNet, Cyc is highly formalised and all the knowledge is described with a language based on first order predicate logic. Its authors claim Cyc is the world's largest and most complete general knowledge base and common-sense reasoning engine.

Cyc was constructed manually, mainly because of a pessimistic view of its authors, who did not believe on the NLP and AI capacities to build such a resource by automatic means. Concepts aiming at describing human reality and knowledge about them, more specifically common-sense axioms on them, were handcrafted and included in Cyc's knowledge base. Common-sense assertions embody fundamental knowledge that is assumed to be already known about the world and is unlikely to be published in books, dictionaries or encyclopedias. Some examples of this kind of knowledge are provided in Lenat (1995):

1. You have to be awake to eat;

2. You can usually see people's noses, but not their hearts;

3. You cannot remember events that have not happened yet.

All knowledge in Cyc is written using CycL, a specific language modelled after first-order predicate calculus, but far more expressive and complex (Cycorp (2002)). With this logical representation, it is possible to define concepts precisely and avoid ambiguity. For instance, in an example given by Cycorp (2002), three assertions are given to the ambiguous word *running*:

- x is running-InMotion → x is changing location

- x is running-DeviceOperating → x is operating

- x is running-AsCandidate → x is a candidate

Bird
"An instance of BiologicalClass, and a specialization of Vertebrate. Each instance of Bird is an air-breathing, warm-blooded, winged animal covered with feathers. Members of most, but not all, species of bird can fly."

- **isa:** KEClarifyingCollectionType, OrganismClassificationType, UniversalVocabularyMt, BiologicalClass
- `genls:` OviparousAnimal, Homeotherm, TerrestrialOrganism, AirBreathingVertebrate, Vertebrate, NonPersonAnimal, Biped

Figure 2.12: The entry for *bird* in OpenCyc's knowledge base.

This makes it possible to place the appropriate rules on their respective concepts. Furthermore, logic has the advantage of offering a calculus of meaning and reasoning capabilities.

The most important terms in CycL are:

- Constants, which denote:
  - Individuals, which can themselves be:
    - Partially tangible individuals, such as `#$BillClinton` or `#$DisneyLand-TouristAttraction`;
    - Relations, such as `#$likesAsFriend`, `#$objectHasColor` or `#$and`;
    - Attribute values, such as `#$RedColor` or `#$Soil-Sandy`.
  - Collections (e.g. `#$Dog`, `#$SnowSkiing`, `#$PhysicalAttribute`)
- Formulas, which are relations applied to some arguments. They can have two types:
  - Sentences, when the relation is a truth function, such as `(#$isa #$GeorgeWBush #$Person)` or `(#$likesAsFriend #$GeorgeWBush #$AlGore)`
  - Non-atomic terms, when the relation is a functional-denotational, such as `(#$GovernmentFn #$France)` or `(#$BorderBetweenFn #$France #$Switzerland)`

In Cyc, each assertion should be considered true only in a certain context, recognisable by the assumptions it makes. For example, in a context of total darkness, there might be an assertion telling it is impossible to see anything and thus contradicting example 2.

The assertions of Cyc are organised in a knowledge base with entries such as the one in Figure 2.12. The two relations shown, `#$isa` and `#$genls`, are the two Cyc taxonomic relations. The difference is that while `#$genls` is transitive, `#$isa` is not. In other words, `#$genls` provides inheritance to all the terms below in the hierachy, while `#$isa` only provides it to the term in the assertion.

OpenCyc[8] is the open source version of Cyc. Its current version (2.0) includes the whole Cyc ontology and virtually all of Cyc's hundreds of thousands of terms, along with millions of assertions relating the terms to each other, forming an upper

[8] http://www.opencyc.org/

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

ontology whose domain is all of human consensus reality. Additionally, links between Cyc concepts and WordNet synsets are available.

Berkeley FrameNet

Berkeley FrameNet (Baker et al. (1998)) is another kind of lexical resource, which constitutes a network of semantic frames (Fillmore (1982)), manually extracted from a systematic analysis of semantic patterns in corpora. Each frame corresponds to a concept and describes an object, a state or an event by means of syntactic and semantic relations of the lexical item that represents that concept.

A frame can be conceived as the description of a situation with properties, participants and/or conceptual roles. A typical example of a semantic frame is *transportation* (see Figure 2.13), 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.*

Besides Inheritance, which is basically the hypernymy relation, frames can be connected with other frames by means of other semantic relations, namely Subframe, Inchoative_of, Causative_of, Precedes, Using and See_also.

MindNet

MindNet (Richardson et al. (1998); Vanderwende et al. (2005)) is a lexical knowledge base created by the Microsoft NLP research group. The resource was created by automatic tools, such as the broad-coverage parser MEG, used in the grammatical verification of Microsoft Word. This parser generates syntactical trees in which logical rules for the extraction of relations between words are applied.

MindNet in not a static resource. It represents a methodology consisting of a set of tools to acquire, structure, access and explore semantic information contained in texts. So, the semantic network was extracted not only from MRDs, such as the LDOCE and the American Heritage Dictionary 3rd Edition (AHD3), but also from encyclopedias, and other kinds of text.

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

MindNet contains a long set of semantic (and syntactic) relations, namely 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 similarity, is the identification of "relation paths" between words. For example, if one looks for paths between *car* and *wheel* a long list of relations will be returned. The returned paths include not only simple relations like *car is a modifier of wheel* but also more complex ones like *car is a hypernym of vehicle and wheel is a part of vehicle*.

Each path is automatically weighted according to its salience. The procedure for determining similarity of two words starts by querying MindNet for the ten top-weighted paths between those words. Then, the configuration of the obtained paths is matched with the configuration of the most frequent paths for similar words. A thesaurus was used in order to obtain the most frequent paths between synonym and hypernym words. After the matching, the similarity potential is the average result of the matching function for the ten paths. In Figure 2.14 the ten top-weighted paths from *bird* to *parrot* are represented.

ConceptNet

Another common-sense knowledge base, similar to Cyc, but generated automatically.

Portuguese lexical ontologies

WordNet.PT (Marrafa (2002); Marrafa et al. (2006)) is an attempt of creating a Portuguese lexical resource from scratch, within the EuroWordNet's framework, which started in 1999. The authors of WordNet.PT explicitly claim that the available resources for Portuguese NLP are not suitable for the automatic construction of such

a resource. Its database includes, among several others, the classical relations of hypernymy or meronymy but, in the later years, its authors have been explicitly interested in cross-categorical relations such as those linking adjectives to nouns (Marrafa et al. (2006)).

WordNet.BR (Dias da Silva et al. (2002); Dias-da-Silva (2006)) is the Brazilian Portuguese version of the "wordnet concept", which started in 2002. Its database is structured around synonymy and antonymy, manually extracted from a reference corpus where several dictionaries are included. Plans for adding more relations in the future have been reported in Dias-da-Silva (2006).

Portuguese is also one of the languages in the european initiative MultiWordNet (Pianta et al. (2002)), presented earlier in this section. This resource, which spans over 17,200 synsets, made of over 16,000 lemmas, is proprietary but can be bought either for research or commercial purposes. MultiWordNet.PT includes the subontologies under the concepts of Person, Organization, Event, Location, and Art works, which are covered by the top ontology made of the Portuguese equivalents to all concepts in the top four layers of the Princeton WordNet and to the 98 Base Concepts suggested by the Global Wordnet Association[9], and also the 164 Core Base Concepts indicated by the EuroWordNet project.

Additionaly, for Portuguese, there is also an electronic thesaurus developed under the principles of Princeton WordNet, Tep, referred in Section 2.2.2. As a thesaurus, the basic unit in TeP is the synset, where synonym terms can be found. The other relation included in TeP is antonymy, which holds between opposing or contradicting concepts. Tep also contains glosses for several synsets and example sentences, both taken from WordNet.BR. Tep's database, which contains 19,888 synsets and 44,678 lexical units, is freely available for download and for browsing in a web interface[10].

More recently, the results of the project PAPEL (Gonçalo Oliveira et al. (2008, 2009a,b)) were made public[11] by Linguateca. PAPEL can be seen as lexical ontology for Portuguese and consists of relations between terms. Among the aforementioned lexical resources for Portuguese, PAPEL is the only one which was extracted semi-automatically. This was achieved after processing the definitions of a major general dictionary, the Dicionário da Língua Portuguesa (dlp (2005)), by Porto Editora. It contains about 200,000 relations organised into main groups, that can be divided into sub-relations, according to the grammatical category of the arguments (see Table 2.2 for the complete relation set of PAPEL).

Comparative view

As we have seen, there are several available lexical databases, following different construction and representation approaches, and with different licenses for utilisation. Despite being useful for many NLP tasks, their different structures can sometimes be seen as optimised for performing different tasks. In Tables 2.3, 2.4, 2.5, 2.6, 2.7, 2.8 and 2.9 the resources referred in this section, both for English and Portuguese, are put side-by-side to ease their comparison.

Table 2.3 contains the core structures of each resource and how these structures are connected, which can be viewed as the nodes and the arcs in the network estab-

[9] http://www.globalwordnet.org/

[10] http://www.nilc.icmc.usp.br/tep2/index.htm

[11] http://www.linguateca.pt/PAPEL

Group	Name	Args.	Qnt.	Examples
Synonymy	SINONIMO_DE	*same*	80,432	(*flexível*, *moldável*)
Hypernymy	HIPERONIMO_DE	n,n	63,455	(*planta*, *salva*)
Meronymy	PARTE_DE	n,n	14,453	(*cauda*, *cometa*)
	PARTE_DE_ALGO_COM_PROP	n,adj	3,715	(*tampa*, *coberto*)
	PROP_DE_ALGO_PARTE_DE	adj,n	962	(*celular*, *célula*)
Cause	CAUSADOR_DE	n,n	1,125	(*fricção*, *assadura*)
	CAUSADOR_DE_ALGO_COM_PROP	n,adj	16	(*paixão*, *passional*)
	PROP_DE_ALGO_CAUSADOR_DE	adj,n	515	(*reactivo*, *reacção*)
	ACCAO_QUE_CAUSA	v,n	6,424	(*limpar*, *purgação*)
	CAUSADOR_DA_ACCAO	n,v	39	(*gases*, *fumigar*)
Producer	PRODUTOR_DE	n,n	932	(*romãzeira*, *romã*)
	PRODUTOR_DE_ALGO_COM_PROP	n,adj	31	(*sublimação*, *sublimado*)
	PROP_DE_ALGO_PRODUTOR_DE	adj,n	348	(*fotógeno*, *luz*)
Purpose	FINALIDADE_DE	n,n	2,095	(*defesa*, *armadura*)
	FINALIDADE_DE_ALGO_COM_PROP	n,adj	23	(*reprodução*, *reprodutor*)
	ACCAO_FINALIDADE_DE	v,n	5,640	(*fazer_rir*, *comédia*)
	ACC_FINALIDADE_DE_ALGO_COM_PROP	v,adj	255	(*corrigir*, *correccional*)
	MANEIRA_POR_MEIO_DE	adv,n	1,433	(*timidamente*, *timidez*)
Place	LOCAL_ORIGEM_DE	n,n	768	(*Japão*, *japonês*)
Property	PROP_DE_ALGO_REFERENTE_A	adj,n	3,700	(*dinâmico*, *movimento*)
	PROP_DO_QUE	adj,v	17,028	(*familiar*, *ser_conhecido*)

Table 2.2: The relations of PAPEL.

lished in the resource. Besides structural information, the general NLP tasks which the resource seems to be optimised for are also referred. Furthermore, Table 2.4 points out the approach followed in their construction (manual or automatic) and also their availability for utilisation. Despite the referred availability, all these resources (or at least representative parts) are freely available through web interfaces. One thing that should be noted is that WordNet's public availability seems to be an important contribution for its high acceptance among the scientific community. On the other hand, there are very few works using MindNet, which is owned by Microsoft.

Tables 2.5 and 2.6 are an attempt to map the relations in all the resources into several broader slots, namely synonymy, antonymy, hypernymy, meronymy, causation, purpose, place and manner. However, this mapping should not be seen as completely straightforward, but only as an approximation exercise, considering the name of the relations, some descriptions and examples of the relations, when available. As one can see, no resource covers all the slots, and, while some of them have only one relation per slot, others make fine-grained distinctions, considering the type (e.g. WordNet.PT 1.5) or the grammatical categories (e.g. PAPEL 1.0) of the arguments of each relation.

Finally, Tables 2.7, 2.8 and 2.9 compare the resources in terms of numbers. Despite these numbers, among the presented resources, OpenCyc 2.0 is the one with more terms and assertions, since it includes hundreds of thousands of terms, along with millions of assertions. However, most of these are nothing by common-sense assertions and not lexico-semantic relations. Besides this fact, Cyc and FrameNet were not considered in these tables since they have different representations, which do not suit this comparison exercise.

As one can see by looking at the tables, resources for English, namely WordNet and MindNet are much bigger (more than 10 times!) than the resources for Portuguese. The only exception is the more recent resource, PAPEL, which was created by semi-automatic tools, but, as opposing to the others, does not establish synsets. This takes us to another important point that should be noticed, which is the size of

Resource	Core structure	Connections	Optimised for...
WordNet	synsets	relations	lexical categorisation, word similarity determination
Cyc	terms	assertions	formalised logical reasoning
FrameNet	frames	relations	*various kinds of NLP processes*
MindNet	words and their dictionary senses	relations and paths	word-similarity determination
WordNet.PT	synsets	relations	*aids in various fields of Computational Linguistics and Language Engineering*
Tep	synsets	*only antonymy connections*	writing aids
MultWordNet.PT	synsets	relations	represent true lexical idiosyncrasies between languages
PAPEL	terms	relations	understand the relations among words of a general dictionary

Table 2.3: Comparative view on lexical databases (core structure, connections and optimised for).

the two resources created semi-automatically, MindNet and PAPEL, compared to the handcrafted ones. The latter which are much smaller and have probably much more intensive work involved.

It should be added that, since MindNet can be viewed as methodology used to build a lexical knowledge base, the results presented are the ones reported in Richardson et al. (1998), which were generated after processing two dictionaries, namely the LDOCE and the AHD3. Since that time, other sources of knowledge might have been processed, however no statistics about newer versions of MindNet have been found.

Another thing about the presented numbers is that, even though the information about WordNet.PT in Tables 2.5 and 2.6 relies on information found in its website, which reports the relations in its latest version, WordNet.PT 1.5, statistics about this version could not be found. So, the later tables are about a previous version, more precisely WordNet.PT 1.0, whose statistics are reported in Marrafa (2002). Nevertheless, in the later years it has been reported (Marrafa et al. (2006), Mendes (2006)) that current directions of the WordNet.PT project are concerned with the inclusion and representation of adjectives. Mendes (2006) refers that WordNet.PT, at that time, had a total of 12,630 synsets including 1,034 adjectives, which doubles the number of adjective synsets shown in Table 2.6.

Still concerning Portuguese lexical resources, Santos et al. (2009) will present a further level of comparison.

[12]OpenCyc, the open source version of Cyc

[13]In Cyc, there over 200 predicates about roles and actor slots, and many relations about the relative positions of objects, nearness and location and also approximately 60 *in* predicates.

[14]These numbers are not correct, but were the ones announced with PAPEL 1.0. They were

Resource	Construction	Availability
WordNet	manual	public domain
Cyc	manual	public domain[12]
FrameNet	manual	licenses: academic (no commercial rights), standard (limited commercialisation) or custom
MindNet	semi-automatic	proprietary
WordNet.PT	manual	proprietary
Tep	manual	public domain
MultWordNet.PT	manual	paid licenses: academic and commercial
PAPEL	semi-automatic	public domain

Table 2.4: Comparative view on lexical databases (construction and availability).

later corrected, in PAPEL 1.1, to 55,372, 24,089, 18,933 and 1,389, respectively.

Resource	Synonymy	Antonymy	Hypernymy	Meronymy
WordNet 3.0	*Words belonging to the same synset*	antonym	(instance) hypernym, hyponym for nouns; hypernym, troponym, for verbs	(part/member/substance) meronym, holonym
Cyc			#$isa, #$genls	#$parts, #$intangibleParts, #$subInformation, #$subEvents, #$physicalDecompositions, #$physicalPortions, #$physicalParts, #$externalParts, #$internalParts, #$anatomicalParts, #$constituents, #$ingredients
FrameNet			Inheritance	Subframe
MindNet	Synonym		Hypernym	Part
WordNet.PT 1.5	*Words belonging to the same synset*	antónimo, quasi-antónimo	hipernimo, hipónimo, instância, instanciado	merónimo, holónimo (parte distinta/membro/-porção/matéria/local)
Tep 2.0	*Words belonging to the same synset*	Antônimo		
MultiWordNet.PT v1	*Words belonging to the same synset*		HAS-HYPONYM, HAS-HYPERNYM	IS-PART-OF, IS-MEMBER-OF, IS-SUBSTANCE-OF, HAS-PART, HAS-SUBSTANCE, HAS-MEMBER
PAPEL 1.0	SINONIMO_DE		HIPERONIMO_DE	PARTE_DE (noun-noun), PARTE_DE_ALGO_COM_PROP (noun-adj), PROP_DE_ALGO_PARTE (adj-noun)

Table 2.5: Mapping attempt for synonymy, antonymy, hypernym and meronymy relations in lexical databases.

Resource	Causation	Purpose	Place	Manner
WordNet 3.0	cause (only between verbs)			
Cyc	#$outputsCreated, #$inputsDestroyed, ...[13]	#$performedBy, #$deviceUsed, ... [13]	#$performedBy, #$deviceUsed, ... [13]	...[13]
FrameNet	Causative_of	Using		
MindNet	Cause, Result	Purpose, Means, Goal		Manner
WordNet.PT 1.5	causa, tem como causa, resulta de, tem como resultado	agente_instrumento, instrumento_agente instrumento_resultado, resultado_instrumento, agente_resultado, resultado_agente, agente_paciente/objecto, paciente/objecto_agente, paciente/objecto_instrumento, instrumento_paciente/objecto, paciente/génese_resultado, resultado_paciente/génese	lugar onde, tem lugar em	feito (modo), modo como feito
Tep 2.0				
MultiWordNet.PT v1				
PAPEL 1.0	CAUSADOR_DE (noun-noun), CAUSADOR_DA_ACCAO (noun-verb), CAUSADOR_DE_ALGO_COM_PROP (noun-adj), PROPRIEDADE_DE_ALGO_QUE_CAUSA (adj-noun), ACCAO_QUE_CAUSA (verb-noun)	FINALIDADE_DE (noun-noun), FINALIDADE_DA_ACCAO (noun-verb), FINALIDADE_DE_ALGO_COM_PROP (noun-adj), ACCAO_FINALIDADE_DE (verb-noun), ACCAO_FINALIDADE_DE_ALGO_COM_PROP (verb-adj)	LOCAL_ORIGEM_DE (noun-noun)	MANEIRA_POR_MEIO_DE (adv-noun)

Table 2.6: Mapping attempt for causation, purpose, place and manner relations in lexical databases.

Resource	Noun	Verb	Adjective	Adverb	Total
WordNet 3.0	11,7097	11488	22141	4601	155,327
MindNet (1998)	159,000 headwords				159,000
MultiWordNet.PT v1	16,205	0	0	0	16,205
WordNet.PT 1.0	9,813	633	485	0	10,931
Tep 2.0	17,276	10,910	15,001	1,138	44,678
PAPEL 1.0[14]	50,201	17,932	14,025	43,713	125,871

Table 2.7: Lexical databases in numbers: unique word forms/category

Resource	Noun	Verb	Adjective	Adverb	Total
WordNet 3.0	81,426	13,650	18,877	3,644	117,597
MindNet (1998)	191,000 definitions				191,000
MultiWordNet.PT v1	17,285	0	0	0	17,285
WordNet.PT 1.0	8,100	424	491	0	9,015
Tep 2.0	8,526	4,145	6,647	566	19,884

Table 2.8: Lexical databases in numbers: synsets

Resource	Number of relations
WordNet 3.0	207,016
MindNet (1998)	713,000
MultiWordNet.PT v1	66,475
WordNet.PT 1.0	11,584
Tep 2.0	4,276 (antonymy)
PAPEL 1.0	200,384

Table 2.9: Lexical databases in numbers: relations

Chapter 3

Related Work

The exploration of textual resources to automatically acquire and structure lexical and semantic information started a long time ago (Calzolari et al. (1980); Amsler (1980)). Many researchers used machine readable dictionaries (MRDs) as their source of knowledge. Besides other advantages, these were, after all, the main sources of lexical knowledge. Nevertheless, it was noticed earlier that the knowledge in MRDs was too general and though not suitable to acquire domain-specific knowledge. So, some authors (e.g. Hearst (1992)) moved on to the exploration of textual corpora. Corpora processing in order to extract lexico-semantic knowledge seemed a good option, especially when this knowledge is used to enrich some generic lexical resource, whether it is a general lexical ontology (e.g. WordNet (Fellbaum (1998))) or the results obtained after extracting knowledge from a dictionary.

In this chapter, some work related to this research is presented, more precisely work on the (semi) automatic extraction of lexico-semantic knowledge from MRDs (Section 3.1) and corpora (Section 3.2), and also work on the evaluation of ontologies (Section 3.3).

Besides the two main sources used in the automatic extraction of lexico-semantic knowledge, several other alternative sources of knowledge, not further developed in this document, have more recently started to be explored for ontology learning, for instance search engine logs (Costa and Seco (2008)), relational databases, software source code (Grcar et al. (2008)) or file directories.

3.1 Extraction of Lexical Knowledge from MRDs

MRDs are electronic versions of dictionaries, especially designed to be used by or through machines and are usually stored in a database that can be queried via some interface. The first dictionaries known to have a machine-readable format were Merry Webster's Pocket Dictionary (MPD) and Webster's Seventh New Collegiate Dictionary (W7)[1] that were manually keyboarded and distributed in multiple reels of magnetic tape, back in the 1960s (Olney et al. (1967)). From that time, the creation of electronic versions of the dictionaries had in mind helping NLP systems.

Besides the aforementioned MRDs, the electronic version of the Longman Dictionary of Contemporary English (LDOCE)[2] is probably the most representative

[1]The current version of the W7 is available for online search in `http://www.merriam-webster.com/`

[2]The current version of the LDOCE is available for online search in `http://www.ldoceonline.`

MRD when it comes to English NLP. It started to be explored during the 1980s, with the purpose of evaluating how useful it could be for NLP (Michiels et al. (1980)). LDOCE contains simple and restricted vocabulary because it is a learner's dictionary made for non-native English speakers. In addition, the LDOCE entries may contain two kinds of codes that revealed to be very helpful concerning semantic information extraction and WSD: box codes and subject codes, both organised into hierarchies. While the former are a set of primitives to assign type restrictions on nouns, adjectives and on the argument of verbs (primitives such as abstract, animate or human, conforming the classical notion of the hypernymy relation), the subject codes consist of headings and sub-headings that classify the words by subject (terms like engineering or economics) (Bruce and Guthrie (1992)).

From the beginning, MRDs started to be used as an important source of lexical information for the construction of lexical knowledge bases. This happened not only because they use restricted vocabulary in simple sentences (suitable to be exploited), but also because they are highly structured, they are a substancial source of general lexical knowledge (Briscoe (1991)) and they the "authorities" of word sense (Kilgarriff (1997)).

3.1.1 In the beginning

Back in the 1970s and through the 1980s, MRDs started to be the target of empirical studies in order to assess the possibilities of using them as a source of semantic knowledge, useful for NLP (Calzolari et al. (1980)). The work of Nicoletta Calzolari includes the exploration of the definitions in order to organise the dictionary into a lexical database (LDB), where morphological and semantic information about the defined words could be obtained directly (Calzolari (1982)). If the created database is well structured, it is easier to automatically identify some syntactic and semantic relations between the entries of the MRD.

Similar work took place for English when the electronic versions of the LDOCE and the MPD were used as a source of information to build such a structure. Michiels et al. (1980) explored the files of the LDOCE, presented its structure and took some conclusions about the properties of its definitions. Like other authors, they concluded that the vocabulary in a dictionary is very limited, easing its processing in order to obtain relations between syntactic or semantic structures.

In the same year, Amsler (1980) explored the structure of the electronic version of the MPD. He noticed that the text of the definitions often consists of a *genus* and a *differentia*:

- The *genus* identifies the superordinate concept of the defined word. In other words, the defined word is a "type of" the *genus* and there is typically a hyponymy relation between the former and the latter.
- The *differentia* consists of the specific properties responsible for the distinction between the respective instance of the superordinate concept and other instances of the same concept.

If the *genus* is extracted and disambiguated, it is possible to build semantic hierarchies based on the hypernymy relation (for nouns) or troponymy (for verbs). The

com/

terms *genus* and *differentia* are used in most of the publications in this research topic.

Having in mind that it was possible to extract a huge amount of semantic information from the dictionary, Amsler (1981) proposed a taxonomy consisting of hierarchies of nouns and hierarchies of verbs. He called them *tangled hierarchies* and created them based on the analysis of the definitions in the MPD and on the manually disambiguated head of each definition. The hierarchies were organised in a way that the most specific words could be found in the lower levels and the most generic (such as *cause*, *thing*, *class*, *being*,...) in the top. Another conclusion taken by Amsler was that the dictionary contains at least two clear taxonomic relations: is-a (hypernymy) and is-part (part-of).

Taking advantage of the restricted and specific vocabulary and of the regular syntactical occurrences in a MRD, Calzolari (1984) also suggests sets of patterns that are regularly used 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 can be easily extracted from an MRD with the identification of the *genus* of the definition.

Some years later, Markowitz et al. (1986) identified a set of textual patterns that occur in the beginning of the definitions of the W7. The presented patterns imply relations between nouns, namely the superordination and the member-set relations; imply that the defined noun is a human being; and identify verbs or adjectives as active or stative. The following are some examples of the identified patterns:

- Superordination: *any*, *any of*;
- Member-set: *member of*;
- Human noun: *one*;
- Information about verbs in the definition of nouns: *act of* <active verb>*ing*, *the act of* <stative verb>*ing*, *the state of being* <adj>;
- Adjectives: *of or relating to* (stative), *being* (active).

3.1.2 First (semi) automatic procedures

Chodorow et al. (1985) proposed two "head-finding" heuristics to identify the *genus* of a definition: one for nouns and another for verbs. Bearing in mind the structure of the definitions and assuming that a defined concept is often a hyponym of its superordinate concept, they took advantage of the restricted vocabulary used in the definitions to developed semi-automatic recursive procedures aiming the extraction and organisation of semantic information into taxonomic trees. 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 a conjunction of verbs following to, they are all heads.* For example:

- `winter (v): to pass the winter` → pass

- `winter (v): to keep, feed or manage during the winter` → keep, feed, manage

When it comes to nouns, the task is much more complex due to their greater variety, but they could still take advantage of the special and predictable style of their definitions and still came up with an heuristic for the extraction of the *genus*. The heuristic is based on the isolation of 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 (thus, introducing a complement to the head noun);
- a preposition-conjunction-preposition configuration (also introducing a complement);
- a present participle following a noun (thus, introducing a reduced relative clause).

After isolating the substring containing the head, the search for the head begins. It is typically the rightmost noun in the substring. Chodorow et al. (1985) claim 98% accuracy for the heuristic for nouns, but we should remind that this heuristic was only capable of identifying the head of the definition whether that was or not the hypernym of the defined word.

Alshawi (1987) analysed the definitions of the LDOCE where syntactic patterns were identified to make possible the construction of semantic structures based on the meaning of the defined words. These 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. Alshawi (1989) also proposed a specific semantic grammar for the derivation of the definitions of the LDOCE. His main concern was to accomplish partial syntactical derivation based on the structure of the definitions of this specific MRD, so the application of the grammars to unrestricted text or to other dictionaries might not be a good option.

3.1.3 Typical problems

One of the first noticed problems when using dictionaries to build a taxonomy is circularity, often present in dictionary definitions (Calzolari (1977)). This phenomenon occurs when starting by processing some entry, then going to the entry corresponding to the head of its definition and, eventually after several levels of recursion, ending up in some entry that had already been processed. The following definitions constitute a made up example of circularity:

- portion - a **part** of a whole;
- part - a **piece** of something;
- piece - a **portion** of some material;

In Amsler (1981)'s work, circularity is referred to as loops (groups of words defined in a circular way) and the importance they have is discussed. He believes that 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. Sometimes these primitives are related with "covert categories" (Ide and Véronis (1993)), which are basically 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, INTRUMENTAL-OBJECT, is created.

When identifying the *genus* term to obtain hypernymy relations, attention should be paid to certain head words that give special information about the defined word and can be related with other types of relation. Chodorow et al. (1985) called them "empty heads" and gave specific examples (e.g. *one, any, kind, class, manner, family, race. group, member, ...*). To deal with this problem, whenever his procedures met an empty head, the noun word following the preposition *of* (as in *kind* ***of*** *boat*) was interpreted as the head. Although it seemed reasonable, Guthrie et al. (1990) argue that since some of the words Chodorow et al. (1985) considered to be "empty" are usually associated with other relations like, for instance, the word `member` which is related with a member-set relation (Markowitz et al. (1986)) or the word `part` which is related with a 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 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 association of words that appear in the definition with their correct sense in the dictionary. For instance, the disambiguation of the *genus* is needed for the extraction of a taxonomy from a MRD. One of the biggest limitations of Amsler (1981)'s and Chodorow et al. (1985)'s work, is that the disambiguation of the *genus* requires human intervention. Concerning this issue, Bruce and Guthrie (1992) worked on an automatic procedure to accomplish this task, that involves two subproblems:

1. identification of the *genus*/hypernym word from a definition;

2. disambiguation of that word into a concept;

While effective methods had been presented to solve the first problem, the second one is more difficult. However, an algorithm was developed for the disambiguation of the *genus*, taking advantage of the box codes, subject codes and frequency of utilisation associated with each entry of the LDOCE. The algorithm, which the authors claim to have 80% accuracy, is stated as follows:

1. Choose the *genus* sense with the same semantic category as the headword, or with the closest more general category.

2. If there is a tie, choose the sense with the same pragmatic code.

3. If a tie remains or no *genus* sense meets the above criteria, choose the most frequently used sense of the *genus* word.

3.1.4 Broad-coverage parsing in MRDs processing

After some discussion about the advantages and the drawbacks of using string patterns or structural patterns to extract semantic information contained in 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, and they proposed the use of a broad-coverage grammar to parse the dictionary definitions in order to obtain rich semantic information. String patterns are based on specific textual constructions of the definitions and were used by Chodorow et al. (1985), Markowitz et al. (1986) and others, while structural patterns are based on the syntactic structure of the definition, obtained after the syntactic analysis, made by a broad-coverage parser. Previous work on the automatic extraction of relations from MRDs (using string patterns) reported very good results, but this work focused mainly on hypernym extraction. When it comes to the extraction of relations depending on the *differentia*, string patterns have several reported limitations:

- When there is an enumeration of concepts at the same level: *to make **laws**, **rules** or **decisions***;

- When there are parentheses in the middle of the definition;

- When it is necessary to identify functional arguments;

- When there are specific relations inside the definition: in *pianta erbacea com **bacche di color arancio*** the color feature should not be extracted as a feature of the defined word.

In spite of seeming an overkill to use a broad-coverage parser for definition text, the authors make the point that there are cases (relative clauses, parenthetical expressions, and coordination) when its use is warranted. The following is an example of an heuristic for the extraction of the *purpose* relation: *if the PP*[3] *with* `for` *is not a post-modifier of a verb* `used`*, then a purpose relation between the defined word*

[3]prepositional phrase

and the head(s) of the PP can be hypothesised if the nearest noun that the PP post-modifies is the genus term.

Although dictionaries have been explored for several purposes, such as parsing, deriving semantic structures or WSD, to our knowledge they have not been converted into an independent resource of its own before the late 1990s (after several publications in that direction (Dolan et al. (1993); Dolan (1994); Vanderwende (1994, 1995))), when MindNet (Richardson et al. (1998)) was presented, which therefore can be said to be a sort of independent lexical ontology in a way that previous work was not.

Dolan et al. (1993) describe a strategy to build a structure of lexical knowledge automatically from LDOCE. Their approach uses a broad-coverage parser to process the dictionary entries, which avoids the need to adapt it for different MRDs. The authors state that much information about a word can be found in the definition of other words. Looking for a specific word in the definitions of other words makes it is possible to obtain many relations that include the first word. In order to create a semantic network, the set of relations to extract must be defined. The definitions are then parsed and searched for patterns that imply semantic relations. Each identified relation is added to the sense entry in a semantic structure representing the definition. The resulting network contains words linked by means of semantic relations. Inferencing over the obtained network can be done to resolve semantic ambiguities on text.

Dolan (1994) also worked on a heuristic approach to automatically identify related senses of the same word in a dictionary, where each word can have definitions divided into more than one sense. Dolan (1994)'s approach consisted of identifying which senses are semantically related and which ones are fundamentally different, offering benefits for semantic processing and for the mapping of word senses across multiple MRDs. This process was then called word sense "ambiguation".

Vanderwende (1994) presents an algorithm for the automatic interpretation of noun sequences in unrestricted text. Her system uses broad-coverage semantic information, acquired automatically by analysing the definitions in an on-line dictionary. Vanderwende (1995) also worked on treatment of lexical ambiguity present in the language used by on-line dictionaries. A dictionary is processed multiple times, each time refining the lexical information previously acquired and identifying new information.

In his PhD thesis, Richardson (1997) discusses the creation of the lexical knowledge base (LKB) that would be known as MindNet. In order to achieve his goal, the entries of the LDOCE are converted into a more formal representation, resulting in a dictionary called MIND (Microsoft Natural Language Dictionary). The syntactical trees of the definitions are obtained with the help of a broad-coverage parser and are then transformed into a logical form. After this, a set of heuristics is applied in order to convert the logical form into relational form, where semantic relations are clear. The relations that include a word can be obtained from all the definitions where that word occurs and the resulting structure can be inverted, giving rise to a lot more relations. With the resulting resource it is possible to browse for relation paths (see Section 2.2.4). Similarity of words can be inferred and other conclusions can be taken from the analysis of the paths.

O'Hara (2005) wrote about the empirical extraction of semantic relations from dictionaries. Special attention was given to the information in the *differentia* to find distinctions between co-hyponyms[4] to accomplish WSD. Relations such as used-for or has-size can be used to learn important information about the concepts. Dictionaries follow lexicography rules that ease the extraction of this kind of knowledge, however definitions are always incomplete or vague when it comes to certain details needed to understand some concept. In his studies, O'Hara uses WordNet (Fellbaum (1998)) as a simple dictionary. Although the usual would be to adapt the parser, all the definitions are pre-processed and transformed in order to be easily interpreted by a general parser. This is done because dictionary definitions are often given by sentence fragments that omit the defined word. For example, the definition for *lock* is "*a fastener fitted to a door or drawer to keep it firmly closed*". This entry is transformed into "*a lock is a fastener...*". There are different transformations, depending on the grammatical category of the words. In his approach, O'Hara 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 are disambiguated into semantic relations between the underlying concepts. Isolating the disambiguation from the extraction allows great flexibility over earlier approaches. After a disambiguation process, the relations are weighted according to their relevance to the assigned concepts, resulting in a labeled direct graph where each link has a probability attached. The network is then converted into a Bayesian network (Pearl (1988)). The author believes that the Bayesian network representation of the differentiating information can be used to improve WSD systems that use both statistical classification as well as probabilistic spreading activation.

3.1.5 Critical work

MRDs are certainly an important source of knowledge about language and the world but their organisation does not favour their direct use as NLP tools, since they were created in order to be read by humans. Wilks et al. (1988) mention three assumptions that should be made to accomplish the objective of automatically extracting knowledge from a dictionary, and transform MRDs into machine tractable dictionaries (MTDs):

1. *Sufficiency*: Determines whether the knowledge is strong enough and contains enough linguistic knowledge on the world to be the target of computational text processing.

2. *Extricability*: Determines whether it is possible to specify a set of computational procedures capable of extracting large scale semantic information from a MRD without human intervention, in a general format suitable for subsequent text analysis processes.

3. *Bootstrapping*: Addresses the initial linguistic knowledge needed to automatically extract knowledge from the definition texts.

The authors say that projects based on the manual construction of semantic structures have pessimistic visions concerning *Extricability* and *Bootstrapping*. Three

[4]Coordinate terms or co-hyponyms, are terms that share one hypernym.

methods based on the former assumptions are presented for the automatic extraction of knowledge. The methods differ in the amount of initial information needed:

- The first approach is based on co-ocurrences that permit the establishment of associations between words, without needing initial linguistic information.
- The second uses a grammar and a collection of linguistic patterns enabling, besides other items, the identification of the *genus* (hypernym) and the *differentia* for each entry in the dictionary.
- The last approach is the one that needs more initial knowledge, but permits the creation of a semantic structure free from circular references. Since circularity makes the knowledge too vague it is better to remove it. Starting with a set of 3,600 semantic units, corresponding to the various senses of the 1,200 words used to define the controlled vocabulary of the LDOCE, the algorithm analyses the additional words in the dictionary. Of the latter, those whose definition uses words with an existing semantic unit lead to the generation of a new semantic unit of the entry. According to the authors, after four iterations all the words are processed.

Ide and Véronis (1994) produced critical work about research on information extraction from dictionaries. The authors affirm that all the research done so far had not achieved significantly more than the extraction of small and limited taxonomies. Two problems concerning the information in dictionaries, that seems to be inconsistent and incomplete, are discussed:

- Dictionaries use inconsistent conventions to represent knowledge. Work around the identification of the conventions turns out to be very time-consuming.
- The definitions are not as consistent as they should be. There are many variations to say the same thing because dictionaries are the result of several lexicographers work for several years, and reviews and updates increase the probability of inconsistencies.

In different dictionaries (or sometimes, even in the same) there are definitions made up from hierarchies with very high levels and it is sometimes difficult to identify terms that belong to the same level.

In order to assess the information extracted from MRDs, Ide and Véronis (1993) performed a quantitative evaluation of automatically extracted hypernymy relations. Hypernymy was chosen because it is the least arguable semantic relation and the easiest to extract. The authors believe that, if the results for hypernymy are poor, they will be poorer for more complex domains and less clearly cut relations. The evaluation methodology consisted in comparing an "ideal" hierarchy, manually constructed, with hierarchies extracted from five dictionaries. The automatic extraction of hierarchies was based on the heuristics by Chodorow et al. (1985), giving rise to *tangled hierarchies* that were later manually disambiguated. After inspection, it was noticed that these hierarchies had several serious problems:

- Incomplete information: some terms are (relatively randomly) attached too high in the hierarchy; some heads of definitions are not the hypernym of the defined word, but the "whole" that contains it; overlaps that should occur between concepts are sometimes missing.

- Difficulties at higher levels: 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 of the hierarchy, possibly when lexicographers lack terms to designate certain concepts.

The authors state that hierarchies with these 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%.

3.1.6 Other approaches

In her PhD thesis, Barriere (1997) presents a method for transforming a MRD for children into a LKB, made from Conceptual Graphs (Sowa (1992)). The American Heritage First Dictionary (AHFD) was chosen due to:

- Its limited size;
- The day-to-day knowledge and simple world knowledge included;
- The complete and simple sentence structure;
- Being a closed world because almost all the words used in the definitions are themselves defined;
- Bootstrapping capabilities, possible because of the closed-world system;
- Naive view of things (in contrast to an adult's dictionary);
- Limited polysemy (limited number of senses for each word).

Conceptual graphs were used because they present a logic-based formalism and are flexible to express the background knowledge necessary for understanding natural language. Most of the structures used during the development of the LKB were based on this formalism. The usage of conceptual graphs allows the coexistence of ambiguous and non-ambiguous information in the LKB. All the definition sentences in the AHFD were transformed into conceptual graphs after being tagged, parsed, parsed to conceptual graph transformations, structurally disambiguated and finally semantically disambiguated giving rise to an automatically created type hierarchy. The LKB is then constructed using those graphs, exploring cluster formations and the expansion of the hierarchy using "covert categories". Concept clusters are large structures used to represent the meaning of a word by its interaction with other words. They consist of groups of words that help define each other.

It was concluded that "covert" classes should be included in the concept hierarchy. These unlabeled categories are often superclasses whose subclasses occupy the case relation to a verb (for example *to live somewhere*). This lead to different relations than the usual synonymy, hypernymy and meronymy.

Nichols et al. (2005) introduced a system that automatically constructs ontologies by extracting knowledge from dictionary definition sentences. Their approach combines deep and shallow parsing of the definition sentences and generates semantic representation by the robust minimal recursion semantics (RMRS). For each

definition, ontological relations are extracted from the most informative semantic representation. Using the deepest possible result, 81,582 relations were extracted from the Lexeed Semantic Database of Japanese and two evaluations were performed:

- An automatic evaluation consisting of the verification of the extracted relations in WordNet Fellbaum (1998) and GoiTaikei (Ikehara et al. (1997)), a manually created Japanese ontology. The results for the relations obtained with the deep parsing had the best confirmation rate, 55.74%, and 63,31% if only nouns were considered. When it comes to relations obtained with the deepest result, the confirmation rates were 50,79% overall and 57,68% for nouns.

- A manual evaluation, consisting of a hand-verification of a set of the acquired relations. 88.99% accuracy is claimed.

WordNet and GoiTaikei seem to lack complete cover, since over half the relations were confirmed with only one resource. This might be what caused the difference between automatic and manual evaluation. The authors claim that their approach is easy to maintain and expand, because it requires few rules, and can be easily extended to cover any language with RMRS resources.

3.1.7 Discussion

Table 3.1 puts side-by-side the attempts to extract and structure knowledge from MRDs referred in this section. It includes the exploited MRD(s), the relations extracted, the method used for the extraction and also, when referred by the authors, the name of the structure produced with the results.

There is no doubt that MRDs are an interesting source of lexico-semantic knowledge, since methods for extracting this kind of information from MRDs have shown a relative success, as discussed in this section. MRDs are possibly the main sources of general lexical knowledge and, since they are created by experts (lexicographers), there is an high degree of confidence about the way words and senses are handled. Moreover, they are easier to process because:

- they are already structured according to words and their meanings;

- they typically contain simple and restricted vocabulary, which can often be predicted, giving rise to less ambiguity.

On the other hand, some works (Ide and Véronis (1993, 1994)) revealed some problems concerning MRDs processing. For instance, they generally contain incomplete information. Everybody agrees that the knowledge in MRDs is limited, since they have a fixed number of entries (Hearst (1992)), as well as broad, not covering specific domains. Handcrafted lexical ontologies, such as WordNet (Fellbaum (1998)) or Cyc (Lenat (1995)), suffer from the same kind of problems concerning the lack of domain-specific knowledge, and seem insufficient or inadequate for most NLP applications (Riloff and Shepherd (1997); Roark and Charniak (1998); Caraballo (1999)). So, authors have moved forward to extract lexico-semantic knowledge, that cannot be found neither in MRDs nor in handcrafted lexical ontologies, from textual corpora.

Work	MRD(s)	Relations	Extraction	Structure
Calzolari et al. (1980); Calzolari (1982, 1984)	Italian Machine Dictionary (DMI)	Hyponymy, "restriction" or "modification"	Textual patterns matching	
Amsler (1980, 1981)	MPD	Hypernymy, troponymy, part-of	Textual patterns matching	Tangled hierarchies
Chodorow et al. (1985)	W7	Hyponymy	Textual patterns matching	Tangled hierarchies
Markowitz et al. (1986)	W7	Superordination, member-set, human, active/stative verb or adjective	Textual patterns (in the beginning of the definitions) matching	
Alshawi (1987, 1989)	LDOCE	Class, purpose, manner, has-part	Syntactic patterns matching, a semantic grammar for the LDOCE definitions	(so called) "semantic structures"
Richardson (1997); Richardson et al. (1998)	LDOCE, AHD3	Hypernymy, Causation, Meronymy, Manner, Location and many more	Broad-coverage parser	Lexical Knowledge Base (MindNet)
Barriere (1997)	AHFD	On, with, hypernymy, part-of, material, instrument, time, goal and more.	Conceptual graphs	Lexical Knowledge Base
O'Hara (2005)	Wordnet	Relations found in the *differentia* (e.g. used-for, has-size)	Broad-coverage parser over preprocessed and simplified definitions	Bayesian network
Nichols et al. (2005)	Lexeed Semantic Database of Japanese	Hypernymy, synonymy, abbreviation, domain, other	Deep and shallow parsing combined	Ontology

Table 3.1: Summary of attempts for knowledge extraction from MRDs.

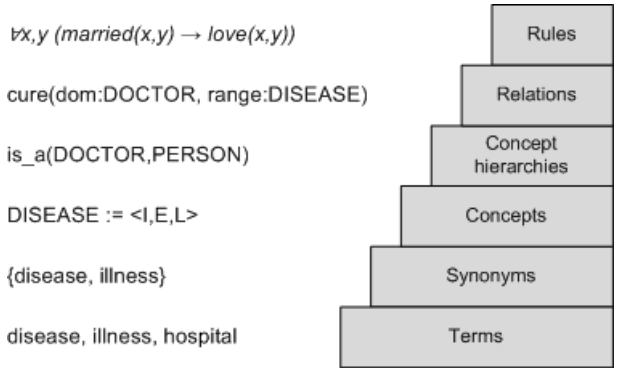

Figure 3.1: Ontology learning layer cake (adapted from Buitelaar et al. (2005))

3.2 Ontology Learning from Textual Corpora

As discussed in the end of the previous sections, in order to overcome some of the limitations of using MRDs, authors have moved on to extract lexico-semantic knowledge from textual corpora. The latter can hence be seen as an important source of domain knowledge (Brewster and Wilks (2004)), which can either be used to enrich existing lexical ontologies or to create new domain-based ontologies.

Concerning ontology learning from text, Buitelaar et al. (2005) establish six subtasks and organise them into a layer cake (see Figure 3.1). Terms are the base of this cake, since they are the smallest units in text and are the linguistic symbols to represent concepts. In order to acquire concepts themselves, terms need to be grouped into synonyms, which are one of the main concerns of ontology construction. Concepts may have several interactions so, ontology construction is concerned with, first identifying inheritance or taxonomic relations among them and then other kinds of relations. In the top of the cake, rules can be defined to derive facts that are not explicitly encoded in the ontology.

Despite the being adequate for describing ontology learning from general text, the layer cake does not capture well the same task, when MRDs are chosen as the source of knowledge. In the latter case, the three base layers are not as complex as in unrestricted text, since in MRDs most of the terms correspond to head words, which are typically divided into their possible senses. Furthermore, in MRDs, many synonyms can be easily extracted by looking upon the structure of the definitions.

Brewster and Wilks (2004) simplify Buitelaar et al. (2005)'s layer cake and define three major steps in ontology construction from text: associating terms, constructing hierarchies and labelling relations. In this section, work where unrestricted text is used to acquire lexical and domain knowledge is presented, according to the three major steps given by Brewster and Wilks (2004). Some of the works rely on linguistic principles (textual or syntactic patterns), others are based on statistic methods and some other have little bit of both paradigms.

3.2.1 Associating terms

The first step for ontology learning consists of identifying which terms are associated with which, in order to start taking some conclusions on similarity and concept formation. Most of the work involved in term association rely on Harris distribu-

tional hypothesis (Harris (1968)), which assumes that similar words tend to occur in similar contexts. Some linguistic constructions can furthermore be combined with the latter assumption in order to improve co-occurrence based clustering algorithms. These constructions comprise:

- conjunctions (*lions and tigers and bears*);
- lists (*lions, tigers, bears...*);
- appositives (*the stallion, a white Arabian*);
- nominal compounds (*Arabian stallion*)

Riloff and Shepherd (1997)'s work is based on the observation that terms of the same category often occur in the aforementioned constructions. However, their definition of context is simply one noun to the left and one noun to the right for head nouns in sentences.

They were the first to apply bootstrapping for building domain-specific semantic lexicons which, in the case of their work, are basically clusters of terms belonging to the same category. The input of their system is a text corpus and an initial set of words strongly related with a chosen semantic category (seed words). For instance, *airplane*, *car*, *jeep*, *plane*, *truck* are the seed words for the category *vehicle*. In order to obtain new member of the category represented by the seed words, the algorithm goes as follows:

1. All sentences in the corpus where one of the seed words occur are identified and parsed into noun phrases (NP), verb phrases (VP) and prepositional phrases (PP);
2. The context surrounding each occurrence of the seed word and where it is the head noun, is collected.
3. A category score of a word W in the category C is computed as follows: $Score(W, C) = \frac{freq.\ of\ W\ in\ C's\ context}{freq.\ of\ W\ in\ the\ corpus}$
4. After removing stopwords[5] and words with corpus frequency less than six, the remaining nouns are sorted by category score and ranked.
5. The top five nouns that are not already in the seed words list are added to this list dynamically and the algorithm goes back to step 1.

After several iterations, the system outputs a ranked list of nouns, supposedly members of the chosen category.

For each category, the top 200 ranked words were selected and rated by human judges in a scale from 1 (no association with the category) to 5 (core member of the category). Considering the amount of words rated 4 (subpart or member of the category) or 5, the algorithm revealed a precision between 12.5% and 22.5%, depending on the category.

[5]Stopwords are general and very frequent words, usually functional, like prepositions, determiners or pronouns.

Roark and Charniak (1998) built on Riloff and Shepherd (1997)'s work by actually focusing on conjunctions, lists, appositives and nominal compounds for noun clustering and also, by changing some other parameters in the algorithm. More precisely, they propose a new ranking measure that allows for the inclusion of rare occurrences and only considers words in co-occurrence situations (conjunctions, lists and appositives), they try to select the most frequent head nouns in the corpus as initial seed words, and they deal with compound nouns in a separate step. Consequently, Roark and Charniak (1998)'s algorithm performed better and a precision between 20% and 40%, depending on the category chosen, is reported.

Pantel and Lin (2002) propose an algorithm, called Clustering by Committee (CBC), for automatically extracting semantic classes, consisting of clustered instances, such 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 terms are found. Then, a set of tight clusters, with representative elements of a potential class (committees), is constructed. The idea is to form as many dissimilar committees as possible. Finally, each element is assigned to its most similar clusters.

The committee members for the previously shown cluster, consisting of elements that unambiguously describe members of the class, would then be:

> *blue, pink, red, yellow*

CBC evaluation was accomplished automatically by mapping clusters with WordNet synsets, with 60.8% precision. Manual evaluation was also performed using a random sample of test data, with 72%. Automatic evaluation was performed for the same sample and agreed with manual evaluation 88% of the time.

Latent Semantic Analysis (LSA) (Deerwester et al. (1990)) is a technique of analysing relationships between sets of documents and the terms they contain, by producing a set of concepts related to the both of them. 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 typical example of weighting the importance of the elements in the matrix is term frequencyinverse document frequency (TF-IDF): the element of the matrix is proportional to the number of times the terms appear in each document, where rare terms are upweighted to reflect their relative importance. PMI-IR is an alternative algorithm for LSA, presented by Turney (2001) for learning synonyms from the Web. PMI-IR uses pointwise mutual information (PMI) to score the similarity between two words, which can be seen as a conditional probability of *word1* occurring, given that *word2* occurs. Both probabilities are calculated using Information Retrieval (IR) and and a web search engine:

$$Score(word2) = \frac{P(word1 \& word2)}{P(word2)}$$

PMI-IR was used to answer to TOEFL[6] questions where given a lead word and four alternative words, the problem is to select the alternative most related in meaning to the lead word. PMI-IR performed better that LSA with a TOEFL evaluation of 74% against LSA's 64%.

According to works referred by Gamallo (2008), syntactically-based methods usually perform slightly better. For Portuguese, Gamallo (2008) compared window and syntax based strategies for the extraction of semantically related words. In window-based techniques, the context of a word consists of its nth adjacent words. The word order can be taken or not into account, giving rise to two context definition methods: word order and bag of words, respectively.

In syntactic-based techniques, the identification of syntactic dependencies are considered to define the context of a word. Sentences are POS-tagged and regular expressions are used to identify syntactic dependencies. Then, lexico-syntactic contexts are extracted from the dependencies and the occurrence of the lemmas in those contexts is counted and stored.

Experiments in a general-purpose Portuguese newspaper were performed. For each member of a sample list of proper nouns, a ranked list with the top-5 similar nouns found was computed. Human judges only had to classify the members of the sample list and the similar nouns as belonging to one of seven categories, namely counties, capitals of countries, Portuguese towns, politicians, organisations, press agencies and football. Nouns are then related if they belong to the same category and are thus co-hyponyms. After measuring the precision of the lists generated by the three tested methods, the syntactic-based method performed better and seems to be the only one that clearly improves for more frequent lemmas. The results obtained suggest that structure information is very helpful for identifying meaningful contexts.

3.2.2 Constructing hierarchies

Most of the works aiming at extracting lexical knowledge from text are more concerned with the extraction of hypernymy relations (Biemann (2005)). Other kinds of relations are examined much less. Concerning this kind of research, the discovery of relations from text using large corpora became the paradigm in ontology construction after Hearst (1992)'s seminal work, where an automatic method to discover lexico-syntatic patterns, used for the acquisition of hyponyms, is proposed.

Besides indicating a hyponymy relation, the discovered patterns must occur frequently and in many text genres, (almost) always indicate the relation of interest and should be recognised with little or no pre-encoded knowledge. The method, which could eventually be adapted to any lexical relation, is yet only applied to hyponymy. After deciding on a lexical relation (e.g. hyponymy), a list of word pairs for which this relation is known to hold is gathered. Sentences in which these words both occur are extracted from a large corpus and the most common contexts found are hypothesised to yield patterns that indicate the target relation.

Following, the list of patterns used to extract hyponymy relations (NP stands for Noun Phrase). The first three patterns were discovered by observation while the other three were discovered using the proposed method.

[6]TOEFL stands for Test of English as a Foreign Language.

1. `NP such as NP , NP ... , (and | or) NP`
   The bow lute, such as the Bambara ndang, ...
   ⇒ (*Bambara ndang* hyponym_of *bow lute*)

2. `such NP as NP ,* (and | or) NP`
   ... works by such authors as Herrick, Goldsmith and Shakespear.
   ⇒ (*Herrick* hyponym_of *author*), (*Goldsmith* hyponym_of *author*), (*Shakespeare* hyponym_of *author*)

3. `NP , NP* , or other NP`
   Bruises, ..., broken bones or other injuries ...
   ⇒ (*bruise* hyponym_of *injury*), (*broken bone* hyponym_of *injury*)

4. `NP , NP* , and other NP`
   ... temples, treasuries, and other important civic buildings.
   ⇒ (*temple* hyponym_of *civic building*), (*treasury* hyponym_of *civic building*)

5. `NP , including NP ,* or | and NP`
   All common-law countries, including Canada and England ...
   ⇒ (*Canada* hyponym_of *common-law country*), (*England*, hyponym_of *common-law country*)

6. `NP , especially NP ,* or | and NP`
   ... most European countries, especially France, England, and Spain.
   ⇒ (*France* hyponym_of *European country*), (*England* hyponym_of *European country*), (*Spain* hyponym_of *European country*)

The results obtained after looking for these patterns were compared to the information in Princeton WordNet (Fellbaum (1998)) and hyponymy relations that are not found in WordNet were suggested as new entries.

Despite the absence of a precision value, Hearst (1992) claims the quality of the results seems high overall, but she admits difficulties, such as the occurrence of metonymy[7] (e.g. in the relation *king* hyponym_of *institution*) and underspecification (e.g. in the relations *steatornis* hyponym_of *species* and *device* hyponym_of *plot*).

Caraballo (1999) also pretended to do more than the automatic creation of clusters of related words and proposes the automatic construction of hierarchies, similar to the ones in manually built lexicons. In his method, which is a combination of pattern detection and clustering methods, noun candidates are obtained from a newspaper corpus using data on conjunctions and appositives. For all nouns, a co-occurrence matrix, consisting of a vector for each noun in the corpus with the number of times each other noun appears in a conjunction or appositive, is set up. Similarity between two nouns is calculated in the following way:

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

For labelling this hierarchy in a post-processing step, *Hearst-like* patterns are used for finding hypernym candidates, which are placed as common parent nodes for clusters, if appropriate. The results were evaluated by human judges and the

[7]Metonymy is a figure of speech in which a thing is not called by its own name, but by the name of something intimately associated with that thing.

method revealed between 33% (considering only the best hypernym accepted by all judges) and 60% (considering any of the second and third best hypernyms accepted by at least one judge) precision.

Cederberg and Widdows (2003) used a variant of Latent Semantic Analysis (LSA) to improve the precision and recall of hyponymy relations extracted automatically from a corpus, also using *Hearst-like* patterns. After hand-checking, a set of extracted hyponymy relations were 40% either correct or needing slight modifications (e.g. depluralization or the removal of an article).

Having in mind that an hyponym and its hypernym are expected to be similar, LSA is used to compute the similarity of terms in the extracted relations. So, relations were ranked according to the similarity of their terms and, in the top 100 relations, precision was 58%, which suggests the effectiveness of this method concerning the reduction of errors.

Regarding that most of the potential hyponymy relations that could be extracted are not expressed by the six *Hearst-patterns*, Cederberg and Widdows (2003) tried to improve the recall of their method using coordination as a cue for similarity. They give the following sentences to illustrate their inference:

> *This is not the case with sugar, honey, grape must, cloves and other spices which increase its merit.*[8]

Provides that:

clove hyponym_of *spice*

> *Ships laden with nutmeg or cinnamon, cloves or coriander once battled the Seven Seas to bring home their precious cargo.*[9]

Suggests that *nutmeg*, *cinnamon*, and *coriander* are also spices, because they appear in a the same list as cloves, which is a cue for semantic similarity. The following hyponymy relations can thus be learned:

nutmeg hyponym_of *spice*
cinnamon hyponym_of *spice*
coriander hyponym_of *spice*

Using the correct relations extracted in the first phase (without the LSA filter), for each hyponym, the top ten most similar words according to this principle were collected and tested for having the same hypernym. The result was that, after manual scoring, precision improved a little, while the number of relations obtained was ten times higher, which shows a clear recall improvement.

After testing the two independent techniques, Cederberg and Widdows (2003) combined both, in order to improve the overall performance. The LSA method was applied to the extended set of relations and a precision of 64% was reached.

In order to obtain a substantial set of hyponymy relations, Pantel and Ravichandran (2004) worked on overcoming one of the limitations of CBC (Pantel and Lin (2002)) – it does not give an actual name to the concepts formed by the committees – and propose an automatic method for labelling word clusters.

[8]Sentence taken from the British National Corpus (BNC)
[9]Sentence taken from the British National Corpus (BNC)

They start with a list of semantic classes, in the form of clusters of words, generated for instance by CBC. At the end, their system outputs a ranked list of possible concept names for each class.

Initially, feature vectors are extracted for each word in a cluster. Then, CBC is used to obtain the committee of each cluster. Each commitee's grammatical signatures are computed. Finally, simple syntactic patterns (e.g. apposition, nominal subject...) are used to discover class names for each signature.

For each term in the syntactical relationships found with a committee of a class, mutual information score is summed up and the highest scoring term is selected as the name of the class. With a name for the class, hyponymy relations can be defined between the instances of the class and its name. For instance, if the previously exemplified class was labeled *color*, the following relations would be extracted:

blue hyponym_of *color*
pink hyponym_of *color*
red hyponym_of *color*
yellow hyponym_of *color*

The results of Pantel and Ravichandran (2004) system were the subject of manual evaluation. 125 randomly selected concept clusters were presented to human judges, together with their top-5 ranked system labels, a human created label and a WordNet label (when possible). All labels were randomly ordered and human judges were asked to classify the correctness of each label. Human created labels revealed 93.6% precision, while system labels 77.5%. WordNet labels had only 19.9% precision because most of the times it was not possible to find a well-suit label in WordNet.

Concerning the evaluation of hyponymy relations, two judges annotated two random samples of 100 relationships: one from all 159,000 hyponyms and one from the subset of 65,000 proper nouns. The total precision was 68% but, if only proper nouns were considered, the precision improved to 81.5%.

Snow et al. (2005) claim that methods for constructing lexicons only by the identification of textual patterns (similar to Hearst's), despite being recurrent, have several problems. For instance, manual identification of the patterns is not very interesting and can be biased by the designer of the patterns. Additionally, most of the approaches use only a small finite set of patterns, that are unlikely to capture all the the occurrences of the target relation(s) in running text.

So, they rely on machine learning techniques to discover hyponymy patterns and propose an automatic classifier that decides if a hypernymy relation holds between two nouns. The training algorithm works as follows:

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 which are good cues for hypernymy.
4. Train a hypernymy classifier based on the previous features.

Automatic discovery of patterns indicative of hypernymy can be achieved by searching, in a corpus, for repeating patterns between words that are classified as

a hypernym-hyponym pair. Besides rediscovering the six *Hearst patterns*, which gives a quantitative justification to Hearsts intuition, Snow et al. (2005) were able to discover the following new patterns:

- `NP like NP`
- `NP called NP`
- `NP is a NP`
- `NP, a NP`

The hypernym-only classifier, based on the intuition that an hypernymy relation is likely to hold between two words if both words occur in one or more of the lexico-syntactic patterns discovered, showed a 132% of f-score[10] improvement over a classifier based solely on the six *Hearst patterns*.

The classification of hypernymy relations can be further improved if a model to classify coordinate terms[11] is combined with the hypernym-only classifier. The probability that a hypernymy relation holds between two terms is thus calculated based not only on the probability given by the hypernymy classifier, but also in the probability that terms coordinated with each of the hyponyms have the same hypernym.

Especially aiming at processing large corpora (e.g. the Web), Pantel et al. (2004) developed an algorithm for the extraction of hypernymy relations that achieves similar performance and efficiency to a linguistically-rich method. The algorithm learns hypernymy textual patterns automatically with a technique based on the minimum edit distance[12].

In the context of (Brazilian) Portuguese, Freitas (2007) discusses the extraction of hypernymy relations from domain and also general corpora. To achieve their purpose, some *Hearst patterns* were adapted to Portuguese, which resulted in the following patterns:

- `NP (tais) como NP , NP ... , (e | ou) NP`
  A tentativa posterior de clonar outros mamíferos tais como camundongos, porcos, bezerros,....
  ⇒ (*camundongos* hyponym_of *mamíferos*), (*porcos* hyponym_of *mamíferos*), (*bezerros* hyponym_of *mamíferos*)

- `NP , NP* , (e | ou) outros NP`
  ... a experiência subjetiva com o LSD-25 e outros alucinógenos.
  ⇒ (*LSD-25* hyponym_of *alucinógeno*)

- `tipos de NP: NP , NP ... , (e | ou) NP`
  Existem dois tipos de cromossomos gigantes: cromossomos politênicos e cromossomos plumulados.
  ⇒ (*cromossomos politênicos* hyponym_of *cromossomos*), (*cromossomos plumulados* hyponym_of *cromossomos*)

[10] F-score is a measure where precision an recall are combined.
[11] Coordinate terms, also known as co-hyponyms, are terms that share one hypernym.
[12] Minimum cost/edit operations for transforming a string into another.

- `NP chamad(o|os|a|as) (de) NP`
  ... a alta frequência da doença mental chamada esquizofrenia.
  ⇒ (*esquizofrenia* hyponym_of *doença mental*)

An important contribution of Freitas (2007) and Freitas and Quental (2007)'s work is the conclusion that even though inferences are only explored a few times, they can be of great value to build domain specific taxonomies. For instance, starting with the following relations:

pineapple hyponym_of *fruit*
fruit hyponym_of *food*

It is possible to infer a new relation:

pineapple hyponym_of *food*

In this context, several taxonomies were generated and validated manually with a precision of 90% and 60%, respectively after processing a health-domain corpus and a general corpus.

Also for Portuguese, Baségio (2006) worked on the semi-automatic extraction, from text, of relevant terms of a domain and taxonomic relations between them. Instances, such as names of people or countries, are not extracted and the extraction of taxonomic relations is based on multiword terms, Hearst patters and other patterns found in the literature (Morin and Jacquemin (2004)). Syntactically-annotated corpora was used to perform this work.

Following the work of Baségio (2006), Ribeiro Junior (2008) developed a plugin for Protégé[13], capable of processing a text and extracting concepts, semantic categories and of suggesting hypernymy relations. The hypernymy patterns used are the ones suggested by Baségio (2006), but this time they are applied to the output of PALAVRAS (Bick (2000)), a broad-coverage parser for Portuguese. One of the methods tested for suggesting hypernymy relations takes advantage of several semantic tags provided by PALAVRAS.

3.2.3 Labelling relations

As we have seen in the previous section, many works aiming at the extraction of hyponymy relations from text were inspired by Hearst (1992). Moreover, Hearst (1992) also inspired works concerning the extraction of other types of relations, such as meronymy (Berland and Charniak (1999); Girju et al. (2003b, 2006)), cause (Girju and Moldovan (2002); Khoo et al. (2000)) and manner (Girju et al. (2003a)).

Cimiano and Wenderoth (2007) present an approach for the automatic acquisition of qualia structures (Pustejovsky (1991)), which are structures aiming to describe the meaning of lexical elements, earlier presented in Section 2.2.1 of this proposal. Considering that *Hearst patterns* occur rarely, Cimiano and Wenderoth (2007) propose looking for them in the Web, willing to decrease the problem of data sparseness.

For each qualia term, a set of search engine queries for each qualia role is generated based on known lexico-syntactic patterns. The first 50 snippets returned are

[13]Protégé is a well-known open source editor for Semantic Web ontologies.

downloaded and part-of-speech tagged. Then, patterns, defined over part-of-speech tags, conveying the qualia role of interest, are matched to obtain candidate qualia elements. Finally, the candidates are weighted and ranked according to one of the following tested measures: Web-based Jaccard Measure, Web-based Pointwise Mutual Information, Web-based Conditional Probability or Conditional Probability.

Besides manual evaluation, a gold standard with the qualia structures of 30 words was created manually by non-linguistic participants that were asked to provide qualia elements for each qualia role. The participant agreement was only 11.8%, which suggests this task is certainly difficult, especially for the telic role, where it was only 7.29%. The ranking measure with the best f-measure was Conditional Probability with 17.1%. This value was similar in the manual evaluation (17.7%).

Kietz et al. (2000) presented a method for semi-automatic domain specific ontology acquisition. In their methodology, GermanNet (which is a lexical ontology for German) is used as a top level generic ontology, that helps to acquire the basic concepts. As for domain-specific concepts, a dictionary that contains important corporate terms, described in natural language, is used to acquire and classify them. Since the authors' decision is to achieve the construction of a domain-specific ontology, too generic concepts are removed, based on frequency measures. It is assumed that domain concepts have both high-frequencies in domain-specific copora and low frequencies in general corpora. However, all dictionary entries were considered as domain-specific concepts and were thus not removed in the latter step.

After obtaining the concepts, a taxonomy is learned from text by pattern matching (again, inspired by Hearst (1992)) and also by the decomposition of compound nouns. Also, frequently co-occurring concept couplings are used to suggest non-taxonomic relations. Kietz et al. (2000)'s method is definitely interesting, but the user seems to play an important role concerning the revision of proposed concepts and also when it comes to accepting and labelling suggested relations.

Regarding the automatic extraction of several relationships, or facts of different types, from corpora, several systems have been developed for English. While some of these systems need some kind of human input with some seeds or other clues about the information they should extract (e.g. Snowball, KnowItAll) others can recognise potentially interesting relational tuples without previous knowledge and even figure out suitable labels for the acquired relations (e.g. TextRunner or Kavalec and Svatek (2005)'s work).

Dual Iterative Pattern Expansion (DIPRE) (Brin (1998)) is a technique for extracting a structured relation from a collection of HTML documents. To achieve its purpose, a set of hand-tagged seed tuples, holding the relations, must be provided. This is though the only required training. For instance, for the extraction of relations of type locationOf(*location*, *organisation*), the following tuples could be provided: (*Redmond*, *Microsoft*), (*Cupertino*, *Apple*), (*Armonk*, *IBM*), (*Seattle*, *Boeing*) and (*Santa Clara*, *Intel*). After finding all close occurrences of both the related entities in the collection, patterns where they occur are learned and can be used to extract new tuples holding the same relation.

Snowball (Agichtein and Gravano (2000)) is a system for extracting structured data from plain-text documents with minimal human participation, built on the idea of DIPRE, but extending it to incorporate automatic pattern and tuple evaluation for extracting relations from large text collections.

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 extraction patterns (some of them adapted from Hearst (1992)).

KnowItAll uses the extraction patterns with classes (e.g. cities, movies, ...) in order to generate extraction rules specific for each class of instances it wants to extract from the Web. A query is then created from keywords in each rule, a Web search engine is queried and the rule is applied to extract information from the Web pages retrieved. The likelihood of each candidate facts is later assessed with a kind of PMI between words and phrases, estimated from the search engine hit counts in a manner similar the PMI-IR algorithm (Turney (2001)). PMI is computed between each extracted instance (I) and automatically generated discriminator phrases (D) associated with the class:

$$PMI(I, D) = \frac{|Hits(D + I)|}{|Hits(I)|}$$

Still concerning the automatic extraction of relations from text, Banko et al. (2007) propose a new paradigm where the system makes a single data-driven pass over a corpus and extracts a large set of relational tuples, without requiring any human input. TextRunner, a fully-implemented system that follows this paradigm, is presented together with some experiments where it is compared with KnowItAll.

In TextRunner, a small corpus sample is given as input in order to get a classifier that labels candidate extractions as trustworthy or not. 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. Experiments show that TextRunner is more scalable, has a lower error rate and, considering only a set of 10 relation types, both systems extract an identical number of correct relations. However, since TextRunner does not take as input the name of the relations, its complete set of extractions contains more types of relations.

The output triples can be used to create an ontology (Soderland and Mandhani (2007)), with WordNet serving as a map of concepts. Furthermore, the relation phrases are normalised and mapped to one of the relations from a predefined set. Then the logical semantics is formalised, the meta-properties of each relation are learned, a correctness probability is given to each relation and, finally, an inference engine combines the derived relations with existing knowledge.

In order to label non-taxonomic, and otherwise anonymous, relations learned between two concepts, Kavalec and Svatek (2005) propose a technique based on the assumption that relational information is, at sentence level, typically conveyed by verbs. Verbs (or verb phrases) frequently co-occurring with each two related concepts are selected and the *concept-verb-concept* triples are ordered by a numerical measure. The top verbs are the candidates for relation labels for the given pair of concepts.

Based on co-occurrence, the ASIUM system (Faure and Nédellec (1998)) builds both a concept and also a verb sub-categorisation frame hierarchy. While the former can be viewed as a simple taxonomy, the latter can be used to cluster the concepts. Yet, this information about co-occurring verbs is not translated into relation names.

3.2.4 Discussion

Work on the various steps needed for ontology learning from text was presented in this section. While for MRDs processing most of the methods are linguistic, the picture is quite different for corpora processing, where the attempts to extract and organise information can be divided into three categories, according to the methods used:

- Linguistic: based on the identification of specific textual patterns and linguistic constructions;
- Statistical: mostly based on frequencies and co-occurrence of tokens;
- Hybrid: where statistical techniques are combined with linguistic approaches.

The first step of ontology learning from text, where the utilisation of statistical methods is more frequent is the first, more precisely, associating terms. Nevertheless, most of the times where such a method is applied without any kind of linguistic knowledge the only information obtained is nothing more that some unlabelled clusters of similar or related terms, lacking interpretation.

On the other hand, it is very difficult to define a finite set of linguistic patterns capable of acquiring all the instances of some relation in corpora, even if the text is syntactically annotated. This is a consequence of using unrestricted text, where there are many possibilities to say exactly the same thing. At the same time, especially when corpora is not domain-specific, there are no boundaries to the vocabulary to be used, which increases ambiguity to a much higher level than in dictionaries. There are many issues that lead to the higher complexity of corpora text, such as:

- Many nouns and verbs are modified, respectively by adjectives and adverbs;
- The use of anaphoras, to refer to entities previously referred in the same text;
- The use of figures of speech, as metaphor where some term is used to refer to a different one, having thus a different meaning.

Combining linguistic and statistical methods, as in many works where relations are extracted (e.g. Caraballo (1999) Cederberg and Widdows (2003)), can sometimes be the best way to deal with these limitations.

Another difference between linguistic and statistical methods is that the latter are language independent. In linguistic methods, the patterns involved are written in some language and need to be changed in order to adapt them to other languages.

It should be added that while most statistical techniques are language independent, linguistic techniques are mostly language dependent. This happens because the patterns involved in linguistic methods are built upon words of that language or, at least, linguistic constructions specific of that language. It is thus needed to change most of these patterns, in order to adapt them to other languages.

3.3 Evaluation of Ontologies

An ontology is a complex structure and not a plain list of classifications, which per se suggests that its evaluation is not straightforward. Furthermore, there is no clear set of knowledge to be acquired (Brewster et al. (2004)). Smith (2004) states that different groups of people from different disciplines have different goals and needs concerning an ontology. They will therefore assess an ontology differently.

In this section, the discussion is focused on ontology evaluation attempts. Regarding their differences and also the focus of this proposal, the evaluation of domain ontologies is addressed separately from the evaluation of lexical ontologies. As we will see, not all evaluation approaches for domain ontologies can be adapted directly to lexical ontologies.

3.3.1 Evaluation of Domain Ontologies

Concerning domain ontologies, Brank et al. (2005) divide evaluation approaches into four groups:

- Manual evaluation, performed by human subjects;
- Comparison with a *golden standard*;
- As for coverage, comparison with a collection of documents about a domain covered by the ontology;
- Task-based evaluation.

Manual evaluation is the most traditional type of evaluation. Due to its eventual complexity, it is sometimes easier to transmit roughly to human judges the aim of the target ontology and the principles that should be considered in its evaluation, rather than encoding a system that automatically evaluates the resource according to those principles. Although, in the end, manual evaluation is the most reliable, it does not take advantage of computer programs and hence cannot be done automatically. It relies heavily on time consuming work from domain specialists which makes it hardly repeatable.

A golden standard is some resource, eventually another ontology, that we know for sure is correct, possibly because it was manually created by specialists. An ontology can be compared to a golden standard according to some criteria in order to assess its quality. In this context, two common measures in IR, precision and recall, are typically and increasingly being applied (Santos (2007b)):

$$Precision = \frac{Correct\,answers}{Given\,answers}$$

$$Recall = \frac{Correct\,answers}{Possible\,answers}$$

Here, an answer can be defined as different things, such as associated terms or existing relationships. These two measures can be further combined giving rise to the so called f-measure.

One limitation is that there are not many golden standards that can be used in ontology evaluation, not only because the creation of machine readable ontologies

is quite recent (Biemann (2005)), but also because of the specific characteristics each ontology has. Most of the times it is necessary to gather a group of knowledge experts to manually build the resource that will serve as a golden standard.

Maedche and Staab (2002) propose an evaluation method that assesses the quality of an ontology by comparing its vocabulary and structure with ontologies created by non-expert subjects and also a golden standard ontology, modelled by a specialist.

Considering its coverage of a domain, an ontology can also be evaluated by measuring how adequate it is for representing the knowledge contained in a collection of available data, usually textual documents on some domain. Brewster et al. (2004) measure the fit between an ontology and a corpus after identifying salient terms in a corpus, and looking for them in an ontology of the same domain. The fit is then proportional to the number of terms found in both corpus and ontology.

Ontologies are commonly used together with applications for the achievement of some specific task. Assuming that the quality of the results of such tasks will depend on the quality of the ontology, the latter can be evaluated indirectly, with conclusions taken from the results of some task performed using the ontology. As one can see, this kind of evaluation has several limitations, starting with the difficulties concerning the generalisation of the results, since the ontology is only being used in one specific task. Furthermore, the results can be difficult to interpret, as the effect of the ontology in the task can be insignificant or indirect. Also, its is only possible to compare different ontologies when they can be both used to accomplish the same task.

Nevertheless, Porzel and Malaka (2004) describe an application for speech recognition, where an ontology is primarily used to find the similarity between the meaning of two concepts. The output of the application can be assessed by comparing the interpretation of sentences produced with a golden standard provided by humans, and used to evaluate the ontology. Their ontology is evaluated with respect to the fit of the vocabulary, the fit of the taxonomy and the adequacy of non-taxonomic semantic relations.

As discussed in the previous sections, an ontology can be a very complex structure, which reflects on the process of its creation where, rather than approaching the ontology as a whole, different levels are focused in different construction phases (as discussed in the previous sections). This should also be considered when attempting to evaluate an ontology, so Brank et al. (2005) proposes the following evaluation levels. It should be stressed that these levels are just an example and are not strict.

- *Lexical, vocabulary, or data layer*: evaluation of the terms, concepts, instances, facts, and others, represented in the ontology, and also the vocabulary used to identify these items.

- *Hierarchy or taxonomy*: assessment of the coherence between the classes and subclasses represented.

- *Other semantic relations*: assessment the quality and adequateness of non-taxonomic relations.

- *Context or application level*: determine how well the ontology suits its contexts, how well it interacts with ontologies in the same context and how well its serves the application it is used in.

- *Syntactic level*: validation of the well-formedness of the syntax used to describe the ontology.

- *Structure, architecture, design*: evaluate whether the ontology meets pre-established design-principles or criteria, and if the structure and organisation of the ontology suits its purpose.

3.3.2 Evaluation of Lexical Ontologies

It seems that the evaluation of dictionaries and handcrafted lexical ontologies on general principles is not a common practice (Raman and Bhattacharyya (2008)). One reason for this to happen is that these kind of resources are created manually, by specialists, and are thus thought of as not prone to errors.

Ide and Veronis (1995) are very critical of this fact and produced work for assessing the quality and usefulness of information extracted from MRDs. After performing a quantitative evaluation of automatically extracted hypernymy relations (described in Section 3.1), they concluded that the structures obtained by applying *Chodorow-like* (Chodorow et al. (1985)) procedures were incomplete and had several other problems but, if they merged the results extracted from several MRDs, the amount of problems decreased drastically.

Despite the lack of evaluation of handcrafted ontologies, we are aware of several independent attempts. For instance, Raman and Bhattacharyya (2008) performed an automatic evaluation of the Princeton Wordnet (Fellbaum (1998)) synsets. In their work, the presence of each word in a synset is validated with the help of the definitions of a dictionary. Their main assumption is that, if a word truly belongs to a synset, there should be a dictionary definition for it, which refers to the hypernym or to other words in the synset. When the presence of a word in a synset is not validated, it is flagged in order to be verified by humans. Since Wordnet was manually created by experts, the authors did not expect a great amount of errors. Among the words found in the dictionary, but not validated, many have rare meanings and usages.

Also concerning the evaluation of a lexical ontology created by hand, Mahesh et al. (1996) report on an empirical assessment exercise of Cyc (Lenat and Guha (1989)) for NLP applications, more precisely they assessed its utility for WSD and coreference resolution problems.

As for the evaluation of lexical ontologies created automatically, methods can be inspired in those for evaluating domain ontologies. However attention should be paid, because lexical ontologies have clearly different characteristics from domain ontologies. While the latter cover only a specific and close domain, the former describe the conceptual model of a whole language.

Again, manual evaluation plays an important role for evaluating lexical information extracted by automatic means. This can be noticed by the amount of works referred in Section 3.2 where evaluation was performed by human judges (e.g. Chodorow et al. (1985), Riloff and Shepherd (1997); Roark and Charniak (1998); Caraballo (1999); Pantel and Ravichandran (2004)).

Additionally, Richardson et al. (1993) hand-checked a random sample of 250 semantic relations automatically extracted from a dictionary, later included in MindNet (Richardson et al. (1998)), and found them to be 78% correct overall. They rely

on common statistical techniques to estimate that this accuracy is representative for all the relations extracted, with an error of +/-5%. However they note that about half of the sample consisted of hypernymy relations, which had 87% of accuracy overall. Also for MindNet, Vanderwende et al. (2005) refer an (incomplete) evaluation of the quality of the semantic relations, but they do not go very far on the description of the evaluation process. One comment they make is that the quality varies according to the relation type.

In order to make human evaluation easier, Navigli et al. (2004) generated natural language descriptions of concepts, based on a grammar with distinct generation rules for each type of semantic relation.

Evaluation using a golden standard, created by human subjects, is typically adopted in joint evaluations such as ACE (Doddington et al. (2004)), SemEval (Agirre et al. (2007)) or, for Portuguese, HAREM (Santos and Cardoso (2007); Mota and Santos (2008))), however these standards only encompass a few examples. For instance, in the HAREM track ReRelEM (Freitas et al. (2009)), users had to discover semantic relations between named entities in a collection of texts. However, named entities where only annotated in a small part of the main collection, the HAREM golden collection, and semantic relations were only marked for a small portion of the latter collection, the ReRelEM golden collection. Consequently, despite the fact that all the texts in the main collection were annotated by the systems, only the texts in the ReRelEM collection were used for evaluation purposes.

One problem is that, for semantic lexicons, it is hard to have an independent golden standard for what should be there in the first place. The knowledge that should be represented is not clear and if we compare it with semantic data extracted from text, we have to remember that different interpretations and different meanings are often possible (Brewster and Wilks (2004)).

Furthermore, several works concerning the extraction of lexical information in English used WordNet as a golden resource to achieve their evaluation (e.g. Hearst (1992), Pantel and Lin (2002), Nichols et al. (2005)) More important than achieving the evaluation of their results, it was noticed that some of the relations extracted were clearly correct, but not present in WordNet, which suggests that WordNet has several gaps and is thus incomplete.

In an alternative evaluation approach, Etzioni et al. (2005) translated their hypernymy relations into natural language patterns and searched for them in the web to evaluate whether a named entity was an instance of a specific class or not. If it was not for the translation process, it could almost be seen as using the web as a golden resource.

As opposing to domain ontologies, we cannot define a clear set of salient terms for general language, which invalidates the application of Brewster et al. (2004)'s measure for lexical ontologies. However, all words of a language could eventually be used. Demetriou and Atwell (2001) refer that the coverage of a lexicon can be measured by looking for words in a corpus that are not found in the lexicon and using one of two base numbers:

- different word forms ("vocabulary type" coverage), which answers to the number of different word types in language, that are covered by the lexicon.

- total words in text ("real text token" coverage), which answers to the number of word tokens that are expected to be covered by the lexicon.

Vocabulary type coverage usually gets a lower coverage proportion when the text is large as opposed to real text token coverage.

An example of indirect evaluation that suits lexical ontologies is given by Cuadros and Rigau (2006), who discuss the evaluation of knowledge resources in the context of a WSD task. Various lexical ontologies, including WordNet and several lexical resources created semi-automatically, were used in this study. Curiously, it was empirically demonstrated that resources that had been created by automatic means surpass the handcrafted ones, both in terms of precision and recall. Also, combining the knowledge contained in the resources studied is very close to selecting the most frequent sense of a word.

It should also be referred that the similarity inference procedure of MindNet was evaluated and is reported by Richardson (1997). This was achieved by calculating the similarity of pairs of nouns and verbs that were associated in a thesaurus, and were thus similar. Among these pairs, 81% had a score that corresponded to similar words.

Chapter 4

Approach

In this chapter, an approach is proposed in order to achieve our goals. Some of the points discussed are just ideas and alternatives that might not be completely clear, but can be viewed as a starting point for further work, comprised by each one of the phases described.

To build Onto.PT, Portuguese textual resources will be exploited, in order to extract lexico-semantic knowledge. More than acquiring terms, we aim to extract several lexico-semantic relations, having in mind that those relations can be expressed by known textual patterns, as referred in Sections 3.1 and 3.2.

It should be stressed that this work will not be bounded by the acquisition of the most typical relations, like synonymy and hypernymy, but it is also its goal to extract other interesting relations such as causation, purpose or manner. As an example, some of the patterns we intend to exploit and the relations they are associated with are shown in Table 4.1.

Relation	Example pattern
Hypernymy	`tipo|género|classe|forma de`
Meronymy	`parte|membro de`
Causation	`causado|provocado|originado por`
Purpose	`usado|utilizado para`

Table 4.1: Examples of patterns indicating semantic relations.

This work will involve the exploration of two kinds of textual resources: starting with MRDs, for obtaining more general knowledge, and then moving on to textual corpora, in order to enrich the ontology with more specific knowledge. However, the methods for extracting information from these kinds of text are quite different, due to the reasons mentioned in Sections 3.1 and 3.2: dictionaries are structured around words and their meanings and typically use simple, restricted and almost predictable vocabulary, while in generic textual corpora the situation is completely different. This is why we believe it is possible to extract useful lexico-semantic information from MRDs with hand-made semantic grammars (Brown and Burton (1975)) that capture indicating patterns. Nevertheless, we have some doubts to what concerns corpora processing, where we will analyse the possibilities of taking advantage of morpho-syntactic information and combine linguistic approaches with statistical text mining techniques.

It should be added that, in opposition to what is followed in most software development projects, the specification of a structure for Onto.PT will only come in a second phase. Since the author of this proposal was one of the developers of PAPEL (Gonçalo Oliveira et al. (2008, 2009a,b)), it seems more natural to start by applying the methodology used for PAPEL's creation, in order to obtain the core elements of Onto.PT – generic terms and relations established between them – and then try to fit these elements in an adequate model. Therefore, this section starts by presenting PAPEL's methodology, and then makes a brief overview on the further phases of the approach to be followed.

4.1 Starting Point: PAPEL

The approach for building PAPEL comprises four stages, briefly described here. It can be said to be a semi-automatic approach, because while stages 2 (the core stage) and 4 are completely automatic, stages 1 and 3 are completely manual or partially manual, respectively.

1. **Creation of the extraction grammars**

   Inspired by Alshawi (1989), semantic grammars are created specifically to parse the dictionary definitions. The grammars, which include textual patterns similar to the ones in Table 4.1, aim at the extraction of specifically predefined relations (see Table 2.2, for the predefined relations in PAPEL) and are based on a previous empirical analysis of the structure of the definitions and of the vocabulary used in the dictionary to be exploited. Each grammar is made to process definitions of words belonging to only one of the four open grammatical categories (nouns, verbs, adjectives, adverbs) and the name of the relation is given based on the grammatical category of its arguments.

2. **Relation extraction proper**

   In this stage, a chart parser, PEN[1], processes the definitions according to the grammars and, if the definition suits the rules, a derivation tree is generated. For each grammar, the extraction tool selects the better tree and outputs eventual relations (identified by the labels of the tree nodes) between words in the definition and the defined word. Two example definitions together with their best derivations according to two different grammars (one for the extraction of hypernymy relations and other for the extraction of meronymy) and the relations extracted from the obtained derivations are presented in Figures 4.1 and 4.2.

3. **Manual result inspection**

   The extraction results are inspected in order to identify systematic problems, and with the two previous stages form a loop that can be repeated at will.

   Results from different extraction runs can be automatically compared, using a regression system, to guarantee that newer results are better than older ones. After this procedure, it is possible to go back to the first stage, in which newer

[1]Freely available for download, under a BSD license, from `http://code.google.com/p/pen/`

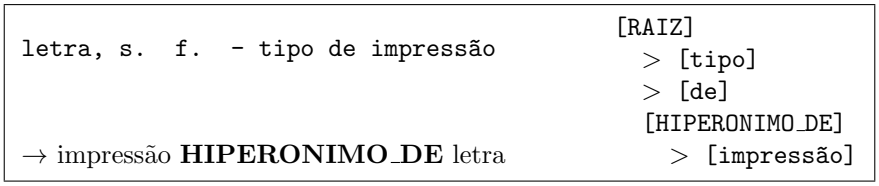

Figure 4.1: Derivation for one definition of *letra*.

versions of the grammars are created, hopefully with some of the identified problems corrected.

4. **Relations adjustment**

   After several loops of processing, a new stage is entered, where the relations with inadequate arguments (i.e. arguments whose grammatical category does not agree with the relation name) are either corrected or discarded. In order to simplify the relation set, all relations are first translated into the type defined as direct. This stipulation is made based on what seemed more natural to the grammar writer, and not on frequency considerations. For example, after this stage in PAPEL, the relation *manga* INCLUI *punho* is automatically translated to *punho* PARTE_DE *manga* and *dor* RESULTADO_DE *distensão* becomes *distensão* CAUSADOR_DE *dor*.

   Then grammatical category of each argument is verified, with the help of the grammatical information in the dictionary and, when the argument is not defined in the dictionary, with the help of the Jspell (Simões and Almeida (2002)), a morphological analyser for Portuguese. If the arguments of a relation are not adequate but there is a relation type that belongs to the same group and suits the categories of the arguments, the relation type is replaced, otherwise the relation is discarded. For example, once again in PAPEL, the relation *loucura* ACCAO_QUE_CAUSA *desvario* becomes *loucura* CAUSADOR_DE *desvario*, because both arguments are nouns. Additionaly, during this verification, if an argument is not in the lemma form, it is automatically changed to it, again with the help of Jspell.

4.2 Extraction of Knowledge from MRDs

In this first phase, the structure and relations of PAPEL will be analysed in order to improve the grammars, the extraction tools and thus the quality of the relations extracted. Then, since we already have the tools and a methodology to extract lexico-semantic relations from MRDs, we are planning to adapt it to other MRDs and later merge the obtained results with PAPEL, in order to refine our results and take some conclusions, as it is suggested by Ide and Véronis (1993).

There are at least two Portuguese MRDs in the public domain, where we believe we can apply the PAPEL tools, namely:

- The Portuguese version of Wiktionary[2], a collaborative dictionary maintained by the Wikimedia Foundation

[2] http://pt.wiktionary.org/

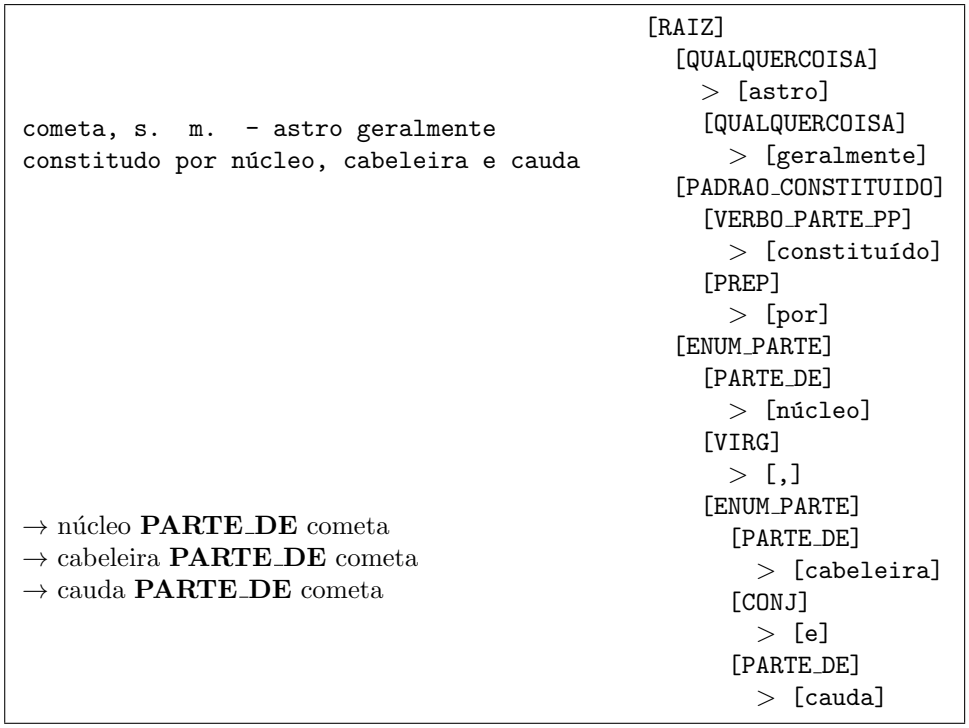

Figure 4.2: Derivation for the definition of *cometa*.

- *Dicionário Aberto*[3], a project supported by the Portuguese National Library[4], the weblog *Página a Página*[5] and the *Natura Project*[6] in University of Minho.

4.3 Resource Structure Specification

Decisions about the resource structure will be made after having the relation set extracted from MRDs. A suitable architecture will be modelled concerning the organisation of the terms and relations obtained in the previous phase. We will start by designing a procedure to merge adequately the results obtained from all the MRDs.

A key decision in this phase will be how to handle polysemy and homonymy (see Section 2.2.1). Considering the structure of the resource, we have two possible ideas in mind. The first one would be to adopt a "wordnet-like" structure, where synonym words are included in the same synset, and the relations occur between synsets. To achieve this kind of structure, WSD techniques would be needed to identify possible different senses of a word. We do not expect this to be easy, due to the lack of consensus concerning WSD (Kilgarriff (1997)) and its dependence on the purpose (Wilks (2000)). There are however some ideas on how to get useful hints to accomplish WSD:

- the sense division in the processed dictionaries;

[3] http://www.dicionario-aberto.net/
[4] http://www.bnportugal.pt/
[5] http://pagina-a-pagina.blogspot.com/
[6] http://natura.di.uminho.pt/wiki/doku.php

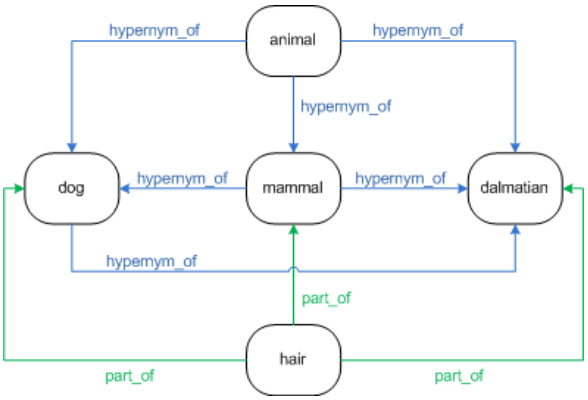

Figure 4.3: Expanded semantic network.

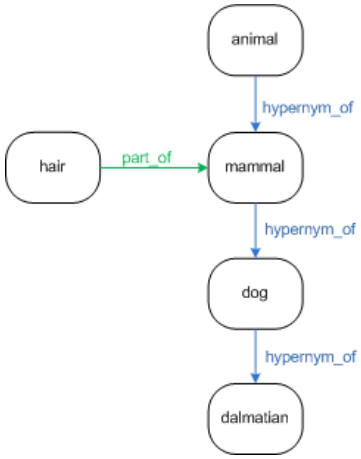

Figure 4.4: Reduced semantic network.

- sentences where the words occur, preferably sentences where we know what sense is being used (e.g. example sentences in Tep (Maziero et al. (2008)), when available);

- exploitation of the (ambiguous) network structure.

Another possibility is to create a structure similar to MindNet (referred in Richardson (1997); Richardson et al. (1998)), where the sense division in the MRDs is used to define the various senses a word can have. The main structure would then be the word, which would contain its possible grammatical categories and senses. Relations would occur between a word sense and a word structure. This approach would require fairly less WSD and would eventually have to deal with word sense ambiguation (Dolan (1994)), in order to group related word senses and simplify the network.

After choosing the right structure to represent the semantic network, some automatic procedures to organise the relations should be developed. For instance, there should be a set of rules for reducing the number of relations without losing information. This can be done by eliminating redundant relations, as it is done to turn the network in Figure 4.3 into the network in Figure 4.4. On the other hand, information should not be lost so, it should be possible to infer all implicit relations, whether they are redundant or not. Looking at the example, it should be possible to infer all the relations in Figure 4.3 from Figure 4.4. To sum it up, the network structure should be as simple as possible, without redundant information, but at the same time it should be possible to calculate information that is not explicit.

The inference mechanism can be done in a similar fashion to what was done in ReRelEM (Freitas et al. (2009)), a task included in the Second HAREM (for a brief description see Santos et al. (2008), and for a full description see the book Mota and Santos (2008)) joint evaluation to identify and classify Portuguese named entities. In ReRelEM, the participants were challenged to develop systems capable of identifying semantic relations among named entities in a collection of Portuguese texts. Neither the golden collection nor the participation files needed to have the full set of relations found annotated, but at least a minimum set that could be used to infer all the relations, after applying the following set of rules:

- A `ident`[7] B ∧ B `ident` C → A `ident` C
- A `inclui`[8] B ∧ B `inclui` C → A `inclui` C
- A `inclui` B ∧ B `sede_de`[9]. C → A `sede_de` C
- A `ident` B ∧ B *any_relation* C → A *any_relation* C

Besides ReRelEM, Freitas (2007) also discusses inference for hypernym relations (referred in Section 3.2.2), while Richardson et al. (1998) refer an inference procedure for MindNet and give an example based on the assumption that hyponyms inherit their hypernyms properties: starting with the relations *watch* HYPERNYM *observe* and *observe* MEANS *telescope*, it can be inferred, for instance, that *watch* MEANS *telescope*.

4.4 Extraction of Knowledge from Textual Corpora

As it is argued by several authors (e.g. Hearst (1992)), in order to find terms and expressions that are not defined in MRDs, we must turn to other textual resources, like textual corpora, that should be viewed as the main source of domain-specific knowledge (Brewster and Wilks (2004)). So, in this stage we will develop tools to extract relations from corpora and use them either to enrich the main ontology in specific domains or to create new domain ontologies based on the texts given as input.

One of the issues about information extraction from corpora, as opposing to dictionary processing, is that the text is generally not structured and it is much more difficult to predict the vocabulary and obtain a fair amount of patterns that indicate each relation. In a dictionary the effort needed to extract terms is reduced as each defined word is a candidate term. On the other hand, in corpora text, the terms are dispersed among free text. Additionally, it is common to find many modifiers and anaphoras, which increase the processing complexity, and also figures of speech, which increase the ambiguity. The latter reasons lead to alternative approaches, either completely based on statistical methods or hybrid, where statistics is combined with more linguistic approaches.

We are thus convinced that a blind adaptation of the same methodology used for the extraction from MRDs will not perform as good as if we combine it with other approaches. Therefore, in order to come up with an extraction methodology, it is our intention to adapt some of the methodologies described in Section 3.2 to Portuguese and take some conclusions on how suited they are for our purpose. The work of Hearst (1992) has already been adapted to Portuguese (for instance by Freitas (2007)), and we believe that it would be interesting to adapt other methods for associating terms (e.g. Roark and Charniak (1998); Pantel and Lin (2002)) and for relation extraction (e.g. Cederberg and Widdows (2003); Pantel and Ravichandran (2004)).

[7] Identity relation, the same as synonymy.
[8] Inclusion relation, a kind of meronymy.
[9] Location relation

On the linguistic side of the methodology to be developed, we believe that having the text syntactically annotated could decrease the processing difficulties, since we would be able to use structural patterns (Montemagni and Vanderwende (1992)). In order to have syntactically-annotated text, there are three options:

- Using syntactically-annotated corpora;
- Using a broad-coverage parser for Portuguese;
- Using a broad-coverage parser that can be trained for Portuguese;

Using syntactically-annotated corpora is however limited to the existing Portuguese annotated corpora, for instance CETEMPúblico (Rocha and Santos (2000); Santos and Rocha (2001)). As for using a broad-coverage parser for Portuguese, PALAVRAS (Bick (2000)) would be a good option if it were free software. Consequently, we will investigate other options of broad-coverage parsers for Portuguese or of parsers that can be trained for this language.

Besides its importance to acquire domain specific knowledge, using corpora will also be useful to find out new patterns indicating relations, not only hypernymy, but also for other relations. Therefore, a pattern discovery method, similar to the one proposed by Hearst (1992), will be applied.

4.5 Resources Evaluation

This stage has the goal to evaluate the results produced by the previous stages, namely the correctness of the represented relations and the coverage of its concepts considering several domains. The utility of the tools and of the created ontology for accomplishing several NLP tasks will also be evaluated.

Looking at the evaluation methodologies discussed in Section 3.3, we will try to avoid manual evaluation, because of the aforementioned reasons – intensive and time consuming human labour, where specialists are usually needed, which leads to difficulties for its repetition. We believe however that, in the end, due to its reliability, manual evaluation of a small but representative part of the results should be performed. An idea to perform some kind of cooperative manual evaluation would be to develop an interactive game (von Ahn (2006)) where the users would be, at the same time, playing and evaluating our results. In this kind of evaluation, the resource could be evaluated by the community, whether they are specialists or not. However, we do not believe it to be a problem because, having in mind that the resource is made to suit the communitys needs, it should also be evaluated by the community itself.

Still concerning manual evaluation, we will study the usage of an adequate scale to rank each relation according to a given level of quality or confidence. For this purpose, Freitas (2007) used the following scale:

Rank	Reason	Given example
3	The relation is correct, exactly how it was extracted	*suco* hypernym_of *bebidas*
2	The relation is almost correct, but there are prepositions, adjectives or others that make it quite strange	*psicólogos* hypernym_of *agentes da equipe*
1	The relation might be correct but is very general or very specific to be useful	*protecção* hypernym_of *valores*
0	The relation is incorrect	*soco* hypernym_of *traumas*

Manually created lexical ontologies for Portuguese would be the strongest candidates for playing the role of golden resources, in an evaluation based on the latter, where the measures of precision and recall can calculated. However, as referred in Section 2.2.2, all the eventual candidates, but Tep (Maziero et al. (2008)), are still not available or not in the public domain. The problem of Tep is that it contains only synonyms and antonyms and can not be used as a golden resource for other kinds of relations. Nevertheless, if the expected output, given a set of texts, were manually created (as it is typically done in joint evaluations), these texts could be used as a golden resource to evaluate our extraction tools, and measure their precision and recall. An interesting way of accomplishing the evaluation of our tools would be to participate in a joint evaluation for Portuguese, similar to ReRelEM (Freitas et al. (2009)), where our tools would be used to detect relations in text, annotate them and participate side-by-side to other systems with similar purposes. We are, unfortunately, not aware of any scheduled campaign of this kind.

Vocabulary coverage can be evaluated using Demetriou and Atwell (2001)'s idea, referred in Section 3.3.2, where words in a corpus are matched to the words in the ontology. It is also possible to evaluate the coverage of some domain, following Brewster et al. (2004)'s fit measure, where salient terms found in domain corpora are looked for in the ontology.

Another interesting way to evaluate the relational triples would be to perform something similar to what was done in PAPEL (Gonçalo Oliveira et al. (2009a,b)), which was then inspired by Etzioni et al. (2004). This evaluation methodology, whose results for PAPEL are described in Gonçalo Oliveira et al. (2009a) (in Portuguese) and Gonçalo Oliveira et al. (2009b) (in English), consists of rendering the relational triples to natural language patterns and look for them in corpora. If at least one pattern is found, the relation is supported. For instance, the following relation

cólera CAUSADOR_DE *diarreia*

Can be validated by searching for a set of patterns, including the following:

- ... `cólera causa|provoca|origina diarreia` ...
- ... `diarreia causada|provocada|originada por cólera` ...
- ... `diarreia devido a cólera` ...
- ... `diarreia resultado de|da cólera` ...

This can almost be seen as a process of reverse engineering.

In the evaluation of Onto.PT, different available corpora can be used or even the whole Web, with the help of a search engine. However, if syntactically annotated corpora is used, it will be easier include more text variations in more simple patterns. Furthermore, in addition to what was done in PAPEL, the number of patterns found for each relation, eventually combined with the frequency of the terms involved, can be used as a confidence indicator for that relation.

We are also planning to evaluate the ontology by using it in one or several applications to perform NLP tasks. A Q&A system based on an ontology, a knowledge extraction system and a creative text generator are among the applications being developed in our research group which are in need of a lexical ontology. We believe that if Onto.PT, or the construction tools, are integrated with some of these applications, their behaviour will be a good indicator to support the utility of our work.

4.6 Resources Deployment and Advertisement

After reaching an adequate level of quality we intend to make the ontology and tools publicly available, together with user documentation. While the main ontology will have the purpose of being integrated with applications that use lexico-semantic knowledge in Portuguese, the extraction tools might be of great utility for information extraction systems. In order to announce the resulting resources for the community, emails will be sent to discussion mailing lists about NLP, such as:

- Corpora List (`corpora@uib.no`), open list for information and questions about text corpora such as availability, aspects of compiling and using corpora, software, tagging, parsing, bibliography, conferences etc;
- Linguist List (`linguistlinguistlist.org`), dedicated to providing information on language and language analysis, and to providing the discipline of linguistics with the infrastructure necessary to function in the digital world;
- Forum-LP (`forum-lp@di.fct.unl.pt`), dedicated to the automatic processing of Portuguese.

It is also our intention to develop some kind of browsing tool, eventually web-based, to ease navigation through the resulting ontology, in a similar fashion to what Princeton WordNet and MindNet both have[10]. Since this tool might be very useful for a quick overview of the ontology and also for debugging, the development of a prototype will be devised some phases earlier. However, a final version is only expected together with the deployment of the final resources.

To extend the potential utilisation scenarios we are devising to export the resource to several data representation formats. For example, there are many Semantic Web (Berners-Lee et al. (2001)) applications based on RDF/OWL (Miller and Manola (2004); McGuinness and van Harmelen (2004)) models, because these are the W3C standard description languages for the Semantic Web. Additionally, these

[10]See respectivelly Wordnet Search in `http://wordnetweb.princeton.edu/perl/webwn` and MNEX in `http://stratus.research.microsoft.com/mnex/InputPath.aspx?l=e&d=d`

languages ease the browsing and visualisation of ontologies and have other useful features like the possibility of creating rules for inference of new relations and reasoning, so there is a strong possibility of developing a RDF/OWL representation of Onto.PT.

Chapter 5

Work plan

In September 2008, while still working for Linguateca (see Santos (2000), Veiga and Santos (2001), Santos (2002), Santos et al. (2004) and Santos (2009) for different snapshots of this project), the author of this proposal enrolled in the Doctoral Program in Information Science and Technology of the Department of Informatics Engineering of the Faculty of Sciences and Technology of the University of Coimbra. During the first year, a set of four courses were completed, namely Research Methods; Advanced Topics in Artificial Intelligence; Art, Music and Creativity and Artificial Intelligence; and Ludic Contexts.

This section is about the work plan for the PhD program, which is scheduled to last between 3 and 4 years. It starts by referring some work made during the last year (Section 5.1), and then exposes a tentative plan for the further to be accomplished until the end of this research (Section 5.2). Finally, some sites we believe to be important to have our work published are presented (Section 5.2).

5.1 Current Work

After finishing the contract with Linguateca in January 2009, a review on background concepts and state of the art work was made with the purpose of starting the research described in this proposal and writing the proposal itself.

Furthermore, work has been done in analysing the results of PAPEL (Gonçalo Oliveira et al. (2008, 2009a,b)) in order to find out problems that can be corrected, both in the grammars and in the relation adjustment, to improve the quality of the relations. In order to validate the relations semi-automatically, a testing system was developed. Synonymy relations are evaluated using the thesaurus Tep (Maziero et al. (2008)) as a golden resource. As for other relations, they are rendered into natural language and the obtained patterns are searched in textual corpora (more precisely CETEMPúblico (Rocha and Santos (2000); Santos and Rocha (2001))), in a similar fashion to what Etzioni et al. (2005) have done to evaluate their hyponymy relations using the Web. The number of patterns found in the corpus gives us an idea of the quality of the relations.

Additionally, shifting to the Semantic Web, PAPEL was converted into a simple OWL model, and then a graphical interface was developed to help us visualise and browse through OWL networks, VisuOWL[1]. This tool (see Figure 5.1) has revealed

[1]Available for download through `http://code.google.com/p/visuowl/`

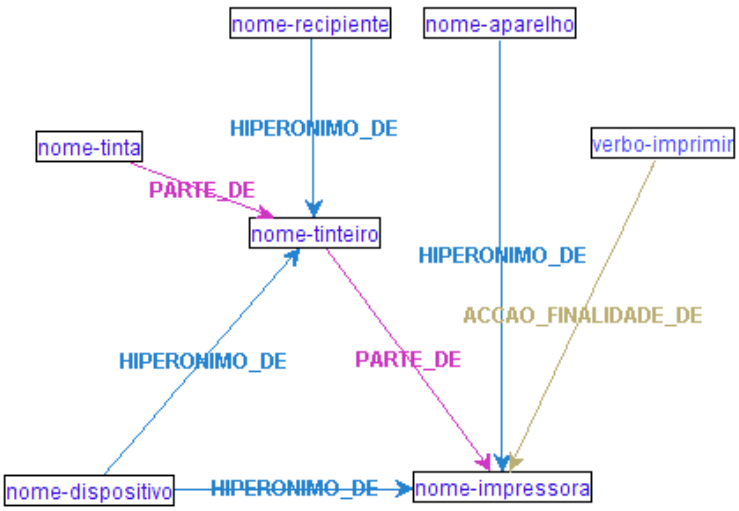

Figure 5.1: The VisuOWL tool, showing the words related to *impressora*.

to be very useful for getting a clearer idea about the results of PAPEL and also for debugging.

5.1.1 Publications

The last improvements to PAPEL and its evaluation originated two publications produced during the last year, namely Gonçalo Oliveira et al. (2009a) and Gonçalo Oliveira et al. (2009b). Besides the latter publications, a paper about the research described in this proposal was written to be presented in Doctoral Symposium on Artificial Intelligence (SDIA) (Gonçalo Oliveira (2009)), a satellite event of the Portuguese Conference on Artificial Intelligence (EPIA 2009).

5.2 Further Plan

For the next 2 or 3 years we are planning to go through the phases referred in Section 4, almost in a sequential way. Figure 5.2 is the result of a scheduling exercise made for this research. The most important points covered by each phase are reminded in the following list:

1. Extraction from MRDs
    - Improvements to PAPEL
    - Adaptation of the methodology to other MRDs
    - Merge the results from different MRDs
2. Experimentation and evaluation 1
    - Assess the advantages and drawbacks of merging results from several MRDs
    - Semi-automatic evaluation using corpora
    - Development of a browser prototype

3. Resource specification
    - Definition of Onto.PT's structure
    - Development of organisation procedures
4. Learning from corpora
    - Adaptation of term extraction techniques
    - Adaptation of relation extraction techniques
    - Design of the methodology
    - Application of the methodology
    - Integration of the results in Onto.PT
5. Experimentation and evaluation 2
    - Automatic population of the ontology from text
    - Assess the advantages and drawbacks of integrating the results extracted from corpora with Onto.PT
    - Assess the coverage of several domains
    - Development of a game for evaluation by the community
    - Manual evaluation of a subset
    - Integration with other application(s)
6. Deployment and advertising
    - Conversion to different formats
    - Browser
    - Final documentation and packaging
    - Public Announcement
7. Writing of the PhD thesis

In each phase, written documentation will be produced in order to describe information that is important for the understanding of this research (e.g. decisions made, techniques used, resource architecture, implemented algorithms, results obtained). While detailed technical information will be included in technical reports, methodologies developed, important results achieved and conclusions taken shall give rise to scientific papers and articles, to be published in conference proceedings or scientific journals. Some of the target publication sites concerning this research are presented in next Section.

In an initial phase, we aim to publish in specialised workshops, so that we can obtain more specific feedback from a more specialised audience. After the work is stable and interesting results are achieved, the main goal will be to publish in the main conferences and journals. In a later phase, a PhD thesis, encompassing all the dealt research topics, will be written.

5.3 Target publication sites

Here, several target publication sites, comprising the two main topics of this research, namely general AI and NLP, are presented. Since some of the work to be done in the latter phases of this research might involve the Semantic Web languages and technologies, the most important conferences on this topic are also referred.

We start by introducing some well-known conferences on general AI:

- *International Joint Conference on Artificial Intelligence* (IJCAI), the main international gathering of researchers in AI.
- *European Conference on Artificial Intelligence* (ECAI), the leading Conference on Artificial Intelligence in Europe.
- *Association for the Advancement of Artificial Intelligence National Conference* (AAAI), whose purpose is to promote research in AI and scientific exchange among AI researchers, practitioners, scientists, and engineers in related disciplines.
- *Conferencia de la Asociación Española para la Inteligencia Artificial* (CAEPIA), the forum where the research community working on AI meet together for presenting and discussing their research and developments.
- *Portuguese Conference on Artificial Intelligence* (EPIA), international conference hosted with the patronage of the Portuguese Association for Artificial Intelligence (APPIA).

Following, the most important conferences on NLP (or computational linguistics) and also the most important conference concerning the processing of Portuguese:

- *Annual Meeting of the Association for Computational Linguistics* (ACL), one of main meetings on computational linguistics which covers a broad spectrum of disciplines working towards enabling intelligent systems to interact with humans using natural language, and towards enhancing human-human communication through services such as speech recognition, automatic translation, information retrieval, text summarization, and information extraction
- *European Chapter of the Association for Computational Linguistics* (EACL), which is, as its name suggests, is an extension of the ACL meant for the European computational linguistics community.
- *North American Chapter of the Association for Computational Linguistics* (NAACL), which is an extension of the ACL meant for the North American computational linguistics community.
- *International Conference on Computational Linguistics* (COLING), one of the main conferences on computational linguistics, which covers a broad spectrum of technical areas related to natural language and computation.
- *Recent Advances in Natural Language Processing* (RANLP), an important conference concerning natural language processing, where recent advances in all aspects of this topic are reported.

- *International Conference on Language Resources and Evaluation* (LREC), the major event on Language Resources and Evaluation for Human Language Technologies.

- *International Conference on Computational Processing of Portuguese* (PROPOR), the main event in the area of Natural Language Processing that is focused on Portuguese and the theoretical and technological issues related to this specific language.

Furthermore, there are some important journals on NLP that worth mentioning:

- *Computational Linguistics*, the leading journal in the field of computational linguistics, published by The MIT Press for the Association for Computational Linguistics (ACL).

- *Natural Language Engineering*, an international journal designed to meet the needs of professionals and researchers working in all areas of computerised language processing, whether from the perspective of theoretical or descriptive linguistics, lexicology, computer science or engineering. Its principal aim is to bridge the gap between traditional computational linguistics research and the implementation of practical applications with potential real world use. Published by Cambridge University Press.

- *Procesamiento de Lenguaje Natural*, the journal of the Sociedad Española de Procesamiento de Lenguaje Natural (SEPLN), which aims to provide a forum to publish scientific-technical articles in the field of NLP, both to the Spanish and to the international scientific community.

- *Linguamática*, an open journal about natural language processing, which gives special focus to the Iberian languages. Published by the Department of Informatics of University of Minho.

Finally, the two most important Semantic Web conferences:

- *International Semantic Web Conference* (ISWC), the main conference on the Semantic Web, a major international forum where visionary and state-of-the-art research of all aspects of this topic are presented.

- *European Semantic Web Conference* (ESWC), the main European conference on the Semantic Web.

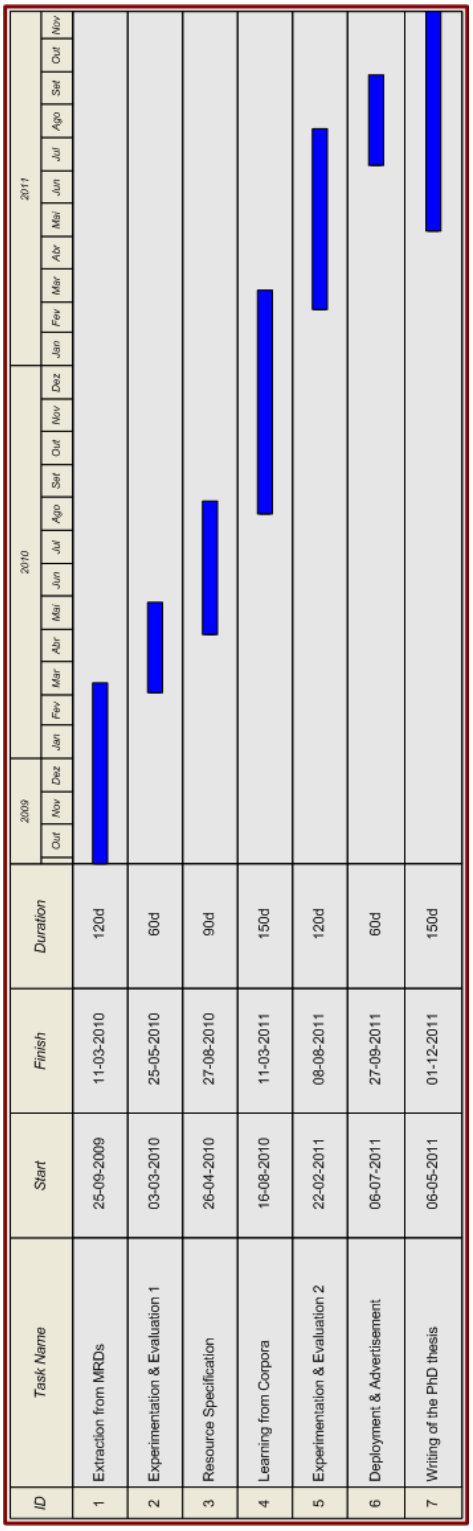

Figure 5.2: PhD further plan

Chapter 6

Conclusions

This research proposal is an answer to the growing demand on semantically aware applications. More precisely, its addresses the lack of public domain lexico-semantic resources for Portuguese. The importance of this kind of resources has been shown and the problems concerning the few existing alternatives for Portuguese were stated. While the existing resources are created manually, this research aims to take advantage of computational tools in order to create a lexical ontology for Portuguese by semi-automatic means. One of the biggest challenges involved is that for Portuguese the amount of existing NLP resources are scarce.

So, this research will be focused in the development of computational tools capable of extracting lexico-semantic knowledge from Portuguese textual resources. In a first phase, MRDs will be exploited in order to acquire general knowledge. The knowledge extracted will be structured into Onto.PT, a lexical ontology for Portuguese. Then, concerning the enrichment of Onto.PT, Portuguese textual corpora will be used as the second source of knowledge.

The approach to be taken will be based on searching for linguistic patterns that are indicative of lexico-semantic relations. Nevertheless, especially to what concerns corpora processing, statistical methods will also be tested. Therefore, whenever the results are improved by using statistical methods, they will surely be integrated in the extraction procedure.

Concerning the assessment of the quality and the utility of the resources developed, some of our work will be dedicated to evaluation, where, inspired by methods for the evaluation of ontologies, a semi-automatic method for the evaluation of Onto.PT will be devised. This can eventually be achieved by taking advantage of existing Portuguese NLP resources, such as corpora or thesaurus. Additionally, manual evaluation of a representative subset of Onto.PT will also take place, as well as the integration of this ontology with one or more real applications, towards the understanding of its potential utility.

In the end of this research, important contributions to Portuguese NLP are expected. Among the latter, the **new lexical resource**, **Onto.PT**, is the headline. It is foreseen that, in a near future, this resource, as well as the tools developed, might be broadly used by researchers and developers that work with Portuguese.

Furthermore, all relevant conclusions and results will be published in technical reports, scientific papers and articles, besides the resulting PhD thesis. There is still a long way to go, but results will come and future Portuguese NLP applications will have a useful resource to complement them and increase their potential.

References

(2005). *Dicionário PRO da Língua Portuguesa*. Porto Editora, Porto.

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.

Agirre, E., Mrquez, L., and Wicentowski, R., editors (2007). *Proceedings of SemEval-2007: 4th International Workshop on Semantic Evaluations*. ssociation for Computational Linguistics.

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.

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 the 19th annual meeting on Association for Computational Linguistics*, pages 133–138, Morristown, NJ, USA. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., and Etzioni, O. (2007). Open information extraction from the web. In Veloso, M. M., editor, *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2670–2676.

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, Canada. Adviser-Fred Popowich.

Baségio, T. L. (2006). Uma abordagem semi-automática para identificação de estruturas ontológicas a partir de textos na língua portuguesa do brasil. Master's thesis, Pontifícia Universidade Católica do Rio Grande do Sul – PUCRS.

Berland, M. and Charniak, E. (1999). Finding parts in very large corpora. In *Proceedings of 37th Annual Meeting of the ACL on Computational Linguistics*, pages 57–64, Morristown, NJ, USA. Association for Computational Linguistics.

Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The Semantic Web. *Scientific American*.

Bick, E. (2000). *The Parsing System PALAVRAS: Automatic Grammatical Analysis of Protuguese in a Constraint Grammar Framework*. PhD thesis, Arhus University, Arhus.

Biemann, C. (2005). Ontology learning from text: A survey of methods. *LDV-Forum*, 20(2):75–93.

Brank, J., Grobelnik, M., and Mladenic', D. (2005). A survey of ontology evaluation techniques. In *Proc. of the Conference on Data Mining and Data Warehouses (SiKDD)*.

Brewster, C., Alani, H., Dasmahapatra, S., and Wilks, Y. (2004). Data-driven ontology evaluation. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 164–168, Lisbon, Portugal. European Language Resources Association.

Brewster, C. and Wilks, Y. (2004). Ontologies, taxonomies, thesauri: Learning from texts. In Deegan, M., editor, *Proc. 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 Schek, H., Saltor, F., Ramos, I., and Alonso, G., editors, *Proceedings of 6th International Conference on Extending Database Technology, EDBT'98*, pages 172–183, London, UK. Springer-Verlag.

Briscoe, T. (1991). Lexical issues in natural language processing. In Klein, E. and Veltman, F., editors, *Natural Language and Speech: Symposium Proc.*, pages 39–68. Springer, Berlin, Heidelberg.

Brown, J. S. and Burton, R. R. (1975). Multiple representations of knowledge for tutorial reasoning. In Bobrow, D. G. and Collins, A., editors, *Representation and Understanding*, pages 311–349. Academic Press, New York.

Bruce, R. and Guthrie, L. (1992). Genus disambiguation: A study in weighted preference. In *Proceedings of the 14th COLING*, pages 1187–1191, Nantes, France.

Buitelaar, P., Cimiano, P., and Magnini, B. (2005). Ontology learning from text: An overview. In Buitelaar, P., Cimmiano, P., and Magnini, B., editors, *Ontology Learning from Text: Methods, Evaluation and Applications*. IOS Press.

Calzolari, N. (1977). An empirical approach to circularity in dictionary definitions. In *Cahiers de Lexicologie*, pages 118–128.

Calzolari, N. (1982). Towards the organization of lexical definitions on a database structure. In *Proceedings of the 9th conference on Computational linguistics*, pages 61–64, Prague, Czechoslovakia. Academia Praha.

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. Association for Computational Linguistics.

Calzolari, N., Pecchia, L., and Zampolli, A. (1980). Working on the italian machine dictionary: a semantic approach. In Zampolli, A. and Calzolari, N., editors, *Computational and Mathematical Linguistics, Proceedings of the International Conference on Computational Linguistics*, pages 49–70.

Cañas, A. J., Valerio, A., Lalinde-Pulido, J., Carvalho, M. M., and Arguedas, M. (2003). Using wordnet for word sense disambiguation to support concept map construction. In Nascimento, M. A., de Moura, E. S., and Oliveira, A. L., editors, *String Processing and Information Retrieval (SPIRE)*, volume 2857 of *LNCS*, pages 350–359.

Caraballo, S. A. (1999). Automatic construction of a hypernym-labeled noun hierarchy from text. In *Proceedings of the 37th annual meeting of the ACL on Computational Linguistics*, pages 120–126, Morristown, NJ, USA. Association for Computational Linguistics.

Cederberg, S. and Widdows, D. (2003). Using lsa and noun coordination information to improve the precision and recall of automatic hyponymy extraction. In Daelemans, W. and Osborne, M., editors, *In Proceedings of 7th Conference on Computational Natural Language Learning (CoNLL)*, pages 111–118, Morristown, NJ, USA. Association for Computational Linguistics.

Chatterjee, N., Goyal, S., and Naithani, A. (2005). Resolving pattern ambiguity for english to hindi machine translation using wordnet. In *Workshop on Modern Approaches in Translation Technologies*.

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*, pages 299–304, Morristown, NJ, USA. Association for Computational Linguistics.

Chomsky, N. (1956). Three models for the description of language. *Information Theory, IEEE Transactions on*, 2(3):113–124.

Cimiano, P. and Wenderoth, J. (2007). Automatic acquisition of ranked qualia structures from the web. In *Proc. 45th Annual Meeting of the ACL, June 23-30*. Association for Computer Linguistics.

Clark, P., Fellbaum, C., and Hobbs, J. (2008). Using and extending wordnet to support question-answering. In *Proceedings of 4th Global WordNet Conference (GWC)*.

Costa, L. F. and Cabral, L. M. (2008). Answering Portuguese Questions. In Teixeira, A., de Lima, V. L. S., de Oliveira, L. C., and Quaresma, P., editors, *Computational*

Processing of the Portuguese Language, 8th International Conference, Proceedings (PROPOR 2008), volume Vol. 5190, pages 228–231. Springer Verlag.

Costa, R. P. and Seco, N. (2008). Hyponymy extraction and web search behavior analysis based on query reformulation. In Geffner, H., Prada, R., Alexandre, I. M., and David, N., editors, *Proceedings of 11th Ibero-American Conference on AI (IBERAMIA 2008)*, volume 5290 of *LNCS*, pages 332–341. Springer Verlag.

Cruse, D. A. (1986). *Lexical Semantics*. Cambridge University Press, Cambridge.

Cuadros, M. and Rigau, G. (2006). Quality assessment of large scale knowledge resources. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 534–541, Sydney, Australia. Association for Computational Linguistics.

Cycorp (2002). Cyc 101 tutorial. `http://www.opencyc.org/doc/tut/?expand_all=1`.

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.

Demetriou, G. and Atwell, E. S. (2001). A domain-independent semantic tagger for the study of meaning associations in english text. In *Proceedings of 4th International Workshop on Computational Semantics (IWCS)*.

Dias-da-Silva, B. C. (2006). Wordnet.br: An exercise of human language technology research. In Petr Sojka, Key-Sun Choi, C. F. and Vossen, P., editors, *Proceedings of the 3rd International WordNet Conference (GWC 2006)*, pages 22–26, Jeju Island, Korea.

Dias da Silva, B. C., Oliveira, M., and Moraes, H. (2002). Groundwork for the Development of the Brazilian Portuguese Wordnet. In Mamede, N. and Ranchhod, E., editors, *Proceedings of Advances in Natural Language Processing: Third International Conference*, Lecture Notes in Artificial Intelligence, pages 189–196, Berlin/Heidelberg. Springer.

Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., and Weischedel, R. (2004). The Automatic Content Extraction (ACE) Program. Tasks, Data and Evaluation. In *Proceedings of 4th International Conference on Language Resources and Evaluation LREC*, pages 837–840.

Dolan, W., Vanderwende, L., and Richardson, S. D. (1993). Automatically deriving structured knowledge bases from online dictionaries. In *PACLING 93, Pacific Assoc. for Computational Linguistics*, pages 5–14.

Dolan, W. B. (1994). Word sense ambiguation: clustering related senses. In *Proceedings of the 15th conference on Computational linguistics*, pages 712–716, Morristown, NJ, USA. Association for Computational Linguistics.

Etzioni, O., Cafarella, M., 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 13th International Conference on World Wide Web (WWW)*, pages 100–110, New York, NY, USA. ACM.

Etzioni, O., Cafarella, M., Downey, D., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D. S., and Yates, A. (2005). Unsupervised named-entity extraction from the web: an experimental study. *Artificial Intelligence*, 165(1):91–134.

Faure, D. and Nédellec, C. (1998). ASIUM: learning subcategorization frames and restrictions of selection. In Kodratoff, Y., editor, *10th Conference on Machine Learning (ECML 98) – Workshop on Text Mining*.

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

Fillmore, C. J. (1982). Frame semantics. In Linguisticsocietykorea, editor, *Linguistics in the morning calm*. Seoul: Hanshin Publishing Co.

Frege, G. (1960). On sense and reference. In Geach, P. and Black, M., editors, *Translations from the Philosophical Writings of Gottlob Frege*. Oxford: Basil Blackwell.

Freitas, C. and Quental, V. (2007). Subsídios para a elaboração automática de taxonomias. In *XXVII Congresso da SBC - V Workshop em Tecnologia da Informação e da Linguagem Humana (TIL)*, pages 1585–1594.

Freitas, C., Rocha, P., and Bick, E. (2008). Floresta Sint (c)tica: Bigger, Thicker and Easier. In Teixeira, A., de Lima, V. L. S., de Oliveira, L. C., and Quaresma, P., editors, *Computational Processing of the Portuguese Language, 8th International Conference, Proceedings (PROPOR 2008)*, volume Vol. 5190, pages 216–219. Springer Verlag.

Freitas, C., Santos, D., Mota, C., Oliveira, H. G., and Carvalho, P. (2009). Detection of relations between named entities: report of a shared task. In *Proceedings of NAACL-HLT Workshop, Semantic Evaluations: Recent Achievements and Future Directions*.

Freitas, M. C. (2007). *Elaboração automática de ontologias de domínio: discussão e resultados*. PhD thesis, Pontifícia Universidade Católica do Rio de Janeiro.

Gamallo, P. (2008). Comparing window and syntax based strategies for semantic extraction. In Teixeira, A., de Lima, V. L. S., de Oliveira, L. C., and Quaresma, P., editors, *Computational Processing of the Portuguese Language, 8th International Conference, Proceedings (PROPOR 2008)*, volume Vol. 5190, pages 41–50. Springer Verlag.

Genesereth, M. R. and Nilsson, N. J. (1987). *Logical foundations of artificial intelligence*. Morgan Kaufmann, San Francisco.

Girju, R., Badulescu, A., and Moldovan, D. (2003a). 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. Association for Computational Linguistics.

Girju, R., Badulescu, A., and Moldovan, D. (2003b). Learning semantic constraints for the automatic discovery of part-whole relations. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 1–8, Morristown, NJ, USA. Association for Computational Linguistics.

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 Haller, S. M. and Simmons, G., editors, *Proc. 15th Intl. Florida Artificial Intelligence Research Society Conference (FLAIRS)*, 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 *Canadian Conference on AI*, pages 537–543.

Gonçalo Oliveira, H. (2009). Ontology learning for portuguese. In *2nd Doctoral Symposium on Artificial Intelligence*.

Gonçalo Oliveira, H., Gomes, P., Santos, D., and Seco, N. (2008). PAPEL: a dictionary-based lexical ontology for Portuguese. In Teixeira, A., de Lima, V. L. S., de Oliveira, L. C., and Quaresma, P., editors, *Computational Processing of the Portuguese Language, 8th International Conference, Proceedings (PROPOR 2008)*, volume Vol. 5190, pages 31–40. Springer Verlag.

Gonçalo Oliveira, H., Santos, D., and Gomes, P. (2009a). Avaliação da extracção de relações semânticas entre palavras portuguesas a partir de um dicionário. In *The 7th Brazilian Symposium in Information and Human Language Technology (STIL 2009)*.

Gonçalo Oliveira, H., Santos, D., and Gomes, P. (2009b). Relations extracted from a portuguese dictionary: results and first evaluation. In *Local Proceedings of 14th Portuguese Conference on Artificial Intelligence (EPIA)*.

Gonzalo, J., Verdejo, F., Peters, C., and Calzolari, N. (1998). Applying eurowordnet to cross-language text retrieval. In *EuroWordNet: a multilingual database with lexical semantic networks*, pages 113–135. Kluwer Academic Publishers, Norwell, MA, USA.

Grcar, M., Grobelnik, M., and Mladenic, D. (2008). Using text mining and link analysis for software mining. In *Proceedings of 3rd International Workshop on Mining Complex Data (MCD)*, volume 4944 of *LNCS*, pages 1–12. Springer Verlag.

Grishman, R. (1997). Information extraction: Techniques and challenges. In *International Summer School on Information Extraction (SCIE)*, pages 10–27.

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.

Guthrie, L., Slater, B. M., Wilks, Y., and Bruce, R. (1990). Is there content in empty heads? In *Proceedings of the 13th COLING*, volume 3, pages 138–143.

Harris, Z. S. (1968). *Mathematical Structures of Language*. Wiley, New York, NY, USA.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics*, pages 539–545, Morristown, NJ, USA. Association for Computational Linguistics.

Hervás, R., Pereira, F. C., Gervás, P., and Cardoso, A. (2006). A text generation system that uses simple rhetorical figures. *Procesamiento de Lenguaje Natural*, 37:199–206.

hui Lee, C. and hwang Juang, B. (1996). A survey on automatic speech recognition with an illustrative example on continuous speech recognition of mandarin. *International Journal of Computational Linguistics and Chinese Language Processing*, 1(1):1–36.

Hutchins, W. J. and Somers, H. L. (1992). *An Introduction to Machine Translation*. Academic Press.

Ide, N. and Véronis, J. (1993). Refining taxonomies extracted from machine readable dictionaries. In Hockey, S. and Ide, N., editors, *Research in Humanities Computing 2*.

Ide, N. and Véronis, J. (1994). Machine readable dictionaries: What have we learned, where do we go. In *In Proceedings of the post-COLING 94 intl. workshop on directions of lexical research, Beijing*, pages 137–146.

Ide, N. and Veronis, J. (1995). Knowledge extraction from machine-readable dictionaries: An evaluation. In Steffens, P., editor, *Machine Translation and the Lexicon, LNAI*. Springer-Verlag.

Ikehara, S., Miyazaki, M., Shirai, S., Yokoo, A., Nakaiwa, H., Ogura, K., Ooyama, Y., and Hayashi, Y. (1997). *Goi-Taikei – A Japanese Lexicon*. Iwanami Shoten, Tokyo, Japan.

Jurafsky, D. and Martin, J. H. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall PTR, Upper Saddle River, NJ, USA.

Kaisser, M. (2005). Qualim at trec 2005: Web-question answering with framenet. In Voorhees, E. M. and Buckland, L. P., editors, *Proceedings of 14th Text REtrieval Conference (TREC)*, volume Special Publication 500-266. National Institute of Standards and Technology (NIST).

Kavalec, M. and Svatek, V. (2005). A study on automated relation labelling in ontology learning. In Buitelaar, P., Cimmiano, P., and Magnini, B., editors, *Ontology Learning from Text: Methods, Evaluation and Applications*. IOS Press.

Khoo, C. S. G., Chan, S., and Niu, Y. (2000). Extracting causal knowledge from a medical database using graphical patterns. In *ACL '00: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 336–343, Morristown, NJ, USA. Association for Computational Linguistics.

Kietz, J., Maedche, A., and Volz, R. (2000). A method for semi-automatic ontology acquisition from a corporate intranet. In *Proceedings of Workshop Ontologies and Text, co-located with the 12th International Workshop on Knowledge Engineering and Knowledge Management (EKAW'2000)*, pages 2–6, Juan-Les-Pins, France.

Kilgarriff, A. (1996). Word senses are not bona fide objects: implications for cognitive science, formal semantics, nlp. In *Proceedings of the 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.

Lenat, D. B. (1995). Cyc: a large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38.

Lenat, D. B. and Guha, R. V. (1989). *Building Large Knowledge-Based Systems; Representation and Inference in the Cyc Project*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Lenat, D. B. and Guha, R. V. (1991). Ideas for applying cyc. Technical report, Microelectronics and Computer Technology Corporation, Advanced Computing Technology Artificial Intelligence Lab. Technical Report ACT-CYC-407-91.

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.

Maedche, A. and Staab, S. (2002). Measuring similarity between ontologies. In *Proceedings of 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web (EKAW)*, pages 251–263, London, UK. Springer Verlag.

Mahesh, K., Nirenburg, S., Cowie, J., and Farwell, D. (1996). An assessment of Cyc for natural language processing. Technical Report MCCS-96-302, CRL, New Mexico State University, Las Cruces, New Mexico.

Mani, I. and Maybury, M. T. (1998). *Advances in Automatic Text Summarization*. MIT 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*, pages 112–119, Morristown, NJ, USA. Association for Computational Linguistics.

Marrafa, P. (2002). Portuguese wordnet: general architecture and internal semantic relations. *DELTA*, 18:131–146.

Marrafa, P., Amaro, R., Chaves, R. P., Lourosa, S., Martins, C., and Mendes, S. (2006). Wordnet.pt new directions. In Sojka, P., Choi, K.-S., Fellbaum, C., and Vossen, P., editors, *Proceedings of the 3rd International WordNet Conference (GWC 2006)*, pages 319–320.

Maziero, E., Pardo, T., Di Felippo, A., and Dias-da Silva, B. (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.

McGuinness, D. L. and van Harmelen, F. (2004). OWL web ontology language overview. Published: W3C Recommendation.

Mendes, S. (2006). Adjectives in wordnet.pt. In Sojka, P., Choi, K.-S., Fellbaum, C., and Vossen, P., editors, *Proceedings of the 3rd International WordNet Conference (GWC 2006)*, pages 225–230.

Michiels, A., Mullenders, J., and Noël, J. (1980). Exploiting a large data base by Longman. In *Proceedings of the 8th conference on Computational linguistics*, pages 374–382, Morristown, NJ, USA. Association for Computational Linguistics.

Miller, E. and Manola, F. (2004). RDF primer. Published: W3C Recommendation.

Mitkov, R. (1999). Anaphora resolution: the state of the art. Working paper, (Based on the COLING'98/ACL'98 tutorial on anaphora resolution), University of Wolverhampton, Wolverhampton.

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*, pages 546–552, Morristown, NJ, USA. Association for Computational Linguistics.

Morato, J., Marzal, M. A., Lloréns, J., and Moreiro, J. (2004). Wordnet applications. In *Proceedings of 2nd Global Wordnet Conference (GWC)*.

Morin, E. and Jacquemin, C. (2004). Automatic acquisition and expansion of hypernym links. *Computer and the Humanities*, 38(4):343–362.

Mota, C. and Santos, D., editors (2008). *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca.

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*, pages 459–464, Morristown, NJ, USA. Association for Computational Linguistics.

Nancy Ide, J. V. (1998). Introduction to the special issue on word sense disambiguation: The state of the art. *Computational Linguistics*, 24:1–40.

Navigli, R., Velardi, P., Cucchiarelli, A., and Neri, F. (2004). Quantitative and qualitative evaluation of the ontolearn ontology learning system. In *Proc. 20th Intl. conference on Computational Linguistics*, Morristown, NJ, USA. Association for Computational Linguistics.

Nichols, E., Bond, F., and Flickinger, D. (2005). Robust ontology acquisition from machine-readable dictionaries. In Kaelbling, L. P. and Saffiotti, A., editors, *Proceedings of the 19th International. Joint Conference on Artificial Intelligence (IJCAI)*, pages 1111–1116. Professional Book Center.

Noy, N. F. and McGuinness, D. (2000). Ontology development 101: A guide to creating your first ontology. *Stanford KSL Technical Report KSL-01-05*.

O'Hara, T. P. (2005). *Empirical Acquisition of Conceptual Distinctions via Dictionary Definitions*. PhD thesis, NMSU CS.

Olney, J., Revard, C., and Ziff, P. (1967). Summary of some computational aids for obtaining a formal semantic description of english. In *Proceedings of the 1967 conference on Computational linguistics*, pages 1–5, Morristown, NJ, USA. Association for Computational Linguistics.

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 Ravichandran, D. (2004). Automatically labeling semantic classes. In *Proceedings of Human Language Technology/North American Chapter of the Association for Computational Linguistics (HLT/NAACL)*, pages 321–328.

Pantel, P., Ravichandran, D., and Hovy, E. (2004). Towards terascale knowledge acquisition. In *COLING '04: Proceedings of the 20th international conference on Computational Linguistics*, pages 771–777, Morristown, NJ, USA. Association for Computational Linguistics.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems : Networks of Plausible Inference*. Morgan Kaufmann.

Peters, C., Jijkoun, V., Mandl, T., Mller, H., Oard, D. W., Peas, A., Petras, V., and Santos, D., editors (2008). *Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers*, volume 5152 of *LNCS*. Springer Verlag, Berlin.

Pianta, E., Bentivogli, L., and Girardi, C. (2002). Multiwordnet: developing an aligned multilingual database. In *1st International Conference on Global WordNet*.

Porzel, R. and Malaka, R. (2004). A task-based approach for ontology evaluation. In Buitelaar, P., Handschuh, S., and Magnini, B., editors, *Proceedings of ECAI 2004 Workshop on Ontology Learning and Population*, Valencia, Spain.

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.

Raman, J. and Bhattacharyya, P. (2008). Towards automatic evaluation of wordnet synsets. In Tancs, A., Csendes, D., Vincze, V., Fellbaum, C., and Vossen, P., editors, *Proceedings of the 4th Global WordNet Conference (GWC 2008)*, Szeged, Hungary. University of Szeged, Department of Informatics.

Reiter, E. and Dale, R. (1997). Building applied natural language generation systems. *Natural Language Engineering*, 3(1):57–87.

Ribeiro Junior, L. C. (2008). *OntoLP: Construção Semi-Automática de Ontologias a partir de Textos da Língua Portuguesa*. PhD thesis, Centro de Ciências Exatas e Tecnológicas, Universidade do Vale do Rio dos Sinos, São Leopoldo.

Richardson, S. (1997). *Determining Similarity and Inferring Relations in a Lexical Knowledge Base*. PhD thesis, The City University of New York, New York, NY.

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., Dolan, W. B., and Vanderwende, L. (1998). Mindnet: Acquiring and structuring semantic information from text. In *COLING-ACL*, pages 1098–1102.

Riloff, E. and Shepherd, J. (1997). A corpus-based approach for building semantic lexicons. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, pages 117–124.

Roark, B. and Charniak, E. (1998). Noun-phrase co-occurrence statistics for semi-automatic semantic lexicon construction. In *Proceedings of the 17th International Conference on Computational linguistics*, pages 1110–1116, Morristown, NJ, USA. Association for Computational Linguistics.

Rocha, P. A. and Santos, D. (2000). CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. In das Graças Volpe Nunes, M., editor, *V Encontro para o processamento computacional da lngua portuguesa escrita e falada (PROPOR)*, pages 131–140, São Paulo. ICMC/USP.

Roget, P. M. (1852). *Roget's Thesaurus of English Words and Phrases*. Longman, London.

Salton, G. and McGill, M. J. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill, New York.

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. (1999). Towards language-specific applications. *Machine Translation*, 14(2):83–112.

Santos, D. (2000). O projecto Processamento Computacional do Português: Balanço e perspectivas. In das Graças Volpe Nunes, M., editor, *V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR 2000)*, pages 105–113, São Paulo. ICMC/USP.

Santos, D. (2002). Um centro de recursos para o processamento computacional do portugus. *DataGramaZero - Revista de Cincia da Informao*, 3(1).

Santos, D., editor (2007a). *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. IST Press, Lisboa, Portugal.

Santos, D. (2007b). Evaluation in natural language processing. European Summer School on Language, Logic and Information (ESSLLI).

Santos, D. (2009). Caminhos percorridos no mapa da portuguesificação: A Linguateca em perspectiva. *Linguamática*, 1(1):25–59.

Santos, D., Barreiro, A., Costa, L., Freitas, C., Gomes, P., Gonalo Oliveira, H., Medeiros, J. C., and Silva, R. (2009). O papel das relações semânticas em português: Comparando o TeP, o MWN.PT e o PAPEL. In *XXV Encontro Nacional da Associação Portuguesa de Linguística*.

Santos, D. and Cardoso, N., editors (2007). *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. Linguateca.

Santos, D., Freitas, C., Oliveira, H. G., and Carvalho, P. (2008). Second HAREM: new challenges and old wisdom. In Teixeira, A., de Lima, V. L. S., de Oliveira, L. C., and Quaresma, P., editors, *Computational Processing of the Portuguese Language, 8th International Conference, Proceedings (PROPOR 2008)*, volume Vol. 5190, pages 212–215. Springer Verlag.

Santos, D. and Rocha, P. (2001). Evaluating CETEMPúblico, a free resource for Portuguese. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 442–449.

Santos, D., Simes, A., Frankenberg-Garcia, A., Pinto, A., Barreiro, A., Maia, B., Mota, C., Oliveira, D., Bick, E., Ranchhod, E., ao Dias de Almeida, J. J., Cabral, L., Costa, L., Sarmento, L., Chaves, M., Cardoso, N., Rocha, P., Aires, R., Silva, R., Vilela, R., and Afonso, S. (2004). Linguateca: um centro de recursos distribuído para o processamento computacional da língua portuguesa. In Luna, G. D. I., Chávez, O. F., and Galindo, M. O., editors, *Proceedings of the international workshop "Taller de Herramientas y Recursos Linguísticos para el Español y el Portugués", IX Iberoamerican Conference on Artificial Intelligence (IBERAMIA 2004)*, pages 147–154.

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)*.

Simões, A. M. and Almeida, J. (2002). Jspell.pm – um módulo de análise morfológica para uso em processamento de linguagem natural. In *Actas do XVII Encontro da Associação Portuguesa de Linguística*, pages 485–495, Lisboa.

Smith, B. (2004). Ontology and information systems.

Smullyan, R. (1995). *First-order logic*. Dover.

Snow, R., Jurafsky, D., and Ng, A. Y. (2005). Learning syntactic patterns for automatic hypernym discovery. In Saul, L. K., Weiss, Y., and Bottou, L., editors, *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 the 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. *Conceptual structures: current research and practice*, pages 3–51.

Strzalkowski, T. and Harabagiu, S. (2006). *Advances in Open Domain Question Answering (Text, Speech and Language Technology)*. Springer-Verlag, Secaucus, NJ, USA.

Turney, P. D. (2001). Mining the web for synonyms: PMI–IR versus LSA on TOEFL. In Raedt, L. D. and Flach, P., editors, *Proceedings of 12th European Conference on Machine Learning (ECML-2001)*, volume 2167, pages 491–502. Springer-Verlag.

van Heijst, G., Schreiber, A. T., and Wielinga, B. J. (1997). Using explicit ontologies in kbs development. *International Journal of Human-Computer Studies*, 46(2–3):183–292.

Vanderwende, L. (1994). Algorithm for automatic interpretation of noun sequences. In *Proceedings of the 15th conference on Computational linguistics*, pages 782–788, Morristown, NJ, USA. Association for Computational Linguistics.

Vanderwende, L. (1995). Ambiguity in the acquisition of lexical information. In *Proceedings of the AAAI 1995 Spring Symposium Series*, pages 174–179. symposium on representation and acquisition of lexical knowledge.

Vanderwende, L., Kacmarcik, G., Suzuki, H., and Menezes, A. (2005). Mindnet: An automatically-created lexical resource. In *HLT/EMNLP*. The Association for Computational Linguistics.

Veiga, P. and Santos, D. (2001). Contributo para o processamento computacional do português: o CRdLP. In Mateus, M. H. M., editor, *Mais Lnguas, Mais Europa: celebrar a diversidade lingustica e cultural da Europa*, pages 103–109. Colibri, Lisboa.

von Ahn, L. (2006). Games with a purpose. *IEEE Computer*, 39:92–94.

Vossen, P. (1997). Eurowordnet: a multilingual database for information retrieval. In *Proceedings of DELOS workshop on Cross-Language Information Retrieval*, Zurich.

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*, pages 750–755, Morristown, NJ, USA. Association for Computational Linguistics.

Zúñiga, G. L. (2001). Ontology: its transformation from philosophy to information systems. In *Proceedings of the International Conference on Formal Ontology in Information Systems (FOIS'01)*, pages 187–197. ACM Press.